Science Innovation
2021; 9(2): 53-62
http://www.sciencepublishinggroup.com/j/si
doi: 10.11648/j.si.20210902.14
ISSN: 2328-7861 (Print); ISSN: 2328-787X (Online)

Analysis and Prediction of COVID-19 Data Quality Based on Benford's Law-- Take Data from 51 Countries and Regions as an Example

Cheng Han Leung[1], Yu Bo Luo[1, 2, *], Tan Cheng Lok[1], Zi Chen Luo[3]

[1]School of Business, City University of Macau, Macau, China
[2]Department of Accounting, School of Economics and Management, Guangdong Institute of Petrochemical Technology, Maoming, P. R. China
[3]Institute of Data Science, City University of Macau, Macau, China

Email address:
tiffangleung@cityu.mo (Cheng Han Leung), lyb0668@126.com (Yu Bo Luo), Stellalok@chwcpa.com.mo (Tan Cheng Lok), 1773293851@qq.com (Zi Chen Luo)

[*]Corresponding author

To cite this article:
Cheng Han Leung, Yu Bo Luo, Tan Cheng Lok, Zi Chen Luo. Analysis and Prediction of COVID-19 Data Quality Based on Benford's Law--Take Data from 51 Countries and Regions as an Example. *Science Innovation.* Vol. 9, No. 2, 2021, pp. 53-62. doi: 10.11648/j.si.20210902.14

Received: March 8, 2021; **Accepted**: April 23, 2021; **Published**: April 26, 2021

Abstract: In the absence of effective drugs to control the COVID-19 epidemic, the main intervention of human factors, namely strict isolation, may be the best prevention and control method at present. The conclusion of data empirical study using Benford's Law is of great significance.Research purpose of this paper analyze COVID - 19 data to predict the authenticity and reliability, and on this basis, the method is to use Benford's Law and the panel model for 51 countries or regions COVID - 19 data statistical analysis, the results of the study found that "other areas" unreliable data, Australia, Pakistan and global data are greatly influenced by artificial factors, Africa, Oceania data with several other states Data according to have significant difference, compared the southern hemisphere and northern hemisphere, the first phase of the data and the second stage also has the obvious difference between the data, The COVID-19 data are also predicted to suggest that the outbreak may have multiple iterations.In conclusion, in most cases, when COVID-19 data deviates from Benford's Law, epidemic prevention and control is better; otherwise, it is worse.

Keywords: Benford's Law, COVID-19, Data Quality, Prediction

基于Benford’s Law的COVID-19数据质量分析及预测

梁静娴[1]，罗玉波[1, 2*]，陆丹青[1]，罗子辰[3]

[1]澳门城市大学商学院，澳门，中国
[2]广东石油化工学院经济管理学院会计系，茂名，中国
[3]澳门城市大学数据科学研究院，澳门，中国

邮箱
tiffangleung@cityu.mo（梁静娴），lyb0668@126.com（罗玉波），Stellalok@chwcpa.com.mo（陆丹青），1773293851@qq.com（罗子辰）

摘要：COVID-19疫情在没有有效控制的药物去前提下，人为因素的主要干预，也就是实行严格的隔离可能是目前最好的防控办法，用Benford’s Law进行数据实证其结论具有重要意义，本文研究目的是分析COVID-19数据可靠性和真实性并在此基础上进行预测，方法是运用Benford’s Law和面板模型对于51国家或地区的COVID-19数据进行统计分析，

研究结果发现“其他地区”数据不可信赖，澳大利亚、巴基斯坦数据和全球数据受人为因素影响较大，非洲、大洋洲数据与其他洲数 据相比有显著性差异，南半球数据与北半球、第一阶段数据与第二阶段数据之间也具有明显的差异，同时预测COVID-19数据认为疫情可能出现多次反复。总之，在大多数情况下当COVID-19数据偏离Benford’s Law是，疫情防控较好，反之，则较差。

关键词：Benford’s Law，COVID-19，数据质量，预测

1．引言

2019年12月或更早，在武汉市华南海鲜市场出现一种传染性很强的病毒，于2020年2月11日，世界卫生组织将该病毒命名为“COVID-19”(Corona Virus Disease 2019)，由于该病毒是一种新型病毒，没有相应的比较有针对性的治疗药物，人们在逐渐认识病毒的过程中，只能加强多方面的防控。该病毒在中国湖北省武汉市大面积爆发。随后，网络上披露了大量关于病毒感染的相关信息，各种数据信息的质量引起了人们普遍的关注，如何判断这些数据的真实、可靠性呢？本文拟用美国物理学家Frank Benford在1938年发现的数据统计规律-Benford’s Law（本福特定律）来分析数据的真实性可靠性。Benford’s Law即在自然条件下，大量数据统计规律遵循：首位是1的统计数据比率性为30.1%，数字越大，比率越小，并依次减少；首位数字为9的统计比率为4.6%，即首位数字越小而统计的比率反而越大。当遇到人为因素影响时，此统计结果会出现相应的偏离。

Benford’s Law在数据信息质量方面的应用越来越广泛，Goodman（2016）认为Benford’s Law较好的鉴别出异常数据和评价数据信息的质量[4]。Benford定律也被应用于公共卫生方面的数据分析。Sambridge [11]等（2010）Nigrini（2019）[14]指出在2007年世界卫生组织报告的传染病数字的分布似乎很符合Benford’s Law。Zhang（2020）[13]利用Benford定律检验中国的COVID-19数据信息，证明数据信息是值得信赖的；而且Collins（2017）[1]等发现利用Benford定律检验数据信息质量的成本可以大大降低。

本文的主要研究目的是利用Benford’s Law分析中国疾控中心公布的病毒相关数据的可靠性，以消除人们对相关数据的质疑。同时，根据研究结果，进一步分析感染新型冠状病毒COVID-19的病人死亡人数、确诊病例和疑似病例数据之间的关系，并对未来相关数据变化的趋势进行预测。

本文主要内容分成以下五个部分：

第一部分为引言，介绍了本文的研究背景。

第二部分进行了文献综述。对国内外应用Benford’s Law检验统计数据真实性、可靠性的文献进行综述，对使用回归模型进行相关数据分析和预测的文献进行梳理。

第三部分分析了Benford’s Law和面板模型设计。进行Benford’s Law理论分析，介绍拟合优度χ^2卡方、相关系数R、距离d^*和m^*等公式。

第四部分进行了实证研究检验。利用Benford’s Law和面板模型，鉴别50多个国家和地区相关数据的可靠性，并将按照各大洲、南北半球以及第一、二阶段分类检验数据的可靠性、真实性依次进行，并比较是否有显著性差异。

第五部分预测了COVID-19在未来发展的趋势。利用相关数据建立回归曲线模型，进行相应数据的未来变化趋势预测。

第六部分是对上述分析内容做相应的总结，并阐述了本文的局限性。

2．文献综述

利用Benford’s Law鉴别统计数据真实性、可靠性的相关文献综述

2.1．Benford’s Law在公共医疗监控系统的应用

Benford’s Law可用于快速评估数据质量和流行病学监测系统的敏感性。当数据的第一位数遵循Benford分布时，这~~强烈~~表明监视系统可靠且有效[2，3，7]，使用Benford’s Law来验证公共卫生监控系统的可靠性的主要优势在于其成本低廉[8]。Kuruppu & Muscat（2019）[6]、Pomykacz & Tantinan（2017）[10]和Collins（2017）[1]提出了可在Excel界面使用Benford’s Law，以大大降低成本。Singleton（2011）[12]、Nigrini（2019)[9]等模拟了环境中人类审计师行为的强化学习模型来改良传统方法，发现其精确度优于传统的Benford定律的研究方法。

Benford’s Law也被应用于公共卫生科学研究。Crocetti[2]证明癌症发病率遵循Benford分布，观测分布与预期分布之间的相关系数非常高。Daniels （2017）[3]使用卡方检验调查了小型和大型数据库中的死亡人数是否都遵循Benford分布，他们的研究结果支持了这一假设。

Idrovo 等（2011）[5]根据Benford’s Law和死亡率，检查了美洲A（H1N1）型流感大流行期间监测系统的数据质量和敏感性。Manrique-Hernandez等[8]（2017）在拉美和加勒比地区2014年寨卡病毒暴发的流行病学监测中采用了类似的方法。这些研究人员都发现不同国家的监测系统在绩效方面表现并不均衡。事实证明，Benford’s Law是一种用于评估公共卫生监视系统性能的新颖、简单、客观和低成本的工具。

Zhang（2020）[13]的最新工作成果是启发我们进行这次研究的核心文献之一。Zhang（2020）[13]使用Newcomb-Benford Law测试了2019年中国新型冠状病毒病的病例数，测试结果并未表明发现欺诈行为，COVID-19的累计案例数遵从Benford Law。此外，作者还指出研究结果的局限：由于缺乏医疗资源和对诊病例的不成熟定义，该测试无法判断是否有一部分患者未被正确纳入研

究。我们的研究结果通过将研究期间延长至五月上旬来补充了Zhang（2020）[13]的工作。

最后，需要指出的是Zhang（2020）[13]曾提到，将Benford's Law应用于公共卫生监控系统的质量保证时，无法识别是否存在因缺乏医疗资源而系统地将一部分患者排除在外的现象，同时，对确诊病例定义的不成熟以致数据偏差亦未能借此方法发现，除非在研究期间相关因素曾发生改变。

2.2．预测模型的文献综述

本文利用收集的数据资料进行回归分析，分析每天新增确诊病例人数与时间t之间的关系，这一方法是采用了如下相关文献的研究成果，吕娜和邹薇（2015）[18]利用CHNS数据考察1991至2006年间我国居民健康投资与健康人力资本和居民收入水平的关系；杜玉忠等（2019）[14]基于向量自回归模型分析了清远市手足口病与气象因素之间的关系,预测结果的误差不大于17.75%，因而气象因素和手足口病发病的VaR模型可以较好地进行清远市手足口病发病的短期预测。

3．Benford's Law和面板模型设计

3.1．Benford's Law

1881年，美国Simon Newcomb偶然发现首位数概率分布，但是他对这一现象并未做进一步研究。后来，美国物理学家 Frank Benford（1938）也注意到了这一现象，并收集了不同类型的数据共计20229项，涉及领域广泛。经过大量的实证研究，他证明了 Simon Newcomb 的理论。Benford研究发现，首位数为"1"的数字统计的频率是0.301，首位数为"2"的数字统计的频率是0.176，也就是说，首位数字越大，统计出现频率依次减少。Hill（1995）从理论上对 Benford's Law给出了令人满意的解释，并进行了严谨的数学证明，同样发现研究的样本数据量越大，统计结果越接近Benford's Law的理论分布。

Benford's Law首位数出现的概率公式：

$$P_n = \log_{10}(1+\frac{1}{n}) \tag{1}$$

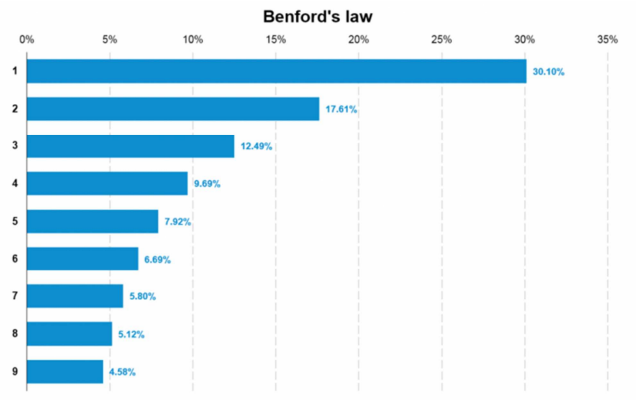

图1 Benford's Law 统计图表。

其中，首位数字 n 是指左边的第一位非零的有效数字。根据公式（1）首位数概率分布如表 1 所示。

表1 Benford 定律首位数概率分布。

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

3.2．Benford's Law的检验方法

现有四种常用的Benford's Law检验方法，验证所取样本数据的分布是否符合 Benford's Law的期望频率分布。

3.2.1 Pearson相关系数

计算样本数据的首位数频率分布与首位数期望频率分布的 Pearson 相关系数，相关系数越接近于1，则越符合Benford's Law理论分布。

$$R = \frac{(\Sigma_{i=1}^{9}(p_i-\bar{p}_i)(e_i-\bar{e}_i))}{\sqrt{(\Sigma_{i=1}^{9}(p_i-\bar{p}_i)^2\Sigma_{i=1}^{9}(e_i-\bar{e}_i)^2)}} \tag{2}$$

3.2.2．拟合优度检验χ2

拟合优度检验χ2为重要的检验方法。下列检验公式中，e_i为首位数是i的实际频率，p_i为 Benford's Law的理论频率。F_e（x）为实际所取样本首位数的累积分布函数，F_p（x）为Benford's Law的理论频率分布下的首位数的累积分布函数。

$$\chi^2 = N\cdot\sum_{i=1}^{9}\frac{(e_i-p_i)^2}{p_i} \tag{3}$$

在显著性水平为 10%、5%和 1%条件下，χ^2的 临界值分别为 13.36、15.51 和 20.09。

原假设：实际样本数据首位数分布符合 Benford's Law的理论分布频率。

备择假设：实际样本数据首位数分布不符合Benford's Law的理论分布频率。

通过计算，如果统计量大于临界值，则拒绝原假设，接受备择假设，表明该样本数据与 Benford's Law理论分布频率不相符，出现偏离Benford's Law统计规律，样本数据或者人为影响因素较大，或者数据有人为因素影响的可能。

3.3.3．修正 Kolmogorov- Smirnov 拟合优度检验

根据 Kolmogorov- Smirnov 检验相关理论，将实际样本首位数累积分布函数减去理论上 Benford's Law分布累计频率之差，再取绝对值，并选取最大值作为统计计算 D 值，然后将 D 值与临界值进行比较，若超过临界值，则说明实际所取样本首位数分布不服从该理论分布。Stephens（1970）对 K- S 拟合优度检验作了修正，检验方法如下：

$$V_N=\max[F_e(x)-F_p(x)]+\max[F_p(x)-F_e x)] \quad (4)$$

后来Giles（2007）对公式（3）的统计量进行修正，改成以下计算方法：

$$V_N^*=V_N\cdot[N^{1/2}+0.155+0.24\times N^{-1/2}] \quad (5)$$

在显著性水平为 10%、5%和 1%条件下，V_N^* 的临界值分别为 1.19、1.32 和 1.58。

3.3.4．修正的距离检测

计算实际所取样本首位数的实际频率分布与Benford's Law理论分布之间的距离，其中距离偏离越大，超过临界值则不符合Benford's Law 理论分布距离计算公式：

$$d=\sqrt{\{\sum_{i=1}^{9}(p_i-e_i)^2\}} \quad (6)$$

$$m=\max_{i=1,2,\cdots,9}\{|p_i-e_i|\} \quad (7)$$

Morrow（2014）对上述距离进行了修正：

$$d^*=\sqrt{N\cdot\sum_{i=1}^{9}(p_i-e_i)^2\}} \quad (8)$$

$$m^*=\sqrt{N}\max_{i=1,2,\cdots,9}\{|p_i-e_i|\} \quad (9)$$

在10%、5%和 1%的置信水平下，d^* 统计量判别值分别为 1.212、1.330 和 1.569；m^* 统计量判别 值分别为 0.851、0.967 和 1.212。

表2 Benford's law 检验方法。

方法	统计量	1%显著性水平	5%显著性水平	10%显著性水平
拟合优度检验χ^2	$\chi^2=N\cdot\sum_{i=1}^{9}\frac{(e_i-p_i)^2}{p_i}$	20.09	15.51	13.36
修正的Kolmogorov-Smirnov	$V_N^*=V_N\cdot[N^{1/2}+0.155+0.24\times N^{-1/2}]$	1.58	1.32	1.19
拟合优度检验	$V_N=\max[F_e(x)-F_p(x)]+\max[F_p(x)-F_e x)]$	1.569	1.330	1.212
修正的距离检验	$d^*=\sqrt{N\cdot\sum_{i=1}^{9}(p_i-e_i)^2\}}$ $m^*=\sqrt{N}\max_{i=1,2,\cdots,9}\{\|p_i-e_i\|\}$	1.212	0.967	0.851
Pearson相关系数	$R=\frac{(\sum_{i=1}^{9}(p_i-\bar{p}_i)(e_i-\bar{e}_i))}{\sqrt{(\sum_{i=1}^{9}(p_i-\bar{p}_i)^2\sum_{i=1}^{9}(e_i-\bar{e}_i)^2)}}$	相关系数越接近1，越符合Benford's law		

面板模型和Benford's Law可以综合运用在数理统计中，面板数据可以提供横截面和时间序列两个维度上的数据信息，并把它们融合在一起。利用Benford's Law来观测样本中存在差异的数据，应用适当的面板回归方法，不仅可以用于模拟因变量与自变量之间的关系，也可以预测数据未来发展趋势。在应用Benford's Law对数据进行检验分析的基础上，再构建面板模型进行拟合，进一步发现哪些国家的和哪个首位数字存在可疑点。

面板模型和Benford's Law结合的思路如下：首先，对研究样本中的各个COVID-19数据进行第一位数字计算，统计出各各国COVID-19数据第一位数的频率分布；然后，将计算出的第一位数字频率分布与Benford's Law第一位数的期望分布进行统计学检验和分析，判断两者之间差异是否显著，具有显著差异的则很可能是存在人为因素影响的COVID-19数据；再次，运用面板模型对很可能存在人为因素影响的数据进行回归模拟和残差分析，并计算置信区间，仅存在极少数样本点数据首位数字偏离出置信区间，偏离的部分第一位数字的相关数据则可能存在问题；最后，根据残差分析，得到的置信区间外数据，找出有可能出现偏离Benford's Law预期值的原因。

4．实证研究检验

4.1．数据收集

我们在中国疾病预防控制中心网站，收集到2020年1月16日至2020年5月19日间，50多个国家和地区新型冠状病毒肺炎的相关数据14221个，数据类型为累计确诊、新增确诊、累计死亡和新增死亡病例数据；本文前后三次收集数据以湖北首次出现COVID-19病例开始到中国周边病例得到基本控制划分为第一阶段，具体时间为2020年1月16日至2020年3月6日，数据1222个（含全国数据294个和湖北267个，其他9个国家和地区661个，第二阶段时间为2020年3月7日至4月18日（有的国家后来增加数据可能到2月份），第三阶段时间为2020年4月19日至5月19日，从三月初到五月其他国家大面积出现病例到目前基本上得到控制划分为第二、三阶段。收集样本国家的依据是新冠肺炎比较严重，感染人数较多，数据来源亚洲13个国家、南美洲11个国家、欧洲16个国家、北美洲3个国家、非洲6个国家和大洋洲2个国家，加上全球合计数据总计有51个国家和全球数据；通过统计首位数字得到下列统计数据，并利用刘云霞等（2013）[16]、韩兆洲和程学伟（2019）[15]的相关系数r、χ^2拟合优度检验、V_N^*和m^*、d^*等来判断数据的受人为因素影响情况[17]。

同时，将统计数据频率信息分成国家或者地区间检验，各大洲之间的差异检验，南半球、北半球之间异同的检验；还有第一阶段和第二、三阶段统计首位数字偏离的检验。在此基础上，本文提出下列假设：

$H1_0$：各国家和地区与COVID-19病例数据Benford's Law统计数字频率之间是独立的（不存在依赖关系）,即不符合Benford's Law统计数字频率规律；

$H1_1$：各国家和地区与COVID-19病例数据的Benford's Law统计数字频率之间不独立（存在依赖关系），即符合Benford's Law统计数字频率。

$H2_0$：各大洲与COVID-19病例数据Benford's Law统计数字频率之间是独立的（不存在依赖关系）,即不符合Benford's Law统计数字频率规律；

$H2_1$：各大洲与COVID-19病例数据的Benford's Law统计数字频率之间不独立（存在依赖关系），即符合Benford's Law统计数字频率。

$H3_0$：分南半球和北半球 （气候）与COVID-19病例数据Benford's Law统计数字频率之间是独立的（不存在依赖关系），即不符合Benford's Law统计数字频率规律；

$H3_1$：分南半球和北半球 （气候）与COVID-19病例数据的Benford's Law统计数字频率之间不独立（存在依赖关系），即符合Benford's Law统计数字频率。

$H4_0$：第一阶段与第二、三阶段数据之间COVID-19病例数据Benford's Law统计数字频率之间是独立的（不存在依赖关系），即不符合Benford's Law统计数字频率规律；

$H4_1$：第一次爆发与第二三阶段数据之间COVID-19病例数据Benford's Law统计数字频率之间不独立的（存在依赖关系），即符合Benford's Law统计数字频率。

4.2．2020年4月18日前COVID-19统计数据首位数字频率分析

我们首先将收集到的数据按照国家或者地区进行分类，统计其数字的首位数据频率，然后通过SPSS和EXCEL计算统计检验量相关系数R、卡方检验值χ^2、距离d^*、m^*和修正的拟合优度V_N^*，结果参见表3。

表3 统计检验量表（4月18日前数据）。

序号	国家或地区	R	χ^2	样本量
1	俄罗斯	0.95916557	6.327783023	131
2	印度	0.953276975	8.684596046	154
3	巴西	0.961039689	12.60380788	149
4	加拿大	0.927315779	16.78531296**	162
5	荷兰	0.95957855	19.1886829**	177
6	爱尔兰	0.954040236	10.33978722	160
7	以色列	0.921055227	12.33552176	139
8	比利时	0.990280827	7.449767176	157
9	葡萄牙	0.99050134	1.958836017	153
10	瑞典	0.969216101	6.980620515	160
11	美国	0.991819436	18.30636666**	232
12	土耳其	0.909592481	16.58586631**	124
13	西班牙	0.87350101	19.70808942**	170
14	德国	0.902375751	13.62967116*	128
15	法国	0.990894163	1.770299793	168
16	意大利	0.921019099	14.99337803*	183
17	英国	0.927046907	13.32596198	169
18	伊朗	0.987157136	9.549613828	233
19	韩国	0.987157136	11.14722613	278
20	沙特阿拉伯	0.942441375	7.04377166	133
21	澳大利亚	0.936373039	25.36012112***	204
22	瑞士	0.919111991	12.34728448	167
23	巴基斯坦	0.916802333	22.64401165***	150
24	秘鲁	0.958363628	9.090130881	128
25	哥伦比亚	0.973843456	9.27206187	129
26	智利	0.937625987	7.579596438	137
27	墨西哥	0.950361777	13.48617114*	141
28	乌拉圭	0.842068539	14.80135036*	90
29	丹麦	0.980013113	8.406221394	167
30	挪威	0.94871826	8.288666217	111
31	日本	0.973877187	5.937236769	263
32	马来西亚	0.964477284	16.79144142**	200
33	新加坡	0.98566336	4.041681252	203
34	乌克兰	0.973061665	4.417078521	133
35	菲律宾	0.97782107	8.909169433	171
36	白俄罗斯	0.868388044	10.62795165	102
37	新西兰	0.862338207	15.7440703**	106
38	南非	0.949420369	16.14670584**	122
39	摩洛哥	0.977488828	9.306790101	154
40	阿联酋	0.925044702	11.88339536	167
41	其他地区	0.304874727	77.41568546***	63
42	全球	0.951967977	31.96151431***	360

注：*、**、***分别表示具有10%、5%和1%显著性。

具体分析如下：

1.按照国家和地区分析统计数据的差异。

按照卡方值分类成符合Benford's Law程度很好的国家或者地区、符合程度一般的国家或者地区和符合程度欠佳的国家和地区。

从各个国家和地区分析，澳大利亚、巴基斯坦、其他地区和全球合计数据首位数字的统计频率与Benford's Law期望频率有显著性差异，其他国家数据与Benford's Law期望频率没有显著性差异。也就是说澳大利亚、巴基斯坦、其他地区的数据首位数频率偏离Benford's Law期望频率大，受人为影响因素较大，但是偏离程度不大，数据还是值得信赖的可靠的。

"其他地区"的数据偏离程度很大，基本上不符合Benford's Law期望，从其他检验值也可以看出，相关系数为0.30487，d^*为1.5682，m^*为1.007，V^{N*}为1.8333均超出临界点。这表明"其他地区"的统计数据并不值得信赖，可靠性存在较大问题。

综上，上述45个国家的数据都是基本可靠和真实的，但是澳大利亚、巴基斯坦两国数据与其他国家数据还是存在明显的差异，全球汇总数据由于受到澳大利亚、巴基斯坦、其他地区数据等因素的影响也出现与Benford's Law期望频率较大的偏离。

2.按照亚洲、欧洲、大洋洲、非洲等分类，分析COVID-19数据符合Benford's Law期望频率有无明显的差异，结果参见表4。

表4 各大洲COVID-19数据Benford's Law卡方值。

	χ^2
亚洲	8.870956183
欧洲	1.790187578
北美洲	18.66845105**
南美洲	14.22424186*
非洲	20.35814609***
大洋洲	27.88813654***
其他地区	77.41568546***

注：*、**、***分别表示具有10%、5%和1%显著性。

将上述国家和地区按照各大洲出现分类计算统计检验数据，拟合优度值（卡方χ^2）如上表，亚洲和欧洲数据很好的符合Benford's Law期望频率，北、南美洲也较好的符合Benford's Law期望频率，大洋洲拟合优度值（卡方χ^2）27.88813654大于临界值20.09，但超过临界值不是很大，COVID-19数据有一定程度的人为因素影响，数据也基本可靠信赖；亚洲、欧洲、北美洲和南美洲COVID-19数据信息是可靠的、真实的，他们之间也没有显著性差异，但非洲、大洋洲与亚洲、欧洲、北美洲、南美洲相比，相关信息还是存在明显的差异。

3.按照南、杯半球（气候影响）分类，分析COVID-19数据符合Benford's Law期望频率有无明显的差异，结果参见表5。

表5 南、北半球COVID-19数据Benford's Law卡方值。

	χ^2
北半球	7.026116705
南半球	26.14481595***

注：*、**、***分别表示具有10%、5%和1%显著性

北半球COVID-19病例数据首位数字频率与Benford's Law期望频率之间存在依赖关系，很好的符合Benford's Law统计期望频率，相关信息是可靠和真实的，但南半球COVID-19病例数据首位数字频率的卡方值26.14481595超过临界值20.09，存在人为因素影响较大，有可能会有气候影响的因素在内部发生作用。

4.按照第一阶段与第二次阶段分类，分析COVID-19数据符合Benford's Law期望频率有无明显的差异，结果参见表6。

表6 第一、二阶段COVID-19数据Benford's Law卡方值。

	χ^2
第一阶段	24.1667724***
第二阶段	8.522832758

注：*、**、***分别表示具有10%、5%和1%显著性。

第二阶段COVID-19病例数据首位数字频率与Benford's Law期望频率之间存在依赖关系，很好的符合Benford's Law统计期望频率，相关信息是可靠和真实的，但第一阶段COVID-19病例数据首位数字频率的卡方值24.1667724超过临界值20.09，存在一定人为因素影响。

4.3．2020年5月19日前COVID-19统计数据首位数字频率分析

表7 样本量统计检验表(2020.5.19年前)。

Num	Continent	Nation	χ^2	R	Size
1	Asian	India	8.418097941	0.972367687	278
2		Israel	13.33691243	0.99077878	250
3		Turkey	14.21168482*	0.969599589	248
4		Iran	36.26265044***	0.912300286	357
5		South Korea	38.70062533***	0.953798374	392
6		Saudi Arabia	7.187616351	0.985707377	257
7		Pakistan	6.869487297	0.986115754	274
8		UAE	15.80756948**	0.940400676	289
9		Japan	3.427664126	0.995113276	387
10		Malaysia	21.0318872***	0.962382982	312
11		Singapore	9.396608	0.98161249	305
12		Philippines	8.234113249	0.98571878	295
13		China	10.28277349	0.96395017	184
Asian total			35.71782062***	0.993016713	3828

Num	Continent	Nation	χ^2	R	Size
14		Russia	17.83204432**	0.961946818	251
15		Netherlands	20.05196057**	0.9453586	301
16		Ireland	27.43153365***	0.966394104	280
17		Belgium	23.00795542***	0.926933739	281
18		Portugal	27.80795707***	0.920496282	276
19		Sweden	13.01087607	0.959568517	284
20		Spain	33.06514598***	0.87130126	290
21	Europe	Germany	21.21102419***	0.970528555	250
22		France	12.54552793	0.985303673	291
23		Italy	16.4250982	0.925739959	307
24		United Kingdom	19.49047905**	0.899175626	293
25		Switzerland	15.48101965*	0.969544094	286
26		Denmark	18.67259509**	0.927135064	290
27		Norway	49.65350845***	0.810540441	220
28		Ukraine	15.44993582*	0.96394288	257
29		Belarus	21.05125365	0.921016582	212
Europe total			22.92793713***	0.993277732	4369
30		Egypt	6.545937126	0.984395693	284
31		Nigeria	13.06372873	0.988457347	193
32	Africa	Ghana	29.05743697***	0.903683819	181
33		Algeria	11.27988471	0.960159451	274
34		Morocco	29.06461136***	0.968584732	267
35		South Africa	8.312736747	0.978830758	243
		Africa total	34.5215878***	0.98852492	1442
36	Oceania	Australia	75.90339085***	0.831270094	315
37		New Zealand	39.20189084***	0.957872161	201
		Oceania total	64.51813121***	0.928472377	516
38		Canada	36.18100079***	0.92396168	286
39	North American	United States	16.07365809**	0.991279505	352
40		Mexico	11.93220996	0.989554882	261
North American total			32.56662938***	0.989243584	899
41		Brazil	9.216691091	0.975068671	273
42		Peru	2.821283015	0.992126079	250
43		Chile	5.813290849	0.985017554	258
44		Uruguay	26.28215766***	0.877629613	189
45		Argentina	4.449166081	0.99135785	277
46	South American	Paraguay	10.64363861	0.963005175	199
47		Bolivia	10.25275801	0.99163084	230
48		Ecuado	21.11037969***	0.950083539	249
49		Panama	13.83905462*	0.962510805	261
50		Dominica	6.51161095	0.954339362	244
51		Colombia	7.43952599	0.990654312	253
South American total			9.619290722	0.997601936	2683
Southern Hemisphere			32.43114605***	0.990929997	2684
North Hemisphere			61.73386398***	0.996882381	11053
ALL TOTAL			50.10053032***	0.931248322	484

注：*、**、***分别表示具有10%、5%和1%显著性。

表8 根据相关系数进行判断的分级标准 许存兴（2009）**错误!未找到引用源。**韩兆洲,程学伟（2019）**错误!未找到引用源。**

表8 相关系数进行判断的分级标准。

分级	相关系数分级标准	说明
正常	0.99<R≤1	完全符合Benford定律
关注	0.97<R≤0.99	存在一定因素的人为因素影响的可能
可疑	R≤0.97	人为影响因素较大，需特别注意

注：作为政府公布的宏观数据R小于0.9特别标出

具体分析如下：

1.按照国家和地区分析统计数据的差异。

按照卡方值分类成符合Benford's Law程度很好的国家或者地区、符合程度一般的国家或者地区和符合程度欠佳的国家和地区。

从上述数据，也就是根据2020年5月19日前从COVID-19数据，从各个国家和地区分析，亚洲有伊朗、韩国和马来西亚；欧洲有爱尔兰、比利时、葡萄牙、西班牙、德国和挪威；非洲有加纳和摩洛哥；大洋洲有澳大利亚和新西兰；美洲有加拿大、厄瓜多尔和乌拉圭个16国家和全球合计数据首位数字的统计频率与Benford's Law期望频率有显著性差异，其他国家数据与Benford's Law期望频率没有显著性差异。也就是说上述国家的COVID-19数据首位数频率偏离Benford's Law期望频率大，受人为影响因素较大，可能是因为有人为因素的控制，使COVID-19疫情得到有效控制，从而出现与Benford's Law期望值出现偏离。

从相关系数R判断，其中有西班牙、英国、挪威、澳大利亚和乌拉圭6个国家的数据与Benford's Law期望频率之间的相关系数小于0.90，受人为因素影响较大，R值最小的是挪威为0.810540441，与Benford's Law期望频率之间符合程度较高，与Benford's Law期望频率之间符合程度最好的国家（R值大于0.97）的国家有印度、以色列、阿联酋、巴基斯坦、日本、新加坡、菲律宾、德国、法国、埃及、尼日利亚、南非、美国、墨西哥、巴西、智利、秘鲁、阿根廷、玻利维亚和哥伦比亚等20个国家。

综上，检验假设H1,从R值判断，所有51个国家和全球汇总数据符合Benford's Law期望频率的程度都大于0.81，但国家之间数据受人为因素影响程度是不一样的，伊朗、韩国、马来西亚、爱尔兰、比利时、葡萄牙、西班牙、德国、挪威、加纳、摩洛哥、澳大利亚、新西兰、加拿大、厄瓜多尔和乌拉圭个16国家和全球合计数据首位数字的统计频率与Benford's Law期望频率有显著性差异，即这16个国家和全球汇总数据受人为因素影响比其他国家的影响要大。

2.按照亚洲、欧洲、大洋洲、非洲等分类，分析COVID-19数据符合Benford's Law期望频率有无明显的差异，结果参见表9。

表9 各大洲COVID-19数据Benford's Law卡方值。

	χ^2	R
亚洲	35.71782062***	0.993016713
欧洲	22.92793713***	0.993277732
北美洲	32.56662938***	0.989243584
南美洲	9.619290722	0.997601936
非洲	34.5215878***	0.98852492
大洋洲	64.51813121***	0.928472377
全球	50.10053032***	0.931248322

注：*、**、***分别表示具有10%、5%和1%显著性。

将上述国家和地区按照各大洲出现分类计算统计检验数据，先从相关系数R值判断，各大洲数据与Benford's Law期望频率符合程度超过0.92，各大洲都符合Benford's Law；从拟合优度值（卡方χ^2）如上表，南美洲数据很好的符合Benford's Law期望频率，并没有超过临界值20.09，说明南美洲COVID-19数据受人为因素较小，其他洲COVID-19数据有一定程度的人为因素影响。

3.按照南、北半球（气候影响）分类，分析COVID-19数据符合Benford's Law期望频率有无明显的差异，结果参见表10。

表10 南、北半球COVID-19数据Benford's Law卡方值。

	χ^2	R
南半球	32.43114605***	0.990929997
北半球	61.73386398***	0.996882381

注：*、**、***分别表示具有10%、5%和1%显著性

从相关关系来看，北半球COVID-19病例数据首位数字频率与Benford's Law期望频率之间存在依赖关系，R很好的符合Benford's Law统计期望频率，但南北半球COVID-19病例数据首位数字频率的卡方值32.43和61.73均超过临界值20.09，存在人为因素影响较大，可能疫情发展必须有严格的人为干预，才能很好的控制疫情。

4.按照第一、二、三阶段分类，分析COVID-19数据符合Benford's Law期望频率有无明显的差异，结果参见表6。

表11 第一、二、三阶段COVID-19数据检验表。

	χ^2	R
第一阶段1.16-3.6	24.1667724***	0.996565547
第二阶段3.1-4.18	8.522832758	0.998873825
第三阶段4.19-5.19	79.4512243***	0.993023269

注：*、**、***分别表示具有10%、5%和1%显著性。

第二阶段COVID-19病例数据首位数字频率与Benford's Law期望频率之间存在依赖关系，很好的符合Benford's Law统计期望频率，相关信息是可靠和真实的，但第一、三阶段COVID-19病例数据首位数字频率的卡方值24.1667724、79.4512243均超过临界值20.09，存在人为因素影响较大。

总之，检验假设H1,从R值判断，所有51个国家和全球汇总数据符合Benford's Law期望频率的程度都大于0.81，但国家之间COVID-19数据受人为因素影响程度是不一样的，伊朗、韩国、马来西亚、爱尔兰、比利时、葡萄牙、西班牙、德国、挪威、加纳、摩洛哥、澳大利亚、新西兰、加拿大、厄瓜多尔和乌拉圭个16国家和全球合计数据首位数字的统计频率与Benford's Law期望频率有显著性差异，即这16个国家和全球汇总数据受人为因素影响比其他国家的影响要大。在第一阶段只有中国湖北数据的卡方（χ^2）值超过临界值，第二阶段有巴基斯坦、澳大利亚和全球汇总数据的卡方（χ^2）值大于临界值，其他国家和地区的卡方（χ^2）值均未超过临界值，第三阶段有16个国家卡方（χ^2）值超过临界值，也就是从第一阶段开始，人为干预疫情的国家越来越多，即疫情需要严格的人为干预，才能较好的控制疫情。

检验假设H2,除南美洲外各大洲都在人为的努力干预COVID-19的疫情传播。

检验假设H3，南、北半球统计检验数据卡方（χ^2）值均超过临界值，人为因素影响原有没有从差异；从南、北半球看全球都在人为的努力干预COVID-19的疫情传播。

H4假设检验，第一、三阶段数据与第二阶段数据在人为因素因素方面应该有较为明显的差异；第一、三阶段数据受人为因素影响较大。

5．预测

表12 偏离benford定律的国家。

Nation	^{2}	R	Size
Iran	36.26265044***	0.912300286	357
South Korea	38.70062533***	0.953798374	392
Spain	33.06514598***	0.87130126	290
Germany	21.21102419***	0.970528555	250
Norway	49.65350845***	0.810540441	220
Ghana	29.05743697***	0.903683819	181
Morocco	29.06461136***	0.968584732	267
Australia	75.90339085***	0.831270094	315
New Zealand	39.20189084***	0.957872161	201
Oceania total	64.51813121***	0.928472377	516
Canada	36.18100079***	0.92396168	286
Uruguay	26.28215766***	0.877629613	189

注：*、**、***分别表示具有10%、5%和1%显著性。

我们通过对全球多个国家每日新增确诊病例人数分析，分析大部分国家的疫情经过三个发展阶段，第一阶段爆发期，相关疫情数据会很好的符合Benford定律，每日确诊人数呈现加速发展态势，这一阶段人为影响因素较小；第二阶段稳定期，也就是经过多方面的医学管控和治疗，每日新增确诊人数增幅逐渐放慢，这段期间主要受人为因素管控，相关数据也开始偏离Benford定律，如下表12；第三阶段结束期，这一阶段主要是从医学管控的角度，在没有全面疫苗预防的情况下，采取严格的人为管控来减少疫情的蔓延。在这一阶段，统计相关数据又开始较好的符合Benford定律。

当然我们也注意到，各国每天新增确诊病例拐点和峰值，可能是过个一个拐点和峰值，会出现另一个一个拐点和峰值，未来还可能出现多个拐点和峰值，也就是疫情会出现多次反复。很难用某一个模型做比较准确的预测，后面我们可能从符合一定条件的情况下，用多个模型来分析探讨疫情的发展。

6．结论与局限

6.1．主要结论

本研究利用Benford's law分析检验COVID-19数据的可靠性，是将Benford定律 law应用到公共卫生管理领域，这是本研究的贡献和创新点。从这次研究数据，我们判定中国疾病预防控制中心官网公布的新型冠状病毒肺炎数据基本上是客观的、可靠，并得出如下结论。

1.第一阶段的其他地区COVID-19相关数据经过R、拟合优度χ^2、V_N^*、d^*、m^*检验等检验，基本不符合Benford's Law期望频率值，受人为因素影响很大，数据基本上不值得信赖；第二阶段澳大利亚、巴基斯坦和全球数据，偏离Benford's Law程度较大，受一定人为影响的因素影响；第三阶段有16个国家和全球汇总数据的χ^2超过临界值。

2.从各大洲看，第二阶段非洲和大洋洲的数据Benford's Law超过临界值，即非洲和大洋洲数据受人为因素影响较大，非洲和大洋洲数据与其他洲数据具有显著性差异。第三阶段，只有南美洲数据的统计检验χ^2没有超过临界值，也就是其他洲数据受人为因素影响较大。

3.从南、北半球看，第二阶段南半球COVID-19数据，超过临界值，可能受人为因素影响或者受气候因素影响。南北半球之间COVID-19数据具有显著性差异。第三阶段，北、南半球的统计检验量χ^2值都超过临界值。受人为因素影响均较大。

4.从第一、二、三阶段数据分析分析，第一、三阶段数据受人为影响因素较大。第一、三阶段数据和第二阶段数据也具有显著性差异。

5、从疫情预测方面分析，大部分符合benford 定律的地方疫情都较严重，而偏离benford定律的地方，疫情得到慢慢的控制，由此得出人为的管控可能是目前控制疫情的最好办法。

6.2．研究局限性

本文中提及的"其他地区"是指在公海上不属于任何国家或者地区的COVID-19相关数据的统计量，当然，COVID-19相关数据也可能存在统计各国统计数据受到标准、所处地理位置、气候、医疗技术水平和经济发达情况等多因素影响的情况。

本文使用的分析数据的收集始于2020年1月16，截止2020年5月19日，当有足够的新数据出现后，还需要做进一步分析和探讨。

致谢

基金资助：广东省教育厅2018年广东省高等教育教学改革项目（2018407）（(ACCA)协同育人产教融合相关研究与实践）。Higher Education Teaching Reform Project of Guangdong Province, 2018 (2018407) ((ACCA) Research and Practice on Collaborative Education Integration of People, Industry and Education)。

广东省哲学社会科学规划"建设粤港澳大湾区"和"支持深圳建设中国特色社会主义先行示范区"专项（GD20SQ20）（粤港澳大湾区价值观认同与经济辐射效应研究）。Philosophy and Social Science's Special Project Planning of Guangdong Province in 2020: Research on Values Identity and Economic Radiation Effect in Guangdong-Hong Kong-Macao Greater Bay Area, GD20SQ20。

广东省教育科学"十三五"规划2019年度高校哲学社会科学专项研究（2019GXJK101）（粤港澳大湾区经济辐射效应及作用机理）。Guangdong Province Educational Science "13th Five-Year Plan" Special Research on Philosophy and Social Sciences in Colleges and Universities in 2019:Economic Radiation Effect and Mechanism of Guangdong-Hong Kong-Macao Greater Bay Area,2019GXJK101。

参考文献

[1] Collins, C. (2017). Using Excel and Benford's Law to Detect Fraud. Journal of Accountancy, April.

[2] Crocetti, E., & Randi, G. (2016). Using the Benford's Law as a first step to assess the quality of the cancer registry data. Frontiers in public health, 4, 225.

[3] Daniels, J., Caetano, S. J., Huyer, D., Stephen, A., Fernandes, J., Lytwyn, A., & Hoppe, F. M. (2017). Benford's law for quality assurance of manner of death counts in small and large databases. Journal of forensic sciences, 62(5), 1326-1331.

[4] Goodman, W. (2016). The promises and pitfalls of Benford's law. Significance, 13(3), 38-41.

[5] Idrovo, A. J., Fernández-Niño, J. A., Bojórquez-Chapela, I., & Moreno-Montoya, J. (2011). Performance of public health surveillance systems during the influenza A (H1N1) pandemic in the Americas: testing a new method based on Benford's Law. Epidemiology & Infection, 139(12), 1827-1834.

[6] Kuruppu, N., & Muscat, O. (2019). The Application of Benford's Law in Fraud Detection: A Systematic Methodology. International Business Research, 12(10), 1-10.

[7] Lu, F., & Boritz, J. E. (2005). Detecting fraud in health insurance data: Learning to model incomplete Benford's Law distributions. In European Conference on Machine Learning (pp. 633-640). Springer, Berlin, Heidelberg.

[8] Manrique-Hernandez, E. F., Fernandez-Nino, J. A., & Idrovo, A. J. (2017). Global performance of epidemiologic surveillance of Zika virus: rapid assessment of an ongoing epidemic. Public health, 143, 14-16.

[9] Nigrini, M. J. (2019). The Patterns of the Numbers used in Occupational Fraud Schemes. Managerial Auditing Journal, 34(5), 606-626.

[10] Pomykacz, M., Olmsted, C., & Tantinan, K. (2017). Benford's Law in Appraisal. The Appraisal Journal, Fall, 274-284.

[11] Sambridge, M., Tkalčić, H., & Jackson, A. (2010). Benford's law in the natural sciences. Geophysical research letters, 37(22).

[12] Singleton, T. W. (2011). Understanding and Applying Benford's Law. ISACA Journal, 3, 4.

[13] Zhang, J. (2020). Testing Case Number of Coronavirus Disease 2019 in China with Newcomb-Benford Law. arXiv preprint arXiv:2002.05695.

[14] 杜玉忠，张铭驱，范秀红，卢文涛，曾茜茜，黄燕，黄燕琼.基于向量自回归模型的清远市手足口病发病预测分析[J].实用预防医学，2019（2）：247:250.

[15] 韩兆洲,程学伟.GDP统计数据质量实证研究:基于Benford法则和空间面板模型[J]. 数理统计与管理,2019,38(03):394-404.

[16] 刘云霞,曾五一.关于综合利用Benford法则与其他方法评估统计数据质量的进一步研究[J].统计研究,2013,30(08):3-9.

[17] 罗玉波,张冬霞,吴亚炳.Benford's Law在财务审计中的实证研究[J]. 中国审计评论,2015（2）:69-78.

[18] 吕娜,邹薇.健康人力资本投资与居民收入——基于私人和公共部门健康支出的实证分析[J].中国地质大学学报(社会科学版),2015,15(01):113-119.

[19] 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）：56:60.