广义洛特卡定律

GUANGYILUOTEKADINGLü

—估计、推论及其在管理中的应用

张贤澳 著

厦门大学出版社
XIAMENDAXUECHUBANSHE

广义洛特卡定律

——估计、推论及其在管理中的应用

张贤澳　著

厦门大学出版社

图书在版编目(CIP)数据

广义洛特卡定律:估计、推论及其在管理中的应用/张贤澳著.
—厦门:厦门大学出版社,2002.6
ISBN 7-5615-1886-2

Ⅰ.广… Ⅱ.张… Ⅲ.①数理统计-基本知识②数理统计-应用-管理学 Ⅳ.O212

中国版本图书馆 CIP 数据核字(2002)第 029190 号

厦门大学出版社出版发行
(地址:厦门大学 邮编:361005)
http://www.xmupress.com
xmup @ public.xm.fj.cn
福州市鼓楼印刷精装厂印刷
2002 年 6 月第 1 版 2002 年 6 月第 1 次印刷
开本:850×1168 1/32 印张:5
字数:125 千字 印数:1—1 500 册
定价:20.00 元

前　言

洛特卡定律（Lotka's Law）、布拉德福定律（Bradford's Law of scattering）、齐夫定律（Zipf's Law）合称情报学或文献计量学的三大定律。布拉德福定律描述和揭示了某一主题或专业文献在原始期刊中发散分布的规律，1934 年由英国著名文献学家塞缪尔·克莱蒙特·布拉德福（Samuel Clement Bradford）提出；齐夫定律描述和揭示了自然语言的词频等级分布的规律，1929 年由美国著名的语言学家和心理学家乔治·琴斯利·齐夫（George Kingslye Zipf）提出；洛特卡定律描述和揭示了不同科学生产率水平的科技工作者分布的规律，1926 年由美国兴趣广泛的数学家阿尔弗莱德·詹姆斯·洛特卡（Alfred James Lotka）提出。

洛特卡（Alfred J. Lotka）1880 年 3 月 2 日出生于奥地利莱姆伯格，父母是美国人。1901 年毕业于英国伯明翰大学并获得理学学士学位；1909 年在康奈尔大学获得文学硕士学位；1912 年在伯明翰大学获理学博士学位；1901—1902 曾在德国莱比锡大学攻读过研究生课程。洛特卡工作经历丰富，曾任美国化学总公司的助理化学师、化学师，康奈尔大学物理学助教，美国专利办公室审定，美国标准局助理物理学家，科学美国人副刊编辑。1924 年起在都市人寿保险公司统计处工作，任监督、总监；1934 年起任该公司的助理统计师。洛特卡学术兴趣广泛，研究的触角涉及多个学科，他是美国统计协会、科学发展协会、人口协会、经济学会、数学学会、物理学会、生态学会、公众健康协会等的会员，瑞典

保险统计学会会员，国际人口调查联合会和华盛顿科学院的成员。洛特卡一生著述颇丰，在科技期刊上发表为数众多的论文，论文涉及人口的数学分析、进化的数学原理等；1925 年出版了著作《自然生物学原理》，1930 年与杜布林（L. I. Dublin）合作，根据美国 1920 年人口普查资料，利用生命表技术建立了一个人的货币价值（即收入能力）随年龄变化的模型，出版了《人的货币价值》（The Money Value of a Man）一书。该书对人力资本投资和收益与生命周期关系进行了系统的科学研究。毋庸讳言，1940—1941 年阿尔伯特·内尔森·马奎斯公司出版的《美国名人录》在介绍洛特卡时，并未提及“科学生产率的频次分布”[1]。事实上，在 1949 年洛特卡谢世之年，他的这个成果才被称为洛特卡定律。无独有偶，布拉德福于 1934 年发表的论文，直到 1948 年他与世长辞时，才引起文献学家维克利（B. C. Vickery）的关注。虽然洛特卡的“科学生产率的频次分布”的研究成果在 1949 年被称为洛特卡定律，但是真正在学术界“重见天日”并被广泛传播，却是在 20 世纪 60 年代初——得益于著名科学学家 D·普赖斯（D. Price）的两部著作《巴比伦以来的科学》（1961）和《小科学、大科学》（1963）的出版。

由于洛特卡“科学生产率的频率分布”所依据的数据学科范围小，拟合的方法粗糙，缺少统计检验等，20 世纪 60 年代后，学术界对洛特卡定律的研究，主要朝着完善洛特卡定律的方向发展。这其中包括验证性研究、普适性研究、拟合方法研究、特征参数研究、合著现象研究等等，还出现了基于狭义洛特卡定律的普赖斯定律和伊格公式等。我国学术界对于洛特卡定律的研究，由于起步较晚（20 世纪 80 年代才介绍到我国）研究的方向虽与国外研究相比有所侧重，但研究的基本趋势是一致的，也颇有进展。

21 世纪的经济是以知识为基础的经济，即在经济活动中，知识已成为交换和应用的主体。我国目前正处于知识经济的萌芽阶

段，其知识经济发展水平在整体上比知识经济最发达的美国要落后40多年（厦门日报，1999年1月11日，第4版）。如果把美国现有水平作为一个发展标准，以1来衡量的话，我国仅约有0.26。如何制定正确的经济发展战略，迎头赶上发达的国家？粗略地讲，激励知识生产，建立知识创新体系尤为重要。洛特卡定律揭示了不同科学生产率水平的科技工作者的分布，为正确评价科技工作者，开发人力资源提供了理论依据，进而激励科技工作者进行知识创新。因此，研究洛特卡定律具有十分重要的现实意义。同时，洛特卡定律精美简洁的数学表达式，与社会科学其他学科中描述“发生源”与“产物”之间关系的数学表达式如出一辙，都服从负幂函数分布，如不同收入水平人群的分布等，从而洛特卡定律参数估计与特征研究，理所当然地为具有相同分布的“发生源”与“产物”之间的研究带来福音。

本书是作者在汇集了近十年的研究成果的基础上写就的，旨在抛砖引玉，推进洛特卡定律及具有与其相同分布的社会科学中的其他规律的研究。书中不当之处，敬请批评指正。

作　者

目　录

第一章　洛特卡定律的由来与发展

1.1　洛特卡定律的由来

1926 年，时任美国纽约都市人寿保险公司统计处监督的洛特卡(Afred J. Lotka，1880—1949)在《华盛顿科学院会刊》(Journal of the Washington Academy of Science) 上发表了一篇当时并不十分引人注目的论文“科学生产率的频次分布”(The Frequency Distribution of Scientific Productivity)，首次研究并揭示了科技文献数量与著者数量之间的关系。20 世纪 60 年代初期，随着普赖斯的两部著作的出版，洛特卡定律才得以广泛流传，并成为文献计量学的三大定律之一(布拉德福定律、齐夫定律)。所谓科学生产率是评价科学工作者对科学发展所做贡献的一个重要指标，假定科学工作者每一篇论文对科学的贡献是一样的，那么在给定的时间内，科学工作者所发表的论文数量就称之为科学生产率(Scientific Productivity)。

为了考察研究科学生产率按著者分布的情况，洛特卡统计了化学和物理学两个学科著者的数据。他选用《化学文摘》和一个 10 年累积索引 (Decennial Index to Chemical Abstract 1907—1916)，统计其中姓名以 A、B 两字母打头的所有著者 (只统计第一著者)。对于物理学，他选用了德国著名学者奥尔巴赫(Auerbach) 所编的《物理学史一览表》 (Geschichtstafeln der

Physik，1910 年莱比锡出版）的人名索引，其中包括 1900 年以来有突出贡献的 1 325 名物理学家。统计数据如表 1—1。

表 1—1　科学生产率频率分布（摘录）

论著数量	撰写人数				占总数百分比（%）					
	化学文摘			物理学史一览表	化学文摘				物理学史一览表	
					观察值			计算值	观察值	计算值
	A	B	A+B		A	B	A+B	A+B	全部作者	
总数	1 543	5 348	6 981	1 325	—	—	—	—	—	—
1	890	3 101	3 991	784	57.68	57.98	57.92	56.69	59.17	60.79
2	230	829	1 059	204	14.91	15.50	15.37	15.32	15.40	15.20
3	111	882	493	127	7.19	7.14	7.15	7.12	9.58	6.75
4	58	829	287	50	3.76	4.28	4.16	4.14	3.77	3.80
5	41	43	184	33	2.66	2.61	2.67	2.72	2.49	2.43
6	42	89	131	28	2.72	1.66	1.90	1.92	2.11	1.69
7	20	93	113	19	1.30	1.74	1.64	1.44	1.43	1.24
8	24	61	85	19	1.56	1.14	1.23	1.12	1.43	0.95
9	21	43	64	6	1.36	0.80	0.93	0.90	0.45	0.75
10	15	50	65	7	0.97	0.93	0.94	0.73	0.53	0.61
11	9	32	41	6	0.58	0.60	0.59	0.61	0.45	0.50
12	11	36	47	7	0.71	0.67	0.68	0.52	0.53	0.42
13	6	26	32	4	0.39	0.49	0.46	0.45	0.30	0.36
14	7	21	28	4	0.45	0.39	0.41	0.39	0.30	0.31
15	3	18	21	5	0.19	0.34	0.30	0.34	0.38	0.27
16	4	20	24	3	0.26	0.37	0.35	0.30	0.23	0.24
17	4	14	18	3	0.26	0.26	0.26	0.27	0.23	0.21
18	5	14	19	1	0.32	0.26	0.28	0.24	—	—
19	3	14	17	0	0.19	0.26	0.25	0.22	—	—
20	6	8	14	0	0.39	0.15	0.20	0.20	—	—
21	0	9	9	1	—	0.17	0.13	0.18	—	—
22	2	9	11	3	0.13	0.17	0.16	0.17	—	—

续表

论著数量	撰写人数				占总数百分比（%）					
	化学文摘			物理学史一览表	化学文摘				物理学史一览表	
					观察值			计算值	观察值	计算值
	A	B	A+B		A	B	A+B	A+B	全部作者	
23	4	4	8	0	0.26	0.07	0.12	0.15	—	—
24	4	4	8	3	0.26	0.07	0.12	0.14	—	—
25	0	9	9	2	—	0.17	0.13	0.13	—	—
26	3	6	9	0	0.19	0.11	0.13	0.12	—	—
27	1	7	8	1	0.06	0.13	0.12	0.11	—	—
28	2	8	10	0	0.13	0.15	0.15	0.11	—	—
29	2	6	8	0	0.13	0.11	0.12	0.10	—	—
30	2	5	7	1	0.13	0.09	0.10	0.09	—	—
…	…	…	…	…	…	…	…	…	…	…
114	0	1	1	—	—	0.02	0.01	—	—	—
115～345	1	0	1	—	—	—	—	—	—	—
346	1	0	1	—	0.06	—	0.01	—	—	—

洛特卡采用直角对数坐标系——坐标轴采用对数刻度的坐标系，以横轴表示发表的论文数量（表1—1中，第1列的数字），以纵轴表示论文的作者数量的百分数，即发表某一数量论文作者数与全部论文作者数比值的百分数（表1—1中，第8列、第10列数字），发现由表1—1中所列相关数字所描成点的连线，非常近似直线。这意味着经过对数处理，发表过 x 篇论文的作者相对百分数 f_x 与论文数量 x 之间存在很密切的线性关系。经研究，洛特卡提出了被后人称为洛特卡定律的数学表达式：

$$f_x = \frac{C}{x^2}$$

式中：f_x 表示发表 x 篇论文的科技工作者占被统计的该学科科学工作者总数的百分比；x 表示科学工作者发表的论文数量；C

为常数，即为发表一篇论文的科学工作者占该学科科学工作者总数的百分比。

由于：$f_1 = \dfrac{C}{1^2} = C$

$$f_2 = \frac{C}{2^2}$$

$$\cdots$$

$$f_n = \frac{C}{n^2}$$

则有：$C \cdot \sum_1^n \dfrac{1}{x^2} = C\left(\dfrac{1}{1^2} + \dfrac{1}{2^2} + \cdots + \dfrac{1}{n^2}\right)$

$$= f_1 + f_2 + \cdots + f_n$$

$$= \sum_1^n f_x$$

$$= 1$$

为了确定 C 的极限值，不妨设 $n \to \infty$，则有：

$$C\sum_1^{\infty} \frac{1}{x^2} = 1$$

将 t^2 在$[-\pi, \pi]$上展成傅立叶级数，有：

$$t^2 = \frac{\pi^2}{3} - 4\left(\frac{\cos t}{1^2} - \frac{\cos 2t}{2^2} + \frac{\cos 3t}{3^2} - \cdots\right)$$

令 $t^2 = \pi$，可得：

$$\sum_1^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6}(x \text{ 为非零自然数})$$

于是：$C \sum_1^{\infty} \dfrac{1}{x^2} = 1$

即：$C \cdot \dfrac{\pi^2}{6} = 1$

从而：$C = \dfrac{6}{\pi^2} \approx 0.607\ 9$

归纳上述结果，洛特卡定律可用文字表述为：写 2 篇论文的

作者的数量约为写1篇论文作者的数量的$\frac{1}{4}$；写3篇论文的作者的数量约为写1篇论文作者的数量的$\frac{1}{9}$；而写1篇论文作者的数量约占作者总数的60%。也有人将其称为“平方反比律”（The Inverse Square Law）。

洛特卡本人将该成果称为“科学生产率的频次分布”，不同的学科和主题领域的科技工作者科学生产率的频率具有相当的稳定性（详见第四章）。同时，如果我们定义：

$$F(x)=\sum_{z=-\infty}^{x} f_z$$

且：$f_z=\begin{cases}\dfrac{C}{k^n} & z\in(k-1,k] \quad k\text{为非零自然数},n>1\\ 0 & z\in(-\infty,0]\end{cases}$

不难看出$F(x)$满足：

(1)$F(x)$是一个单值实函数；

(2)非减且至少左连续；

(3)$F(-\infty)=0,F(+\infty)=1$。

从而，用累积的科学生产率作者频率表达的$F(x)$具有分布函数的形式，因此，有时我们也将洛特卡定律用累积的科学生产率的作者频率描述的$F(x)$，称为洛特卡分布。

1.2 洛特卡定律的发展

洛特卡使用物理、化学两个学科科学生产率的数据，应用非常初始的方法——在以对数为标度的平面直角坐标图上描点，用肉眼确定能用于计算回归方程的点的方法，归结出一个十分精美的数学公式：$f_x=\frac{C}{x_2}$，在人类历史上第一次提出了不同著述水平的科学工作者的频率服从平方反比律，第一次揭示了科学生产率

与某种分布之间的关系。其成果对科学的贡献不言而喻。但洛特卡定律必竟是从经验数据经统计分析而得出的规律，其学科覆盖面、拟合方法都存在不足，当洛特卡定律在学术界广泛传播后，很自然就引发了学术界对洛特卡定律普适性研究的兴趣，即在不同的学科领域采集科学生产率的数据，验证它们是否符合洛特卡分布。而要验证，不可避免地要涉及最优拟合方法、统计检验等问题。当初洛特卡发表“科学生产率的频次分布”的论文时，文中并没有提供一种统计检验的方法和检验结果，后来者要使同行信服，要验证自己采集的科学生产率数据是否服从洛特卡分布，必须去寻找一种适合自己手头问题的统计的评判方法。有多大的可信度来判断一组科学生产率数据符合洛特卡分布，这个问题在20世纪70年代由科尔（R.C.Coile）解决了。最优拟合方法到20世纪90年代才彻底解决。

20世纪80年代，较大规模有代表性地对洛特卡定律进行经验性检验的有美国伊利诺斯大学图书馆学院篇名记录数据的统计分析、美国国会图书馆机读目录数据的统计分析，以及M·L·鲍的工作。从美国伊利诺斯大学图书馆学院的250万条篇名记录中随机抽取2 345条记录，科学生产率数据通过统计分析，得出该数据严格服从洛特卡分布的结论。从美国国会图书馆1969—1979年的机读目录中抽取的数据，经过相同方法的统计分析，却得不出与伊利诺斯大学数据相同的结论。1986年，M·L·鲍收集了可供利用的前人使用过的20多个学科领域和3个大型研究图书馆的目录共计48组数据，经过数学处理，发现仅有7组数据符合平方反比律，即$n=2$的洛特卡分布；而不拘泥于平方反比律，即n可以取不包括2在内的其他值，则有32组（事实上是34组）数据服从洛特卡分布[2]。

由于出现大量与洛特卡初衷不相符的情况，引起了学术界的思考与争论。对于伊利诺斯大学图书馆学院的数据，有人认为，数

据的跨度大，几乎覆盖了有史以来出现的作者和论著，因而服从洛特卡分布。对于美国国会图书馆机读目录的数据，有人认为统计时限仅 10 年，不足以全面反映科学生产率的分布情况，因而不服从洛特卡分布。这种说法实际上也殃及洛特卡本人使用过的《化学文摘》的数据，洛特卡当年使用的《化学文摘》就是 10 年的累积索引，其数据的统计跨度当然也仅有 10 年。M·L·鲍收集的 48 组数据中，包含了洛特卡本人使用过的《化学文摘》以姓名首字母为 A、B 及 A 和 B 混并起来的科学生产率数据，以及伊利诺斯大学的数据和国会图书馆的数据（参见附录）。M·L·鲍认为国会图书馆的数据包含了所有形式出版物的著者，即图书、连续出版物、地图及商业性出版物等著者，虽然除图书之外的出版物的著者所占比例很小，但这些数据的渗入会影响洛特卡分布，即，即使 n 不囿于 2，经过 K－S 检验，这些数据也不服从洛特卡分布。

鲍使用的 48 组数据，其涵盖的学科大大超过洛特卡本人使用过的数据，符合洛特卡定律参数 n 等于 2 的仅占 14.6%，而符合洛特卡定律参数 n 不等于 2 的却占 70.8%（在鲍的论文中，实际上仅占 66.7%，有两组数据被误判，见第二章）。事实上，洛特卡本人使用的《化学文摘》的数据其洛特卡定律的参数 n 也不等于 2。因此可以说，当年洛特卡发现科学生产率频次分布的规律时，其分布参数 n 本来就不都等于 2。也许是先入为主，学术界对于参数 n 等于 2 的洛特卡定律钟爱有加，特别是当采集的数据经过拟合，当发现其不符合参数 n 不囿于 2 的洛特卡定律时，往往就退缩到参数为 2 的胡同里去，去寻找似乎能自圆其说的理由。

进入 20 世纪 90 年代，尽管偶然还能在学术刊物上看到参数为 2 的洛特卡定律的争鸣，但必竟是秋蝉之声了，学术界已普遍认为 n 等于 2 仅仅是洛特卡定律的特例，为区别起见，将参数不囿于 2 的洛特卡定律称为广义洛特卡定律。与此同时，参数 n 只取 2

的洛特卡定律也就被称为狭义的洛特卡定律。

20 世纪 70 年代弗拉奇(J. Vlachy)的研究表明，n 值在 1.2～3.5 之间，这也许是经验之谈，缺乏理性的依据。据 20 世纪 80 年代鲍的研究，29 种期刊的数据，n 值为 3.666 9，情报科学数据的 n 值高达 3.774 7，已突破了弗拉奇的范围了。

第二章　广义洛特卡定律参数的估计

如何估计广义洛特卡定律的参数？这个问题始终困扰着洛特卡定律普适性的研究。参数的估计，就其本质而言，就是运用一组科学生产率的数据，采用某种方法拟合洛特卡分布。拟合结果就确定了洛特卡分布的参数。

任何曲线的拟合，总希望拟合的曲线与经验数据（统计数据）所描绘的曲线越“贴切”越好。衡量拟合的曲线与经验数据描绘的曲线的“贴切”程度，可以用观察值 y_k 与估计值 $\hat{y}_k$ 之差，即残差 $\varepsilon_k = y_k - \hat{y}_k$ 来描述。任何一种曲线的最优拟合方法都是相对于由残差构成的优化标准而言的，如果一种曲线的拟合方法，符合某一种由残差构成的优化标准，那么就可以说这种拟合方法是最优的，由这种拟合方法所确定的参数也就是最优的。通常有以下三种标准：

（1）拟合的曲线使残差的绝对值最大最小化，即：

$$Q_1 = \min \max_k |\varepsilon_k|$$

（2）拟合的曲线使残差的绝对值和达到最小，即：

$$Q_2 = \min \sum_k |\varepsilon_k|$$

（3）拟合的曲线使残差平方和达到最小，即：

$$Q_3 = \min \sum_k \varepsilon_k^2$$

如果按照合理性和直观性程度依次排序，上述三种优化标准排列顺序恰好是（1）至（3）。

按拟合的优化标准，可将广义洛特卡定律的拟合方法，或广