本福特定律和美国公共教育财政调查

林昊

南京大学计算机科学与技术系 *171860673*

摘 要 本福特定律预言了自然形成的数据中，数字 1 ～ 9 出现的概率。本文应用本福特定律检测美国公共教育财政调查数据质量，通过 Pearson 相关系数检验、Chi-squared 检验和 K-S 检验确定美国公共教育财政调查数据符合本福特定律。文章最后，给出对本福特定律的进一步思考。

关键词 本福特定律，美国公共教育财政

1 本福特定律

1.1 本福特定律的发现

1881 年，美国天文学家 Simon Newcomb 发现，图书馆的对数表册上前面的页码比后面的页码磨损程度更加严重，并且磨损程度随着数字的增大而逐渐加深。他据此得出结论：在一个数据集当中，更小的数字出现的次数往往比大一点的数字出现次数多。1881 年，Newcomb 在《美国数学学报》上发表论文，描述他观察到的结果，并提出一个数字 N 作为一串数字的首位出现的概率近似等于 $log(N+1)-log(N)$。

1938 年，物理学家 Benford 再次观察到这一现象。他对来自 20 个不同领域的共 20229 条数据进行了测试，在印证这一结果的同时也让这一定律得到高度的认可。

1.2 本福特定律的定义

如果一组 m 进制数字当中的开头第一个数字 $d(d \in 1,2,..,m-1)$ 出现的概率满足

$$P(d) = log_m(d+1) - log_m(d) = log_m(\frac{d+1}{d}) = log_m(1+\frac{1}{d}) \tag{1}$$

那么我们认为这一组数据满足本福特定律。[1]

1.3 十进制下的本福特定律

考虑十进制，将 $m=10$ 及 $d=0,1,2,...,9$ 代入式 (1) 并把计算得到的概率分布律可视化，得到如图 Figure1 所示的计算结果：

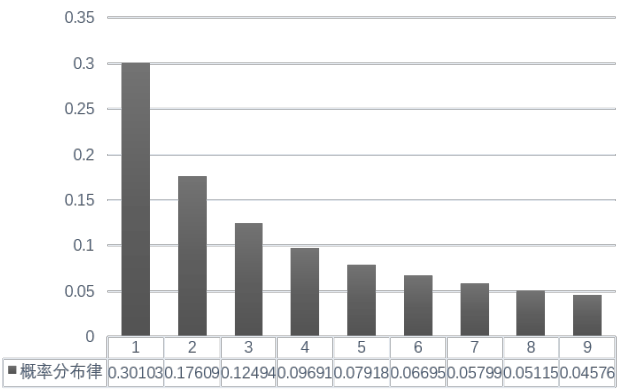

Figure 1: 本福特定律的概率分布（十进制）

1.4 本福特定律的应用范围

本福特定律如此强大，但也不是任何数据都可适用，比如门牌号、街道号、身份证号码等人为制定编号规则的数据就不适用。相反，本福特定律一般对来源于自然、经济、政治等的数据具有很好的适用性。它倾向于最准确地应用于跨越较多数量级的数据。根据经验，数据均匀覆盖的数量级越多，本福特定律越准确。

例如，在 1972 年，Hal Varian 建议本福特定律可用于检测为支持公共规划决策而提交的社会经济数据清单中可能存在的欺诈行为。基于合理的假设，即制造数字的人倾向于相当均匀地分配他们的数字，根据本福特定律对数据的第一位数频率分布与预期分布的简单比较应该显示异常结果；同样，在美国，基于本福特定律的证据已在联邦，州和地方的刑事案件中被接纳，并且被用作 2009 年伊朗选举欺诈的证据；本福特定律甚至逐渐用于科学欺诈检测、基因组数据分析、价格数字分析、宏观经济数据检测当中。[1]

2 美国公共教育财政

2.1 什么是美国公共教育财政

美国国家公共教育财政调查的主要是向公众提供每年用于幼儿园至 12 年级公共教育的国家级收入和支出。美国每一个州的教育机构有一年的时间来生成这些数据，并在财政年度结束后进行审计以提交给全国教育统计中心（NCES）。

2.2 数据简介

由于截止 2019 年 4 月 20 日，美国全国教育统计中心网站（https://nces.ed.gov/ccd/stfis.asp）上只有截止 2016 年的公共教育财政数据。因此，我们选取 2007 到 2016 财年这十年间的所有数据，对里面所有有效数据的首位数字的出现频数和频率进行统计，并观察是否与本福特定律的预测相符合。

2.3 数据初步处理

首先要对数据集进行清理。为了数据处理方便，将所有 EXCEL 文件当中的数据填补成 $56 * 155$ 的矩阵：若原始 EXCEL 当中没有这个数据，那么将 0 填入；其次，根据网站给出的帮助文档提示，清洗掉值 $x < 0$ 的数据，然后将剩余所有的数据读入一个三维数组，数组的高度代表年数，长度和宽度分别代表数据矩阵的行和列。

2.4 数据统计

对三维数组当中所有数字的首位数字进行统计（详见附录 A 中给出的 MATLAB 脚本文件 AnalyzeData.m），得到下面 Table1 所示的结果：

数字	1	2	3	4	5	6	7	8	9
频数	20776	12466	8303	6536	5661	4839	4065	3621	3234
频率	0.29893	0.17936	0.11947	0.09404	0.08145	0.06963	0.05849	0.05210	0.04653

Table 1: 数据处理结果

将得到的频率和本福特定律所预言的概率放入同一柱形图进行对比，直观上易知：对于美国公共教育财政支出的每一个首位数字，均和本福特定律的预言十分接近。(如图 Figure2)

为了进一步确定得到的概率分布是否符合本福特定律，我们选择 Pearson 相关系数检验、Chi-squared 检验和 Kolmogorov-Smirnov 检验同时进行检验。

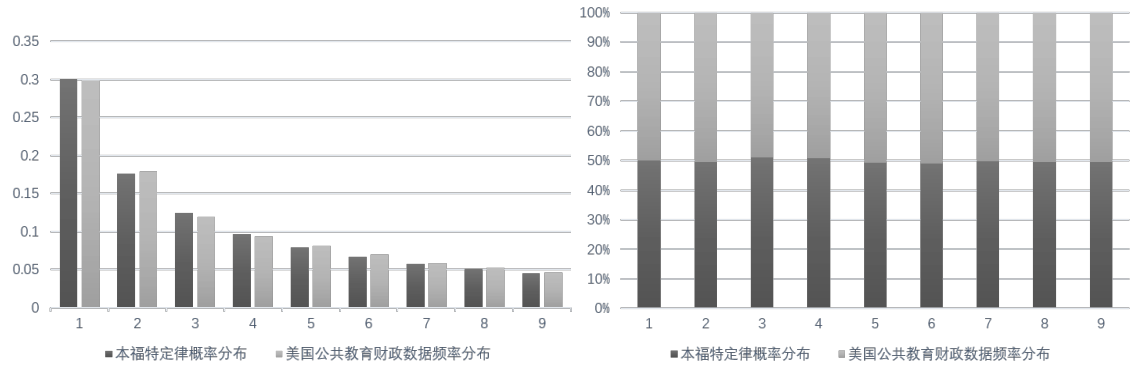

Figure 2: 本福特定律概率分布和美国公共教育财政数据频率分布柱形图

2.5 Pearson 相关系数检验

相关系数，是研究变量之间相关程度的量，一般用字母 r 来表示 [3]。本文中采用 Pearson 相关系数，用以度量两个变量 X 和 Y 之间的线性相关性，值介于 -1 与 1 之间。首先，用下式表示总体相关系数：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{2}$$

由于样本相关系数是总体相关系数的估计值，因此我们估算样本的协方差和标准差，可以得到 Pearson 相关系数如下 [5]：

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{3}$$

将本福特定律预言的十进制下首位数字出现的概率分布和 2007-2016 年间美国公共教育财政的首位数字出现的频率分布去中心化以后分别作为 X 和 Y 代入式（3）并利用 SPSS 软件进行验证，可以得到 Pearson 相关系数 $r = 0.999$。因此 X 和 Y 有很好的正相关性，即两个分布几乎完全相同。

2.6 Chi-squared 检验

Chi-squared 检验是以 χ^2 分布为基础的一种常用假设检验方法，它的零假设 H_0 是：观察频数与期望频数没有差别。其基本思想是：首先假设 $H0$ 成立，基于此前提，利用公式

$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i-E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(A_i-np_i)^2}{np_i},\ i = 1,2,3,...,k \tag{4}$$

计算出 χ^2 值，它表示观察值与理论值之间的偏离程度。其中，A_i 为 i 水平的观察频数，E_i 为 i 水平的期望频数，n 为总频数，p_i 为 i 水平的期望频率。i 水平的期望频数 E_i 等于总频数 $n \times i$ 水平的期望概率 p_i，k 为单元格数。[2]

根据 χ^2 分布及自由度可以确定在 H_0 假设成立的情况下获得当前统计量及更极端情况的概率 p。如果 p 值很小，说明观察值与理论值偏离程度太大，应当拒绝无效假设，表示比较数据之间有显著差异；否则就不能拒绝无效假设，尚不能认为样本所代表的实际情况和理论假设有差别。

该问题中，做出假设 H_0：2007-2016 年美国公共教育数据的首位数字分布满足本福特定律。利用 SPSS 软件对首位数字分布律和本福特定律指出的分布律进行 Chi-squared 检验，并设显著性值为 0.05，可得检验结果如下 Table2：

可以看出，Pearson 卡方数值 72 即为卡方值，其渐近显著性值即为 p 值，在显著性 0.05 的前提下，可以接受假设 H_0，即 2007-2016 年美国公共教育数据的首位数字频率分布满足本福特定律。

	数值	df	渐近显著性（2 端）
皮尔森 (Pearson) 卡方	72	64	0.23
概似比	39.55	64	0.993
线性对线性关联	7.991	1	0.005
有效观察值个数	9		

Table 2: SPSS 软件 Pearson Chi-squared 检验结果

2.7 Kolmogorov-Smirnov 检验

Kolmogorov-Smirnov 检验，简称 K-S 检验，是一个基于累计分布函数，用以检验两个经验分布是否不同或一个经验分布与另一个理想分布是否存在差异的检验方法。此处考虑一个经验分布和另一理想分布的情况，K-S 检验认为：若原假设 H_0：观测数据的频率分布 $F(x)$ 符合理论分布 $G(x)$，若检验统计量 D_{max} 在样本总数 $N > 35$ 满足公式

$$D_{max} = max\{|F(x) - G(x)|\} < \frac{1.36}{\sqrt{N}} \tag{5}$$

或者渐近显著性 p 满足条件

$$p > 0.05 \tag{6}$$

时，接受原假设 H_0，否则不接受原假设 H_0。[4]

该问题中，做出假设 H_0：2007-2016 年美国公共教育数据的首位数字分布满足本福特定律。将 Table2 中第三行首位数字出现的概率分布和本福特定律所预言的首位数字出现的概率分布作为 K-S 检验的两个分布代入 SPSS 软件计算，有如下 Table3 的结果：

	本福特定律理论分布	实际样本频率分布
Kolmogorov-Smirnov 检定	1.247	1.247
渐近显著性（双尾）	0.089	0.089

Table 3: SPSS 软件 K-S 检验结果

因为渐近显著性 $p > 0.05$，因此，我们接受零假设“2007-2016 年间美国公共教育财政数据满足本福特定律”。

3 结论

基于以上讨论，2007-2016 年间美国公共教育财政数据在显著性 $\alpha = 0.05$ 的条件下满足本福特定律，并且考虑不存在数据造假问题。

4 更多的思考

4.1 本福特定律——第二位数字？

本福特定律是否可以拓展到第一位以外的数字呢？事实上，如果把一个数串的前 n 位全部看作本福特定律的首位数字，我们就可以求出第 n 位数字为 d 的概率。若一个 m 进制数串以数字 d 开头的概率由公式（1）给出，那么该数串当中第 n 个数字为 d 的概率为

$$P(d) = \log_m(1+\frac{1}{1*m+d})+...+\log_m(1+\frac{1}{(m-1)*m+d}) = \sum_{k=m^{(n-2)}}^{m^{(n-1)}-1} log_m(1+\frac{1}{m*k+d}) \tag{7}$$

以十进制数为例，式（7）可以表达为：

$$P(d) = \sum_{k=10^{(n-2)}}^{10^{(n-1)}-1} log_{10}(1+\frac{1}{10k+d}) \tag{8}$$

将十进制不同位的数字代入计算得到 Table4，我们可以明显观察到随着选取的数字位数 n 的增加，第 n 个数字的分布迅速接近均匀分布。

位数	$P(0)$	$P(1)$	$P(2)$	$P(3)$	$P(4)$	$P(5)$	$P(6)$	$P(7)$	$P(8)$	$P(9)$
1		0.3010	0.1761	0.1249	0.0969	0.0792	0.0670	0.5799	0.0512	0.0458
2	0.1197	0.1139	0.1088	0.1043	0.1003	0.0967	0.0934	0.0904	0.0876	0.0850
3	0.1018	0.1014	0.1010	0.1006	0.1002	0.0998	0.0994	0.0990	0.0986	0.0983
4	0.1002	0.1001	0.1001	0.1001	0.1000	0.1000	0.0999	0.0999	0.0999	0.0998

Table 4: 前 4 位数字情况下本福特定律计算所得概率分布律

References

[1] Wikipedia. Benford's law. https://en.wikipedia.org/wiki/Benford's_law.

[2] Wikipedia. Chi-squared test. https://en.wikipedia.org/wiki/Chi-squared_test.

[3] Wikipedia. Correlation coefficient. https://en.wikipedia.org/wiki/Correlation_coefficient.

[4] Wikipedia. Kolmogorov–smirnov test. https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test.

[5] Wikipedia. Pearson correlation coefficient. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

A AnalyzeData.m

```
%%
clc;clear;
for i=1:10
    Mat(:,:,i)=zeros(56,155);
end
%%
Mat(:,:,1)=xlsread('F:\midterm\Dataset\stfis071b.xls',1,'E2:FC57');
Mat(:,:,2)=xlsread('F:\midterm\Dataset\stfis081b.xls',1,'E2:FC57');
Mat(:,:,3)=xlsread('F:\midterm\Dataset\stfis091b_xls.xlsx',1,'E2:FC57');
Mat(:,:,4)=xlsread('F:\midterm\Dataset\stfis101a_xls.xlsx',1,'E2:FC57');
Mat(:,:,5)=xlsread('F:\midterm\Dataset\stfis11_1a.xls',1,'E2:FC57');
Mat(:,:,6)=xlsread('F:\midterm\Dataset\Stfis12_1a_xls.xls',1,'E2:FC57');
Mat(:,:,7)=xlsread('F:\midterm\Dataset\Stfis13_1a.xls',1,'E2:FC57');
Mat(:,:,8)=xlsread('F:\midterm\Dataset\Stfis14_1a.xlsx',1,'E2:FC57');
Mat(:,:,9)=xlsread('F:\midterm\Dataset\stfis15_1a.xlsx',1,'E2:FC57');
Mat(:,:,10)=xlsread('F:\midterm\Dataset\Stfis16_1A.xlsx',1,'E2:FC57');
%%
count(1:9)=zeros(1,9);
for i=1:56
    for j=1:155
        for k=1:10
            if(Mat(i,j,k)<0)
                Mat(i,j,k)=0;
            end
            numStr=num2str(Mat(i,j,k));
            firstNum=numStr(1)-'0';
            switch firstNum
                case{1}
                    count(1)=count(1)+1;
                case{2}
                    count(2)=count(2)+1;
                case{3}
                    count(3)=count(3)+1;
                case{4}
                    count(4)=count(4)+1;
                case{5}
                    count(5)=count(5)+1;
                case{6}
                    count(6)=count(6)+1;
                case{7}
                    count(7)=count(7)+1;
                case{8}
                    count(8)=count(8)+1;
                case{9}
                    count(9)=count(9)+1;
            end
        end
    end
end
```

```
countTotal=sum(count);
proportion(1:9)=zeros(1,9);
theoProportion(1:9)=zeros(1,9);
theoCount(1:9)=zeros(1,9);
for i=1:9
    proportion(i)=count(i)./countTotal;
    theoProportion(i)=log10(i+1)-log10(i);
    theoCount(i)=theoProportion1(i)*countTotal;
end
```