Classical Explanation And Application Of Benford's law

张子杰, 数学系,161110120
Zijie Zhang, Department of Mathematics

November 2017

1 Introduction

本福特定律，也称为本福特法则，说明一堆从实际生活得出的数据中，以 1 为首位数字的数的出现机率约为总数的三成，接近直觉得出之期望值 1/9 的 3 倍。推广来说，越大的数，以它为首几位的数出现的机率就越低。它可用于检查各种数据是否有造假。
在 wikipedia 找到的不完整的解释：

一组平均增长的数据开始时，增长得较慢，由最初的数字增长到另一个数字起首的数的时间，必然比起首的数增长到，需要更多时间，所以出现率就更高了。

从数数目来说，顺序从 1 开始数，1,2,3,...,9，从这点终结的话，所有数起首的机会似乎相同，但 9 之后的两位数 10 至 19，以 1 起首的数又大大抛离了其他数了。而下一堆 9 起首的数出现之前，必然会经过一堆以 2,3,4,...,8 起首的数。若果这样数法有个终结点，以 1 起首的数的出现率一般都比 9 大。

1881 年，天文学家西蒙 · 纽康发现对数表包含以 1 起首的数那首几页较其他页破烂。57 年后，物理学家法兰克 · 本福特重新发现这个现象，还通过了检查许多数据来证实这点。从 20 个不同的表格中找到了超过 20,000 条的数据支持，包括 335 条河流的区域数据，1389 种化学物质的比热，美国联盟基准球统计和从读者文摘和报纸的头版收集的数字。

2 Proof

下面给出一个比较经典的解释 (A Classical Explanation)

从 empirical significant-digit law 可以知道，在没有指定明确的统计实验的样本空间中，大多数都在试图证明这个定理的是纯粹的数学性质。这个想法是首先证明实数集满足 Benford's law，然后试着解释该定理的存在性。

$\{D_1 = 1\} = \{1, 10, 11, 12, 13, ..., 19, 100, 101, ...\}$ 这是一个以 1 为首位的所有正整数的集合，但是

$$\lim_{n\to\infty} \frac{1}{n}\{D_1 = 1\} \cap \{1, 2, ..., n\}$$

这个极限并不存在。

可以看出集合 $\{D_1 = 1\}$ 的经验密度在 $\frac{1}{9}$ 和 $\frac{5}{9}$ 中间震荡, 从而可以认为 [1/9, 5/9] 中的任意一个数都是这个集合可能的数值。

Jech（1992）发现的充分必要条件为有限可加条件集函数是对数函数。

但是，如果每个单整数出现等概率然后可列可加性意味着整个空间一定概率为零或无穷大。下面用离散的理论可以得倒，Fourier analysis 和 Banach measures 的正实数的连续密度是

$$\{D_1 = 1\} = \bigcup_{n=-\infty}^{\infty} [1, 2) \times 10^n$$

在这方面一个普遍的假设是尺度不变性 (scale invariance)

下面涉及的理论 THE NATURAL PROBABILITY SPACE, RANDOM SAMPLES FROM, RANDOM DISTRIBUTIONS 太深了，看不懂，所以还是看应用吧。

3 Application

1972 年，Hal Varian 提出这个定律来用作检查支持某些公共计划的经济数据有否欺瞒之处。1992 年，Mark J. Nigrini 便在其博士论文"The Detection of Income Tax Evasion Through an Analysis of Digital Frequencies."（Ph.D. thesis. Cincinnati, OH: University of Cincinnati, 1992.）提出以它检查是否有伪帐。

推而广之，它能用于在会计、金融甚至选举中出现的数据。该定律被华盛顿

邮报上的一篇文章引用，该文章以此为基础声称 2009 年伊朗总统大选中有造假。
若所用的数据有指定数值范围；或不是以机率分布出现的数据，如常态分布的数据；这个定律则不准确。
本福德定律的应用条件是：

1. 数据不能是规律排序的，比如发票编号、身份证号码等；
2. 数据不能经过人为修饰。

Examples

合同金额首位数	合同数量	占合同总量比例 (%)	本福特概率值 (%)
1	655	22.32	30.1
2	480	16.36	17.60
3	438	14.93	12.50
4	620	21.13	9.70
5	155	5.28	4.16
6	122	4.16	6.70
7	162	5.52	5.80
8	152	5.18	5.10
9	150	5.11	4.60
合计	2934	100.00	100.00

多次取样后发现，吻合度相当高。从而说明本次审计项目应用本福特定律抽取样本取得了较好的效果。

以上为应用。

4 REFERENCES

A Statistical Derivation of the Significant-Digit Law
本福特定律在审计抽样中的应用研究