关于综合运用 Benford 法则和面板模型检测统计数据质量的研究*

刘云霞　吴曦明　曾五一

内容提要：本文介绍了如何利用 Benford 法则来检测统计数据质量的一般方法。在此基础上，进一步探讨了如何将其与面板模型相结合从而找出可能存在质量问题的具体地区和时间序列数据的方法。并利用上述方法对我国多个国家级开发区的主要经济指标的数据质量进行了实证分析。

关键词：数据质量；Benford 法则；面板模型
中图分类号：O212　　文献标识码：A　　文章编号：1002－4565(2012)11－0074－05

Detecting Statistical Data Anormality by Combining Benford's Law and Panel Data Models

Liu Yunxia　Wu Ximing　Zeng Wuyi

Abstract: This article describes a general method that can detect statistical data abnormalityby Benford's law. In addition, the article discusses how to combine Benford's Law with panel data models to identify the observations that may have data quality problems. We demonstrate the applicability of the proposed method method with an examination on major economic indicators of Chinese national development zones.

Key words: Data Quality; Benford's Law; Panel Model

数据质量是统计工作的生命线。近年来，我国统计数据的质量问题已成为各级政府和社会各界关注的热点。如何利用科学的方法来诊断统计数据的质量，也成为统计学界重点探讨和研究的一项课题。

本文拟对如何利用 Benford 法则来检测统计数据质量的方法做一些介绍，在此基础上，进一步探讨如何将其与面板模型相结合，进一步找出可能存在质量问题的具体地区和时间数据的方法，并利用我国国家级开发区有关经济指标的数据开展实证分析，验证该方法的适用性。

一、Benford 法则

Benford 法则是由美国数学家、天文学家 Simon Neweomb 在 1881 年首次发现的。在 1851 年的一天，他在使用对数表做计算时，注意到对数表的第一页要比其他页更为破旧。奇怪的现象激发了他的研究兴趣，经过大量的统计分析，他发现许多类型的数字都很好地符合这样的规律：以 1 为第一位数的随机数要比以 2 为第一位数的随机数出现的频率要大，而以 2 为第一位数的随机数又比以 3 为第一位数的随机数出现的概率要大，并可以此类推。当时 Simon Neweomb 关注这一数学现象完全是出于好奇，并没有对这一规律做出解释。

到了 1938 年，美国通用电器(GE)的物理学家 Frank Benford 注意到了同样的现象。他收集并验证了 20229 个数字，其中包括篮球比赛的数字、河流的长度、湖泊的面积、各个城市的人口分布数字、在某一杂志里出现的所有数字，发现在这些数字中，整数 1 在数字中第一位出现的概率大约为 30%，整数 2 在数字中第一位出现的概率大约为 17%，整数 3 在数字第一位出现的概率约为 12%，而 8 和 9 在数字中第一位出现的概率约为 5% 和 4%。经过研究后，

* 本文为国家社科基金重点项目“国家统计数据质量管理问题研究(09AZD045)”阶段成果之一；同时获得中央高校基本科研业务费专项资金资助(0140zk1008)。

Frank Benford 得出这样一个结论: 大量自然数据的首位数字的出现频率符合这个规律,这就是 Benford 法则[1]。Benford 法则主张: 在不同种类的统计数字中,首位数字是数字 d_1 的概率是:

$$P(\textit{First digit is } d_1) = \log_{10}(1 + (1/d_1)) \quad (1)$$

其中,首位数字是指左边的第一位非零的有效数字。根据公式(1),Benford 法则中首位数分别出现 1 ~ 9 的概率如下表:

表 1　Benford 法则中首位数的概率分布

首位数	1	2	3	4	5
概率	0.3010	0.1761	0.1249	0.0969	0.0792
首位数	6	7	8	9	
概率	0.0669	0.0580	0.0512	0.0458	

Benford 法则提出之后引起了人们的极大关注。1996 年美国学者 Hill 从理论上对 Benford 法则给出了满意的解释,并进行了严谨的数学证明[2]。后来有学者根据公式(1),还推导出了第二位数为 d_2 以及第三位数为 d_3 的概率,并且这种对数规律可以类推至第四位、第五位数出现的概率[3]。

$$P(\textit{Second digit is } d_2) = \sum_{d_1=1}^{9} \log_{10}\left(1 + \left(\frac{1}{d_1 \cdot d_2}\right)\right) \quad (2)$$

$$P(\textit{Third digit is } d_3) = \sum_{d_1=1}^{9}\sum_{d_2=1}^{9} \log_{10}\left(1 + \left(\frac{1}{d_1 \cdot d_2 \cdot d_3}\right)\right) \quad (3)$$

$$P(\textit{Fourth digit is } d_4) = \sum_{d_1=1}^{9}\sum_{d_2=1}^{9}\sum_{d_3=1}^{9} \log_{10}\left(1 + \left(\frac{1}{d_1 \cdot d_2 \cdot d_3 \cdot d_4}\right)\right) \quad (4)$$

根据 Benford 法则,高质量的数据首位数字的出现应该遵循上述概率,并且数据规模越大,数据首位数字的概率分布就越应该符合 Benford 法则。如果存在弄虚作假或者拼凑、修饰数据的行为,这种规律有可能被破坏。因此,如果一组统计数据的首位数字的概率分布与 Benford 法则下的首位数字概率分布存在差异时,该数据的准确性就值得怀疑了。

也正是因为这个特点,国内外的税务、会计和审计领域都已经将此法则作为检测数据是否有修饰、篡改、舞弊的方法之一。例如,Mark J. Nigrini (1992)[4] 提出该法则可用于检查是否有伪账,并且可以推而广之用于会计、金融甚至选举中出现的数据检测; 张苏彤(2005)[5]、王福胜等(2007)[6] 将该法则用作舞弊审计的分析方法; 狄为等(2010)[7] 将该法则用于会计舞弊的发现研究; 在统计领域,也有学者将此法则用于检验数据的准确性。如 George Judge 等(2009)[1] 将此法则用于检测调查数据的质量; 许涤龙、金瑛(2010)[3] 将该法则用于对 M2 统计数据准确性的研究。

二、Benford 法则的检验方法

目前有四种方法可以检验一个统计数据集首位数字的概率分布是否服从 Benford 法则的分布。

(一) χ^2 拟合优度检验

通过 χ^2 拟合优度检验,可以检测统计数据中首位数的频率分布是否与 Benford 法则下的分布有显著差异。χ^2 统计量为:

$$\chi^2 = N \cdot \sum_{i=1}^{9} [e_i - b_i]^2 / (b_i) \quad (5)$$

其中,e_i 是统计数据中首位(第二位或者第三位)出现数字 i 的实际频率,b_i 是 Benford 法则下首位(第二位或者第三位)出现数字 i 的理论频率。显著性水平分别为 10%、5% 和 1% 时,χ^2 检验的临界值分别是 13.36、15.51 和 20.09。如果 χ^2 统计量的值大于临界值,则接受备择假设,说明统计数据首位数字的频率分布不符合 Benford 分布,即说明该数据可能存在质量问题,应引起注意。

(二) 修正 Kolmogorov-Smirnov 拟合优度检验

K-S 检验是用来检验单一样本是否来自某一特定理论分布的方法。它是以样本数据的累积分布函数与特定理论分布的累积分布函数作比较,求这两个累积分布函数的差的绝对值中的最大值 D。然后,通过查表以确定 D 值是否落在所要求对应的置信区间内。若 D 值大于临界值,说明被检测的数据不服从这一特定理论分布。K-S 拟合优度检验的统计量为:

$$D = \max |F_e(x) - F_b(x)| \quad (6)$$

其中,$F_e(x)$ 是实际的统计数据中首位数的累积分布函数,$F_b(x)$ 是理论分布即 Benford 法则下首位数的累积分布函数。

Kuiper 对 K-S 拟合优度检验作了修正[8],得到如下统计量:

$$V_N = \max [F_e(x) - F_b(x)] + \max [F_b(x) - F_e(x)] \quad (7)$$

Stephens 对公式(7)的统计量再作修正[9],有:

$$V_N^* = V_N [N^{1/2} + 0.155 + 0.24N^{-1/2}] \quad (8)$$

该拟合优度检验在 10%、5% 以及 1% 显著性水平下的临界值分别为 1.19、1.32 和 1.58。

(三) 距离检测[9]

通过计算统计数据首位数字的频率分布与 Benford 分布之间的距离,可以检测该数列是否符合 Benford 法则。这样的距离有:

$$m = \max_{i=1,2,\cdots\cdots,9}\{|b_i - e_i|\} \quad (9)$$

$$d = \sqrt{\{\sum_{i=1}^{9}(b_i - e_i)^2\}} \quad (10)$$

(四) Pearson 相关系数

通过计算统计数据中首位数字的频率分布与 Benford 法则下首位数字的频率分布的 Person 相关系数,也可以判断两个分布是否有差异,其判断标准见表 2。

表 2 根据相关系数进行判断的分级标准[10]

分级	相关系数分级标准	说 明
正常	$0.99 < r \leqslant 1$	完全符合 Benford 法则
关注	$0.97 < r \leqslant 0.99$	存在一定程度篡改数据的可能性
可疑	$r \leqslant 0.97$	有篡改数据的迹象,需特别注意

三、Benford 法则和面板模型的综合

虽然 Benford 法则在数据质量的诊断中已经得到不少运用,但是应当指出其仍然存在不少有待进一步研究改进的问题。

第一,并不是所有的数据样本都一定服从 Benford 法则。能够用 Benford 法则来进行分析的数据应该符合以下条件:①数值既不是完全随机的,也不能过度集中于某个区间;②数值不能存在上下限;③数值在一个很宽的范围里连续变动,不存在间断点或间断区间;④数字没有被特别赋值;⑤数值的形成受多种因素的影响,是多种因素综合作用的结果。

第二,就 Benford 法则本身来说,如果数据检测结果符合 Benford 法则的频率分布,也并不意味着一定不存在数据质量问题。因为当数据总量非常大的时候,并且有质量问题的数据发生次数不多时,它们就会淹没在大样本的规律之中,而不能被发现。

第三,在现实中,人们更希望了解的不仅是何类统计数据可能存在质量问题,而是哪一个单位、哪一个时间的数据可能存在问题。

对于上述问题,我们提出以下进一步完善的思路:

首先,利用 Benford 法则检验何种统计指标有可能存在质量问题。

其次,利用面板模型对上述可能存在质量问题的统计指标作进一步分析。

最后,检查面板模型诊断发现的异常点的数据的首位数与 Benford 法则检验中发现存在的出现频率偏大的首位数是否相同,如果相同则可有较大的把握判断该异常点的数据确实存在质量问题。如果不同,则可以认为尽管存在异常点,但这种异常可能并非由于数据质量造成的。

以上将 Benford 法则和面板模型加以综合运用的方式,不仅可以解决单纯的 Benford 法则检验无法判断具体样本点的数据是否存在质量问题的难点,而且还可弥补单纯利用面板模型诊断数据质量方法的不足。利用面板模型诊断统计数据质量的基本思想是:任何一种统计指标与其相关的一组(或一项)指标之间的关系,都可以用面板模型来近似反映。如果回归估计的结果,整体模型拟合得很好,仅有个别数据严重偏离既定模型,则可以认为处在这些点(奇异点)上数据的准确性可能存在问题,有必要作进一步的观察与分析。利用面板模型诊断统计数据质量的最大难点在于:当诊断结果出现异常时,实际上难以判断这一异常是由于数据质量引起的,或是该点的实际情况并不符合所选用的模型引起的。Benford 法则和面板模型的综合运用可以从另一个侧面找出可能存在质量问题的数据,从而明显提高了统计诊断结论的可靠性。

四、实证分析

(一) 数据来源

实证分析采用的数据来源于两个方面:2002 – 2008 年的数据来源于 2003 – 2009 年版的《中国开发区年鉴》;2009 – 2010 年的数据来源于中国开发区网站的统计公报(http://www.cadz.org.cn/)。在上述资料来源中,各开发区公布的指标不尽相同。因此我们选取了各开发区都发布的地区生产总值、工业总产值(现价)、工业增加值(现价)、税收收入、出口总额、进口总额六个重要的经济指标作为分析的对象。另外,我国国家级开发区在 2002 – 2008 年之间为 54 个,2009 年以后扩大为 90 个,考虑到各年数据的一致性,这里我们只采用 2002 – 2010 年均有数据的 54 个开发区作为研究对象。

表 3　　各指标首位数字的频率分布

首位数字	Obs	1	2	3	4	5	6	7	8	9
Benford Law		30. 103	17. 609	12. 494	9. 691	7. 918	6. 695	5. 799	5. 1151	4. 576
地区生产总值	476	31. 513	15. 126	12. 395	9. 034	9. 664	5. 462	6. 723	4. 832	5. 252
工业总产值	467	29. 764	18. 415	14. 989	10. 493	6. 638	6. 852	4. 711	5. 567	2. 57
工业增加值	462	30. 952	17. 316	11. 255	9. 74	7. 576	7. 143	5. 628	5. 628	4. 762
税收收入	463	31. 965	15. 983	12. 527	6. 695	6. 695	6. 479	8. 639	5. 616	5. 4
出口总额	467	30. 835	14. 989	11. 991	10. 707	10. 493	7. 709	6. 21	4. 069	2. 998
进口总额	461	31. 67	19. 306	11. 497	7. 592	9. 111	6. 941	4. 555	5. 423	3. 905

(二) 6 个指标首位数字的频率分布及 Benford 分布的检验

由于 Benford 法则具有样本量越大,效果越明显的特点,所以我们将这 6 个指标 9 年的数据合在一起来观察它们的首位数字的频率分布,从而更好地验证它们是否符合 Benford 法则。表 3 是 6 个指标 9 年数据首位数字出现的频率分布表。

从表 3 可以看出,各指标首位数字的频率分布与 Benford 法则的频率分布有一定差别。但这种差别是否显著还需要进行一定的检验。我们根据公式 (5) ~ (10),计算有关统计量,用来检验各指标数据首位数字的频率分布是否符合 Benford 法则,计算结果见表 4。

表 4　2002 – 2010 年各指标首位数字频数分布与 Benford 分布的拟合优度检验

	r	χ^2	V_N^*	m	d
地区生产总值	0. 9871	6. 3596	0. 7121	0. 0248	0. 0368
工业总产值	0. 9883	9. 0590	0. 8935	0. 025	0. 037
工业增加值	0. 9975	1. 2042	0. 3952	0. 0124	0. 0167
税收收入	0. 9791	12. 9646	1. 3068*	0. 03	0. 0485
出口总额	0. 9818	10. 7981	1. 092	0. 0262	0. 0433
进口总额	0. 9921	6. 2312	0. 7062	0. 021	0. 0365

注: 表中带* 的数据表示大于显著性水平 10% 的临界值。

表 4 的数据表明,6 个指标的相关系数中,除"税收收入"为 0. 9791 外,其他 5 个指标都在 0. 99 左右。从 χ^2 统计量来看,6 个指标的 χ^2 值都小于 10% 显著性水平的临界值,"税收收入"的 χ^2 值是其中最大的; 从 V_N^* 检验来看,只有"税收收入"的 V_N^* 统计量值大于显著性水平 10% 的 V_N^* 临界值; 另外,"税收收入"的 m 值和 d 值在 6 个指标中也是最大的。所以,虽然各种拟合优度检验的结果不大一样,但可以推断出"税收收入"这个指标的数据可能存在一定的质量问题。从其首位数分布情况看,该指标首位数为 1、7、8 及 9 的数据的频率分布比 Benford 法则的频率分布要大,这说明,出现质量问题的数据很有可能就出现在首位数为 1、7、8 及 9 的数据中。因此审查数据时,对那些首位数为 1、7、8 及 9 的税收收入数据的开发区应多加考察。

(三) 建立面板数据模型

为了充分利用 54 个国家级开发区在不同时间上的数据信息,我们将通过面板模型来进一步诊断哪些开发区在哪些年份的"税收收入"指标可能存在数据质量问题。

在模型的建立中,考虑到各国家级开发区的具体情况如产业结构、税收优惠政策等不同,因而导致不同开发区的"税收收入"指标与其他指标之间的关系也存在差异。为了体现这种差异,我们采用面板数据的变系数模型来对现有数据进行拟合。另外,由于地区生产总值、出口总额、进口总额、工业总产值、工业增加值等 5 个指标之间相关性比较高,如果都加入模型作为自变量将存在多重共线性。为降低多重共线性的影响,本文经过筛选,选取了地区生产总值作为模型的解释变量,税收收入作为被解释变量。从本文主要目的是筛选数据质量存在问题的开发区这一角度来看,这种处理方法是合适的。具体模型为:

$$TAX_{it} = \alpha + \beta_i GDP_{it} + u_{it} \qquad (11)$$

其中,$i = 1, 2, \cdots, 54$; $t = 1, 2, \cdots, 9$。TAX_{it} 为各个开发区在某一年的税收收入,GDP_{it} 为各开发区在某一年的生产总值,α 是模型的截距项,β_i 为斜率系数,其随开发区的不同而不同,u_{it} 是随机误差项。

我们利用广义最小二乘法对上述模型进行了估计。从检验结果可以看出,面板数据模型中的截距项和各开发区的斜率系数非常显著,同时调整后的 R^2 达到了 0. 976,模型整体拟合效果较好。这为我们下一步分析提供了较好的基础。

(四)根据残差分析查找税收收入异常的开发区

在面板数据模型的结果中,利用残差所提供的信息可以对数据的质量进行诊断。一般情况下,如果模型可靠,则残差特别大的样本点数据出现质量问题的可能性较大。

根据 2002 – 2010 年的残差数据,我们计算了每个开发区每年残差的标准化数值,即:

$$z_{ij} = \frac{|x_{ij} - \bar{x}_i|}{\sigma_i} \quad (12)$$

其中,x_{ij}是第 i 年第 j 个开发区的残差,$\bar{x}_i$ 是第 i 年所有开发区残差的平均值,σ_i 是第 i 年所有开发区残差的标准差。如果某个开发区的 z_{ij} 值大于 2,就可以认为该开发区税收收入数据很可能是异常数据。据此,我们发现 9 年间各开发区共有 28 个异常数据。观察这些异常数据的首位数分布,结合前述 Benford 法则的分析结果(即首位数为 1、7、8 及 9 的数据可能存在质量问题),我们发现面板模型检验发现的 28 个异常数据中有 18 个数据同时也是 Benford 法则诊断可能存在问题的数据。对这些开发区这些年份的税收数据有必要做进一步的检查,查出其可能存在问题的原因。

这 18 个数据主要集中在 9 个开发区,即与其他开发区相比,这 9 个开发区 GDP 的回归系数明显较高。将各开发区 GDP 的回归系数从高到低排序之后发现,排名前 6 位的开发区中有 5 个开发区属于税收数据可能存在问题的开发区。对此,一个可能的解释是:由于各个开发区所处的地区不同,区内企业类型也不同,导致开发区之间 GDP 与税收的关系本来就存在差异。相比平均税负比较低的开发区而言,平均税负较高的开发区人为调低税收的冲动更大,这就导致这些开发区在某些时期上报的税收收入可能会低于其应有的真实水平。

参考文献

[1] George Judge, Laura Schechter. Detecting Problems in Survey Data Using Benford's Law [J]. The Jounal of Human Resources, 2009, 44: 1 – 24.

[2] Hill T. P. A Statistical Derivation of the Significant-Digit Law [J]. Stat. Sci, 1996, 10: 354 – 363.

[3] 许涤龙,金瑛. 基于 Benford 法则的 M2 统计数据准确性研究 [J]. 统计与信息论坛,2010(8).

[4] Mark J. Nigrini. The Detection of Income Tax Evasion Through an Analysis of Digital Frequencies [D]. Ph. D. thesis. Cincinnati, University of Cincinnati, 1992.

[5] 张苏彤. 奔福德定律:一种舞弊审计的数值分析方法 [J]. 中国注册会计师,2005(11).

[6] 王福胜,李勋,孙逊. 奔福德定律及其在审计中的应用研究 [J]. 财会通讯,2007(3).

[7] 狄为,施鹏仙. 基于 Benford 定律的会计舞弊发现研究 [J]. 会计之友,2010(9).

[8] Giles, David E. Benford's Law and Naturally Occurring Prices in Certain EBay Auctions [J]. Applied Economics Letters, 2007, 14(3): 157 – 61.

[9] Stephens, Michael A. Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics without Extensive Tables [J]. Journal of the Royal Statistical Society, Series B, 1970, 32(1): 115 – 22.

[10] 许存兴,王大江,张芙蓉. 上市公司审计意见实证分析 – 基于 Benford 法则的造假检测 [J]. 南京财经大学学报,2009(4).

作者简介

刘云霞,女,34 岁,山西省人,厦门大学经济学院统计系助理教授,硕士生导师。研究方向为统计分析与数据挖掘。

吴曦明,男,37 岁,厦门大学经济学院统计系讲座教授、美国得克萨斯农机大学农业经济学系副教授,(美)《美国农业经济》杂志副主编。研究方向为计量经济学、金融计量、宏观经济、劳动经济。

曾五一,男,59 岁,福建省人,厦门大学经济学院统计系教授、博士生导师,中国统计学会顾问、教育部统计学教学指导分委员会副主任委员、国家统计局咨询委员。研究方向为国民经济统计、统计理论与方法。

(责任编辑:程 晞)