1.

科学计量学与情报计量学中的排序问题

埃格赫(Dr. Leo Egghe)

一、引言

为了说明这个问题,首先让我们设定有一组期刊及其被引条数这样一个具体案例,也就是某一学科(例如由美国科学情报研究所所规定的某学科)及其被引条数。

《科学引文索引》/《期刊引用报告》(SCI/JCR)和《社会科学引文索引》/《期刊引用报告》(SSCI/JCR)将其所收录期刊按学科进行了分类。《期刊引用报告》的第四部分给出了同类期刊按影响因子的排序表。利用这些表计算美国科学情报研究所收录的某学科期刊的平均影响因子是不困难的。不过,了解整个该学科的影响因子,也就是总影响因子,是更重要的。的确,上述某学科总影响因子,与其诸子学科(其中必有一个子学科是活跃的)的平均影响因子作比较,乃是在对诸子学科作评估时最有代表性的指标之一[1],[2]。如果每种期刊发表的论文数一样,显然其平均影响因子与总影响因子是相同的。不过,在参考文献[3]中(也可见文献[4])我们已经说明,通常情况并非如此。事实上,按我们所能了解的情况,文献[4]是明确提请注意某子学科的平均影响因子与该学科总影响因子之间差别的第一篇论文。

本文后面,我们将在广泛的意义上使用"子领域"和"影响因子"这两个概念。在本文中,各个子领域不一定与《期刊引用报告》的子学科相同(当然,这是重要的应用之一),而影响因子不一定就是"公认的""加菲尔德影响因子"。它也可能是在文献[5],[6]意

义上的广义的影响因子。因此，我们将用“影响因子”（记作 I）这个术语来表示（在一个确定的时期内）引文数（记作 C）与相应的论文发表数（记作 P）的商。如果我们要强调引文数是发表数的函数，就记为 C(P)，类似地影响因子记为：

$$I(P)=\frac{C(P)}{P} \tag{1}$$

注意，C(P)是 P 的增函数。在所考虑的子领域内所有期刊的集合称为一个元期刊（meta-journal）[3],[5]。这个元期刊的第 i 种期刊的影响因子于是就用 I_i 来表示。由 N 种期刊组成的元期刊的平均影响因子，记作 AIF，定义为：

$$AIF=\frac{1}{N}\sum_{i=1}^{N}\frac{C_i}{P_i}=\frac{1}{N}\sum_{i=1}^{N}I_i \tag{2}$$

同一个元期刊的总影响因子，记作 GIF，则定义为：

$$GIF=\frac{\sum_{i=1}^{N}C_i}{\sum_{i=1}^{N}P_i}=\frac{\mu_C}{\mu_P} \tag{3}$$

其中 μ_C 是平均引文数：

$$\mu_C=\frac{\sum_{i=1}^{N}C_i}{N} \tag{4}$$

而 μ_P 是平均发表数：

$$\mu_P=\frac{\sum_{i=1}^{N}P_i}{N} \tag{5}$$

最后，比例 GIF/AIF 用 ρ 来表示：

$$\frac{GIF}{AIF}=\rho \tag{6}$$

二、ρ = GIF/AIF 与 I(P)对 P 的回归线斜率之间的关系

GIF 可能比 AIF 小，也可能比它大，由下面的定理可以证明。

[定理 II.1][8]如果用 r_P 表示 I(P)对 P 的回归线的斜率,则

$$r_P > 0 \Leftrightarrow \rho > 1 \tag{7}$$

而

$$r_P = 0 \Leftrightarrow \rho = 1 \tag{8}$$

证明:这里我们只证明(7)。

$$\rho > 1 \tag{9}$$

$$\Leftrightarrow \sum_{i=1}^{N} \frac{C_i}{P_i} < N \frac{\mu_C}{\mu_P} \tag{10}$$

另一方面,I(P)对 P 的回归线的斜率 r_P 大于 0

$$\Leftrightarrow N \sum_{i=1}^{N} P_i \cdot \frac{C_i}{P_i} - \left[\sum_{i=1}^{N} P_i\right]\left[\sum_{i=1}^{N} \frac{C_i}{P_i}\right] > 0 \tag{11}$$

(例如可见文献[7],p.66)

$$\Leftrightarrow \mu_C > \mu_P \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{C_i}{P_i} \tag{12}$$

$$\Leftrightarrow \sum_{i=1}^{N} \frac{C_i}{P_i} < N \cdot \frac{\mu_C}{\mu_P} \qquad \text{证毕} \tag{13}$$

[推论 II.1][8](i) 如果 I(P) 随 P 增加而递减,则 $\rho < 1$,

(ii) 如果 I(P) 随 P 增加而递增,则 $\rho > 1$。

无论是 GIF < AIF 还是反过来 GIF > AIF,这本身都不是一个坏性质,因为比较是对同一个组进行的。当然,如果上述的 GIF 与 AIF 之间的差别会导致不同组的排序颠倒,那就不好了。的确会发生这种情况,见文献[9]。作为第二个应用领域,我们将针对国家在作者统计(author - counts)方面的得分指标,来说明这一(非常不好的)性质。

三、有关作者发表论文数量的国家计量指标

对多作者的论文进行作者数计量或引文计量是一个非常复杂的问题。与此有关的另一个问题是从作者所得到的国家得分值的

确定。

人们可能相信，只要明确规定了他使用的是哪种计量方法，就足以对这个问题毫不含糊地进行研究或写出论文了。遗憾的是，并非如此。首先我们指出，在多作者论文情形中有四种不同类型的作者(国家，机构，……)指标计量方法。

(1)第一作者统计(First author counting)：只给一篇论文的 N 个作者(N = 1,2,3,……)中的第一个打分，给的分数为 1，参见[7]，[10]。该法有时称为简单统计(straight counting)。

(2)总量统计(Total counting)：N 个作者每个都得到 1 分，参见[11]。这种统计法也称为标准统计(normal 或 standard counting)，又参见[7]。

(3)分数统计(Fractional counting)：N 个作者各得 1/N 分，见[7]，[11]，[12]，[13]，[14]。这种统计法有时称为校正统计(adjusted counting)。

(4)比例统计(Proportional counting)：如果一个作者在一篇有 N 个作者的论文中排在第 R 位(R = 1,2,……,N)，那么他/她得到的分数为：

$$\frac{2}{N}\left[1-\frac{R}{N+1}\right] \tag{14}$$

在比例统计的情况下使用的这个计分公式，是将绝对权重 N + 1 - R 除以全部序号的和

$$1+2+\cdots+N=\frac{N(N+1)}{2}$$

得到的，参见[15]。

这里我们主要讨论第(2)，(3)和(4)三种统计法。因为我们觉得，第(1)种方法与其他统计步骤在研究方法上不一致，不过对于这个问题作一个随机模型也是有意思的。下面我们将不再考虑它。

我们对几个问题感兴趣。首先，我们想知道统计方法是如何影响作者(国家，……)的相对得分的。确实，只有相对得分是重要的：绝对得分是不可比的，因为各种统计方法所分配的权重的总和是不相同的。因此，我们要知道每个作者(国家，……)在某一统计法中的个别得分与该方法中给出的分数总和相除的商数。

我们将给出这些相对得分的公式。我们将用公式和实例证明，极有可能在一种计量方法中作者 a 比作者 b 有较高的分数，而在另一计量方法中，情况正好相反。其结果是，由于改变计量方法而使相对重要性(例如用得到的排序表示的)颠倒了。

不仅如此——这几乎是一个悖论——甚至有可能，当从一种计量法到另一种计量法，一个作者的相对得分有所增加的同时，他的排序也有增加。也就是说，根据他/她的相对得分，这个作者变得更重要了，而根据他/她的排序却变得更不重要了。

(一)各统计步骤的公式

首先我们规定一些通用的符号和术语。我们考虑 N 篇论文，对于每篇论文 i = 1，……，N 而言，用 a_i 表示论文 i 的合作者数目。设 c 是一个国家。那么论文 i 中来自 c 国的合作者数就用 $a_i(c)$ 表示，“国家”这个词必须做广义的理解：它可以是一个实际的国家名或是机构名，甚至可能是作者名。这时后者就是这种一般表达方式的特例，也就是 $a_i(c) = 0$ 或 1 的情况。一般说来，对于一个真实国家或机构，$a_i(c)$ 可能(原则上)是任何自然数(包括零)。这就是为什么我们要用这种更普遍的框架。显然，$\forall i(i = 1, \cdots\cdots, N)$：

$$a_i = \sum_c a_i(c) \tag{15}$$

. 总量统计

每个作者得到 1 分，因此这个系统中的总分数为

$$W_2 = \sum_{i=1}^{N} a_i \tag{16}$$

按照符号规定，国家 c 的总分数是

$$W_2(c)=\sum_{i=1}^{N}a_i(c) \tag{17}$$

因此,其相对指标是

$$Q_2(c)=\frac{W_2(c)}{W_2}=\frac{\sum_{i=1}^{N}a_i(c)}{\sum_{i=1}^{N}a_i} \tag{18}$$

注意,(18)相当于 GIF(公式(3))。

2. 分数统计

由于一篇论文的总分数是 1,我们得到这一系统中的总分数为 $W_3=N$。论文 i 中每个合作者得到 $1/a_i$ 分,因此国家 c 在论文 i 中得到的绝对分数是

$$\frac{a_i(c)}{a_i} \tag{19}$$

(有来自国家 c 的 $a_i(c)$个合作者)。

国家 c 的总分数是

$$W_3(c)=\sum_{i=1}^{N}\frac{a_i(c)}{a_i} \tag{20}$$

因此相对分数为

$$Q_3(c)=\frac{1}{N}\sum_{i=1}^{N}\frac{a_i(c)}{a_i} \tag{21}$$

注意,(21)相当于 AIF(公式(2))。

3. 比例统计

我们不再去推导比例统计的公式,因为它非常复杂,而且后面我们也不再用到它。我们可以利用下面的定理(可在文献[9]中找到)得到关于 $Q_4(c)$(利用比例统计得到的国家 c 的相对指标)的结论。

[定理 III.1]对于若干国家(机构,作者,……)之间的任何合作系统,我们可以构造出另一个系统,使得后者的比例统计体系等

于前者的分数统计体系。而且两个系统中的总量统计体系也是一样的。

只要有了 $Q_2(c)$与 $Q_3(c)$比较的结果，这个定理就可使我们得到关于 $Q_2(c)$与 $Q_4(c)$比较的结果。因此，我们只需要讨论总量统计与分数统计之间的比较，无需另外的工作即可得到关于总量统计与比例统计之间比较的一个类比结果。

(二)总量与分数统计指标之间以及总量与比例统计指标之间异常的重要实例

我们感兴趣的是，看看有没有这样的例子，其中(c，c′：国家，作者，机构，……)：

$$Q_2(c) > Q_2(c') \tag{22}$$

$$Q_3(c) < Q_3(c') \tag{23}$$

这将是这两种计量方法意义含糊不清的第一个标志：在总量统计系统中 c 比 c′重要(由于我们讨论的是相对指标：c 在绝对指标中占的分数比 c′要大)；而在分数统计系统中情况恰好相反。

如果这个例子还能给出

$$Q_2(c) < Q_3(c) \tag{24}$$

那就是一个完全的悖论了。

这时(24)表明 c 在总量统计系统中的权重比在分数统计系统中要小，但这个较小的权重却比相应统计系统中较大的权重有更高的重要性(与 c′比较)，因为有(22)和(23)！

如果能够给出一个(22)，(23)和(24)同时发生的实例，那将是对所使用方法的确定性打击，并且会使许多据此得出结论的研究工作成为(至少是)可疑的。在这节中我们证明，这样的实例确实存在，尽管是对于相对复杂的(论文总数的)系统。不过我们能够在 3 个作者(或与此相当地，我们甚至可以假定 3 个国家，对于每个 $i = 1, \cdots\cdots, N$ 都有 $a_i(c) = 1$)的简单情形中给出例证。根据前

一节中 $Q_2(c)$与 $Q_3(c)$的公式，并根据(22)，(23)和(24)，我们可以构造出这样的例子来。用 a，b，c 表示作者(或国家)，各横行表示出版物。

[例 1]

	a	b	c
1		x	
2		x	
3	x	x	
4	x	x	
5	x	x	
6	x	x	
7	x	x	
8	x	x	
9	x	x	
10	x	x	
11	x	x	
12	x	x	
13	x	x	
14	x	x	
15	x	x	
16	x		x
17	x		x
18	x		x
19	x	x	x
20	x	x	x
21	x	x	x
22	x	x	x
23	x	x	x
24	x	x	x
25	x	x	x
26	x	x	x
27	x	x	x
28	x	x	x
29	x	x	x
30	x	x	x
31	x	x	x

这导致下列的 Q_2 和 Q_3 值：

Q_2	Q_3
$Q_2(a) = 0.3973$	$Q_3(b) = 0.4140$
$Q_2(b) = 0.3836$	$Q_3(a) = 0.3978$
$Q_2(c) = 0.2192$	$Q_3(c) = 0.1882$

注意，正如所要求的那样 $Q_2(a) > Q_2(b)$，$Q_3(a) < Q_3(b)$，和 $Q_3(a) > Q_2(a)$。

现在我们来看，是否存在一个 a、b 和 c 未一起合作的例子。我们有下面的解。

	a	b	c
1	x		
2	x		
3	x		
4	x		
5	x		
6		x	
7		x	
8		x	
9		x	
10		x	
11		x	
12		x	
13			x
14	x	x	
15	x	x	
16		x	x
17		x	x
18		x	x
19		x	x
20		x	x
21	x		x
22	x		x
23	x		x
24	x		x
25	x		x
26	x		x
27	x		x
28	x		x

现在我们有下列的 Q_2 和 Q_3 值：

Q_2	Q_3
$Q_2(a)=0.3488$	$Q_3(b)=0.3750$
$Q_2(b)=0.3256$	$Q_3(a)=0.3571$
$Q_2(c)=0.3256$	$Q_3(c)=0.2679$

注意，再次出现 $Q_2(a)>Q_2(b)$，$Q_3(a)<Q_3(b)$，和 $Q_3(a)>Q_2(a)$。

四、排序异常的一种解决方法

根据作者在文献[16]中的结果，我们有，Q_2 和 Q_3 的几何“形式”（用 Q_2^g 和 Q_3^g 表示）：

根据(18)，Q_2 是

$$Q_2(c)=\frac{\{a_1(c),\cdots,a_N(c)\}\text{的算术平均}}{\{a_1,\cdots,a_N\}\text{的算术平均}}$$

因此

$$Q_2^g(c)=\frac{(a_1(c)a_2(c)\cdots a_N(c))^{\frac{1}{N}}}{(a_1a_2\cdots a_N)^{\frac{1}{N}}} \tag{25}$$

根据(21)，Q_3 是

$$Q_3(c)=\left\{\frac{a_1(c)}{a_1},\cdots,\frac{a_N(c)}{a_N}\right\}\text{的算术平均}$$

因此

$$Q_3^g=\left[\frac{a_1(c)}{a_1}\cdot\frac{a_2(c)}{a_2}\cdot\cdots\cdot\frac{a_N(c)}{a_N}\right]^{\frac{1}{N}} \tag{26}$$

现在从(25)和(26)可清楚看出

$$Q_2^g(c)=Q_3^g(c) \tag{27}$$

对于任何系统和任何 c 都成立。进而言之，如果是这样，显然总量统计系统中的所有排序与分数统计系统中的都一样了。这样所有模糊性就都不存在了。

我们留下 $Q_4^5(c)$的研究，也就是比例统计系统在这种情况下特性的研究，作为一个待解决的问题。

本文中讨论的这些问题有许多应用，不仅与引文评价或发表物评价有关。事实上，我们甚至可以给出情报计量学和科学计量学范围之外的实例。

五、其它应用

(一)普赖斯指数

对于某篇论文，统计参考文献的总数和出版年龄不老于 d 年的参考文献数，其中出版那一年计为第一年。(普赖斯[17]用的是 d =5)。二者的商就是普赖斯指数。莫德[18]利用平均的普赖斯指数作为对一个领域的计量，而普赖斯本人则使用总体指数。当然，同样可以对于一种期刊计算普赖斯指数。沃特斯和莱德道夫[19]讨论了两种方法的差别。他们注意到，期刊《科学计量学(Scientometrics)》五年平均的普赖斯指数是 0.514，而总体指数是 0.43。从总体指数小于平均指数这一事实，我们得出的结论是，普赖斯指数相对于参考文献总数的回归曲线的斜率是下降的。在参考文献[20]中对普赖斯指数有进一步的研究。

(二)正文/参考文献比

在文献[21]中，随机地选择了一些期刊，对于 1980 年到 1987 年期间发表的每篇论文，得到了(估计的)字数和参考文献数。其商给出了每篇论文的正文/参考文献比。据我们所能看到的，作者未说明他们是如何获得这些期刊数据的；也就是说，究竟是论文数据的平均值还是总体的正文与参考文献之比。我们假定他们使用的是总体方法。

(三)吸收因子(receptivity factor)

在这一应用中，学科领域是固定的。对于所研究的每篇论文，统计其参考文献总数和其中本国的人所写的文章数。逐个国家都

这样做,这里同样可能采取平均的或是总体的观点。将这一结果除以本领域总产出中该国家的份额,就可得到外国文献的吸收因子(参见[22])。

(四)期刊价格

对于每种期刊,统计其(一年)出版的页数和定购价格。商就是每页的价格。类似地可计算每个字符的价格或每条引文的价格,作为一类"金钱价值"(Value for money)的指标。然后可以把结果按领域或按出版社归纳在一起。对于某些出版社而言,这似乎是一种非常可争议的方法[24,25]!

在文献[4]中计算了每篇论文的价格。这些作者提到的是一个学科或组织中每篇论文的加权的,即总体的价格,他们对影响因子的情况也是这样做的。

(五)老化

对于每篇论文,统计 c_j,即发表年龄是 j 年的参考论文数,$j = 0,1,\cdots\cdots,10$。这里用数 10 是为了方便;我们进一步假设所有 c_j 均不为 0。于是一篇论文的老化率就确定为商 c_{j+1}/c_j 的平均值,$j = 0,\cdots\cdots,9$)。请注意文献中还有老化的另外一些定义。上述定义只是用作一个例子。

然后,为了确定一种期刊的老化率,可以使用所有论文老化率的平均值(AAR:期刊的平均老化率),也可以用一种总体的方法,即取所有 c_j 的和,求出商数,然后取所有商数的平均值(GAR:期刊的总体老化率)。

(六)地区生产总值(GRP)

如本文中所讨论的,在许多领域都会出现平均值。一个出自经济计量学的众所周知的例子是,对每个国家统计其国民生产总值(GNP)和居民的数目。它的商就得出人均的 GNP。例如当讨论欧盟的人均 GRP 时,可以取每个成员国人均 GNP 的平均,也可以计算总体的人均 GRP。同样,我们认为第二个指标更有意义。

关于其他应用(学科影响指标,表示图书馆开馆状况的满座率),建议参见[16]。

参 考 文 献

[1] R.E. De Bruin, H.F. Moed and E. Spruyt. Antwerpse analyses. Rapport ten behoeve van de bestuursorganen van de Universiteit Antwerpen, 1993.

[2] R. Rousseau. A scientometric study of the scientific publications of LUC. Report, 1995.

[3] R. Rousseau and G. Van Hooydonk. Journal production and journal impact factors. Journal of the American Society for Information Science, 47(10), p.775 - 780, 1996.

[4] G. Van Hooydonk, R. Gevaert, G. Milis - Proost, H. Van De Sompel and K. Debackere. A bibliotheconomic analysis of the impact factors of scientific disciplines. Scientometrics, 30, p.65 - 81, 1994.

[5] R. Rousseau. Citation distribution of pure mathematics journals. In : Informetrics 87/88. L. Egghe, R. Rousseau, (Eds.), Amsterdam, Elsevier, p.249 - 262, 1988.

[6] R. Rousseau. A note on maximum impact factors. In : Information as a Global Commodity : Communication, Processing and Use, CAIS/ACSI '93, 21st Annual Conference, 11 - 14 July 1993, p.120 - 125, 1993.

[7] L. Egghe and R Rousseau. Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science, Elsevier, Amsterdam, 1990.

[8] L. Egghe and R. Rousseau. Average and global impact of a set of journals. Scientometrics, 36, p.97 - 107, 1996.

[9] L. Egghe, R. Rousseau and G. Van Hooydonk. The behavior of scores obtained through total, fractional or proportional counting. Preprint 1998.

[10] J.R. Cole and S. Cole. Social stratification in Science. The University Press, Chicago, 1973.

[11] D. Lindsey. Production and citation measures in the sociology of science : the problem of multiple authorship. Social Studies of Science, 10, p. 145 - 162, 1980.

[12] D.J. De Solla Price. Letter to the Editor. Science, 212, p.987, 1981.

[13] Q. Burrell and R. Rousseau. Fractional counts for authorship attribution : a numerical study. Journal of the American Society for Information Science, 46, p.97 - 102, 1995.

[14]L. Egghe. Source – item production laws for the case that items have multiple sources with fractional counting of credits. Journal of the American Society for Information Science, 47, p. 730 – 748, 1996.

[15]G. Van Hooydonck. Fractional counting of multi – authored publications consequences for the impact of authors. Journal of the American Society for Information Science, 48, p.944 – 945, 1997.

[16]L. Egghe and R. Rousseau. Averaging and globalising quotients of informetric and scientometric data. Journal of information Science, 22, p.165 – 170, 1996.

[17]D. De Solla Price. Citation measures of hard science, soft science, technology, and nonscience. In : C.E. Nelson and D.K. Pollack (eds.), Communication Among Scientists and Engineers, Heath, Lexington, MA, p.3 – 22, 1970.

[18]H.D. Moed. Bibliometric measurement of research performance and Price's theory of differences among the sciences. Scientometrics, 15, p.473 – 483, 1989.

[19]P. Wouters and L. Leydesdorff. Has Price's dream come true : is scientometrics a hard science? Scientometrics, 31, p.193 – 222, 1994.

[20]L. Egghe. The Price Index and Its relation to the mean and median reference age. Journal of the American Society for Information Science, 48, p. 564 – 573, 1997.

[21) A.E. Little, R.M. Harris and P.T. Nicholls. Text to reference ratios in scientific journals. In : L. Egghe and R. Rousseau (eds.), Informetrics 89/90, Elsevier, Amsterdam, p. 211 – 216, 1990.

[22]I.L. Herman. Receptivity to foreign literature : a comparison of UK and US citing behavior in librarianship and information science. Library and Information Science Research, 13, p.37 – 47, 1991.

[23]H.H. Barschall and J.R. Arrington. Cost of physics journals : a survey. Bulletin of the American Physical Society, 33, p.1437 – 1447, 1988.

[24]C. Holden. Gordon and Breach impanels a journal jury. Science, 249, p.298 – 299, 1990.

[25]A.L. O'Neill. The Gordon and Breach litigation : a chronology and summary. Library Resources and Technical Services, 37, p. 127 – 133, 1993.

（杨虹 译　王存诚 校）