COVID-19 感染人数数据准确性的 Kuiper 检验方法

陈潇　　任博轩　　徐郑钰

指导老师：张鸿燕

海南师范大学信息科学技术学院

版本：0.16

日期：2022 年 5 月 9 日

摘　　要

作品以 Benford 定律为基础，利用多元线性回归模型建立 Kuiper 检验统计量与国家发展水平参数的统计模型，用以研究新型冠状病毒每日新增确诊数据的准确率与国家发展水平参数向量的关系。探究了中国疫苗上市及英国封城对疫情带来的影响。

关键词： COVID-19，Benford 定律，Kuiper 检验，多元线性模型，防控策略

1　引言

自 2020 年起，新冠疫情便不断肆虐，席卷全球，在这两年多的时间里，新冠病毒彻彻底底改变了人们的生活，以前是天天奔忙在自己的岗位上，而现在居家隔离自我防护则成了生活常态。为什么新冠能够困扰我们这么久？我们认为，新型冠状病毒本身的超强传播能力是一方面，相关数据的准确性也是不可忽视的原因之一。

首先，新冠数据的准确性最先影响的就是国家对国内疫情形势的判断[1-2]，国家政策也会跟随数据的波动而变更，政策实施有效与否将直接影响疫情导致的感染率、死亡率。

其次，媒体舆论也会受到新冠数据的影响，若新闻媒体得到了错误的数据，将这些错误数据反映到媒体报刊上，民众因此接收到错误的信息，就有可能引起不必要的恐慌，更有甚者或许会举行大规模的暴乱活动，这样不仅会急速增加新冠病毒的感染人数，对国家控制疫情更是有百害而无一利[1, 3]。

并且，也有人对世界上各个国家的新冠数据准确性提出了疑问[1, 4-5]，为了解答这个疑问，我们采用 Benford 定律[6] 以及 Kuiper 检验[7] 的方法，Kuiper 检验是一个适用于小样本的拟合优度检验方法；通过这二者结合，从而对数据的准确性进行描述。

2　Benford 定律

2.1　概念与解释

在日常生活中，以 “1” 为首位数字的数 (如 12、135、1083 的首位数字为 1) 的出现概率并不总是 $\frac{1}{9}$，而是 $\frac{1}{3}$ 左右。推广来说，越大的数，以它为首位数字出现的概率就越低。

在十进制首位数字的出现概率中，“1” 最高 (30.1%)，逐渐递减，一直到 “9” 最低 (4.6%)。这个规律可以写成数学公式的形式

$$P(d) = \log_{10}(1 + \frac{1}{d}), d \in (1, 2, \cdots, 9) \tag{1}$$

在实际生活中，有许多数都符合这个定律，包括全球各个国家的人口、全球各个国家的 GDP、国家国土面积等等。这些数据有一些共同的特征：**自然增长**、**不被人为所操控**。Benford 定律现已被广泛地应用于:

- 会计学[8]
- 科学研究[9-10]
- 宏观经济学[11-12]
- 法医分析[13]
- 税务分析[14]

并且已经有相关研究证明，传染病期间的数据报道[1, 15] 也符合 Benford 定律。因此，我们认为，在新型冠状病毒自然增长时期，每日新增确诊的数据符合 Benford 定律；而在各个国家开始对疫情进行大力度的管控之后，每日新增确诊的数据则不符合 Benford 定律。

2.2 数据分析与处理

我们收集了从 2020 年 1 月 22 日至 2022 年 3 月 11 日世界上 158 个国家新型冠状病毒每日新增确诊的数据，对这些国家逐一计算 Kuiper 检验统计量 V 的值，当 Kuiper 检验统计量 V 小于 2 时，我们认为它符合 Benford 分布，并且这个值越小，报道的数据就越准确；当 Kuiper 检验统计量 V 大于 2 时，我们认为它不符合 Benford 分布，而这其中有两种原因：

1. 该国的新冠数据报告准确率低，存在数据造假的可能。
2. 该国的防疫措施好，导致新冠新增不符合 Benford 定律适用的前提条件—**自然增长**。

为了判断是二者中的哪一种，我们选取了各个国家新冠新增确诊达到顶峰之前的数据。在一个国家的数据达到平稳前，若它的首位数字分布偏移了 Benford 分布，则认为它的数据存在造假的可能性；而在达到平稳期之后，若仍符合 Benford 分布，则表明该国的防疫措施不够恰当，没能有效地控制住疫情。

为了判断数据是否达到一个顶峰，我们使用了经济学中常用的方法：移动平均线。移动平均线是指一段窗口期内的算术平均线，并且这个窗口是滚动向前的。我们选取 7 天作为这个窗口期，每日新增确诊作为数据，取移动平均线的最大值的日期作为增长期的结束日。

我们使用国家级的数据（各国每日新增数据），计算 158 个国家新冠新增达到平稳之前的 Kuiper 统计量 V，为了直观地体现各个国家的统计量大小，我们绘制了图 1 [1]，颜色从青色到红色，统计量依次增大，即对 Benford 分布的偏移程度越大：

[1]灰色地区的数据因数据缺失等没有被选用

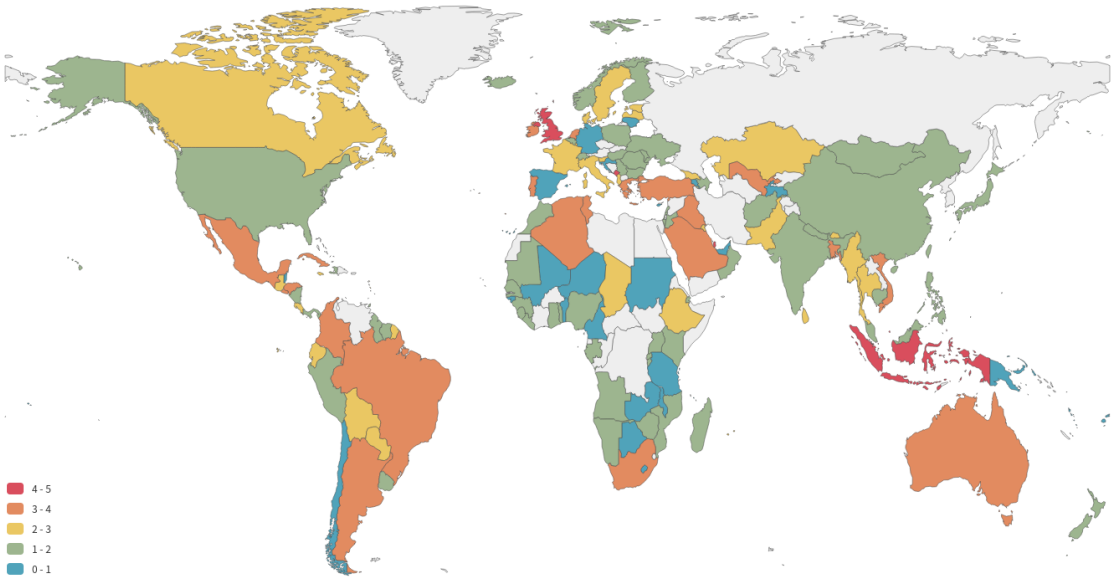

图 1: 全球部分国家的 Kuiper 检验统计量 V

为了更明确地体现单个国家对 Benford 分布的拟合程度，对他们的地区级数据进行分析，绘制了具体的柱状图。

2.2.1 对中国新增确诊人数的分析

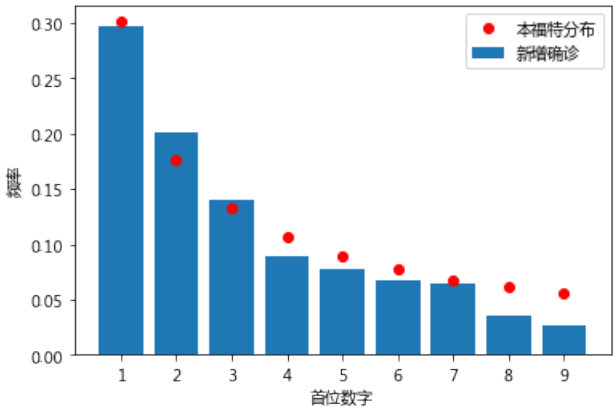

图 2: 中国 2020 年 1 月 22 日至 2020 年 2 月 14 日新增确诊首位数字分布，$V = 1.06799$

从图 2 可以直观地看出，疫情爆发初期的每日新增确诊和 Benford 分布较为拟合；从统计量可以看出，在中国疫情爆发的初期，有 99% 的把握说明中国每日新增确诊符合 Benford 分布，并且是一个较为准确的数据。

但如果把时间范围扩大，拉长到 2020 年整年的数据，我们发现全年的数据就不符合 Benford 分布了，“1” 作为首位数字的概率高了 10% 左右，统计量 V 为 6.98094，远大于 2，可以认为偏离 Benford 分布。

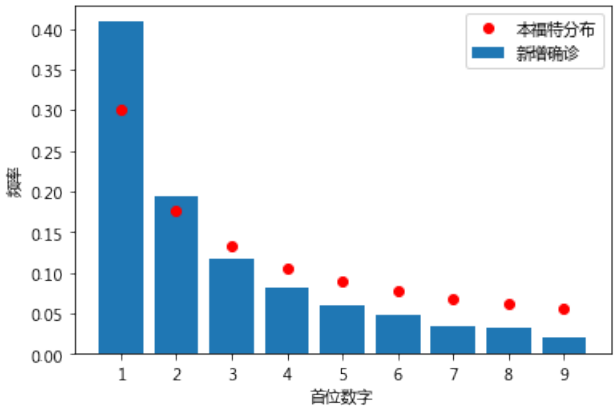

图 3: 中国 2020 年全年新增确诊首位数字分布，$V = 6.98094$

我们认为，全年的数据不符合 Benford 定律的原因是我国的疫情防控策略卓有成效，导致不符合 Benford 定律**“自然增长”**的条件，所以对 Benford 分布产生了偏移。而当每日新增确诊人数自然增长的时候，我国的数据符合 Benford 分布，可以说明我国在数据报道方面没有伪造、欺骗。

2.2.2 对英国新增确诊人数的分析

由于英国疫情爆发的时间较中国来的晚，爆发期相对延长，因此增长期也有所不同。我们选取了 2020 年 1 月 22 日至 2020 年 4 月 13 日作为增长期，2020 年 4 月 13 日为英国首相约翰逊实施群体免疫满一个月的日期。

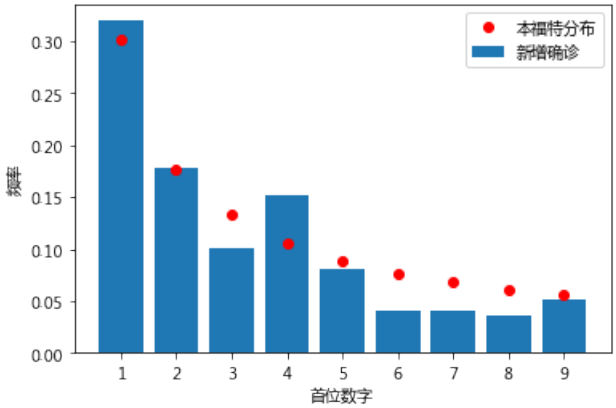

图 4: 英国 2020 年 1 月 22 日至 2020 年 4 月 13 日新增确诊首位数字分布，$V = 0.83826$

计算得到，英国这段时间内的统计量相较中国更小，我们认为，在这段时间内，英国的数据比中国更准确。客观来说，英国的医疗水平比中国更高，疫情爆发的时间晚，数据比中国更准确也是正常的。

对于英国 2020 年全年的新冠新增确诊数据，图 5 表现出了和中国 2020 年全年类似的趋势，也是“1”作为首位数字的概率偏高了 10% 左右，统计量也处在置信区间外。

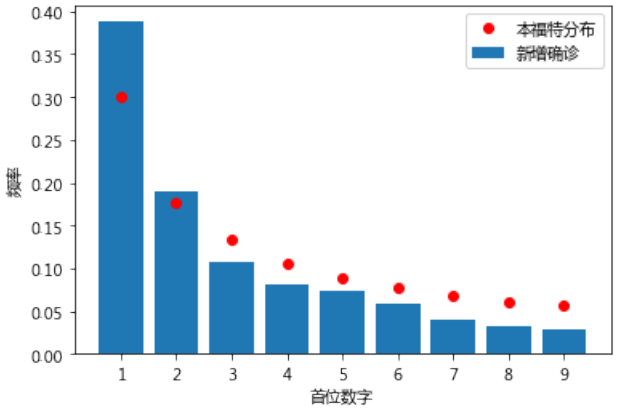

图 5: 英国 2020 年全年新增确诊首位数字分布，$V = 3.50904$

就中国与英国的数据分析对比而言，我们提出了一个疑问：国家的发达程度和数据的准确性是否关联?

3 新冠数据准确性是否与国家发展挂钩

为了分析国家发展程度与数据报告准确性之间的关系，我们建立起一个定量描述的线性回归模型[16]。

线性回归是利用数理统计中的回归分析，来确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法之一，运用十分广泛。而最小二乘法是线性回归中最常用的方法。

但是，数据的准确性会受到较多方面的影响，我们将引入发展指数、人口总数的自然对数值（使得序列更加平稳）、非零新增病例的天数，期望从统计学的角度找出这些指标与数据准确之间的关联。

3.1 构建发展指数

发展指数: 发展指数的构建我们用了三个比较主要的经济指标:

1. 人均国内生产总值 (Gross Domestic Product per capita)
2. GDP 中的医疗占比 (Healthcare Expenditures as a percentage of GDP)
3. 全民健康覆盖指数 (Universal Health Coverage Index)

这三个指标都是较为由于三个指标都是不同量级的数据，我们通过对每个经济指标采用四分位数排名 (quartile rank) 的方式 (从 0 到 3，0 是最低位，3 是最高位) 进行处理。各个国家的发展指数就是三个经济指标的四分位分值的和。

这么处理之后，我们将这 158 个国家的发展程度分为 10 个等级，从 0 到 9，0 为最低，9 为最高。通过这个指标，来描述一个国家大致的发展水平。

下面是筛选过后全世界各个国家发展指数图[2]，颜色从青色到红色，发展指数依次增大。

[2]灰色地区的数据因数据缺失等没有被选用

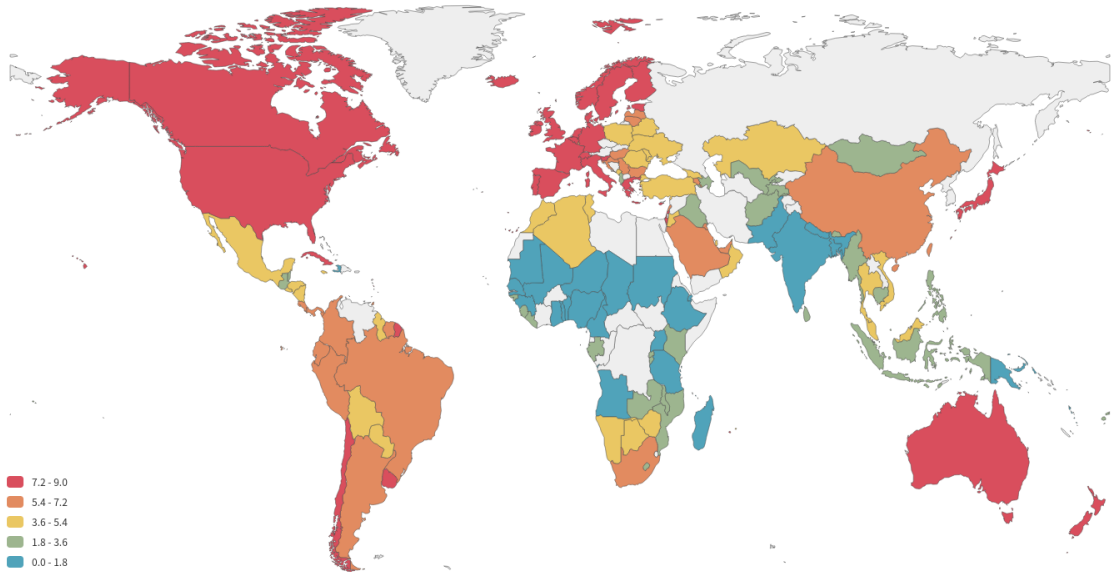

图 6: 全球部分国家的发展指数

从图 6 中可以看出构建的发展指数能够反映出国家整体大致的发展状况[3]。发达国家基本上为红色，而对于一些比较落后的国家颜色大部分为青色或偏绿色。

3.2 回归模型的创建

我们使用最小二乘线性多元回归模型研究因变量 (检验量) 和多个自变量 (发展指数、人口总数的自然对数值、非零新增确诊的天数) 之间的关系。进行回归后我们得到了表 1 中的结果。

表 1: 回归分析结果

	相关系数	标准误	t	$P > \|t\|$
发展指数	0.0703	0.035	1.985	0.049
$\ln(population)$	0.1112	0.041	2.679	0.008
非零新增确诊的天数	0.0020	0.001	4.059	0.000

最终的结果与我们的设想大相径庭，据回归的结果，发展指数、人口都与统计量呈显著正相关，这表明国家越发达，统计量越高，即数据越不准确。但是，当我们把时间缩短到 2020 年 6 月，发展指数又与统计量呈显著负相关.

结合各国对新冠疫情的态度，我们可以认为发达国家熬过疫情最初的爆发时期之后，对新冠病毒检验有所放松。当然，也有可能是因为发达国家的防疫措施做得好，但结合实际的感染人数来看，我们认为这个可能性很小。

[3]需要注意的是，这个指数无法准确地描述一个国家的发展程度，只能大概描述一个国家相较其他国家的发展水平。

从这个结果可以看出，目前的情况下，我们还是需要保持对外来入境的严加管控，不可掉以轻心。

4 部分防疫措施带来的影响

4.1 中国—疫苗上市

2020 年 12 月 31 日，中国首款新冠疫苗正式上市。在 2021 年 3 月 23 日，国家卫健委首次公布了我国疫苗接种数据：各省总计接种 8284.6 万剂。在 2021 年 4 月 23 日，我国接种剂次已达 21608.4 万剂次，2021 年 5 月 23 日，我国接种剂次已达 49727.2 万剂次。我们分别计算了由疫苗上市到首次公布数据 (2021 年 3 月 23 日)，以及接着两个月每个月的统计量变化情况。

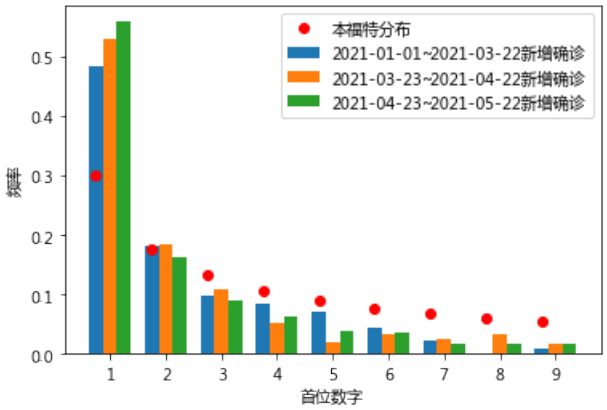

图 7: 2021 年 1 月 1 日至 2021 年 5 月 22 日新增确诊首位数字分布情况

从图 7 中可以直观地看出，随着疫苗接种人数增加，新增确诊对 Benford 定律的偏移程度也在增加。这三段时期的检验统计量 V 分别为2.82700、3.79609、4.15556，是逐渐上升的。当然，我们无法将统计量的增加完全归功于疫苗接种，但这也表明了随着时间推移，我们的数据也越来越偏移 Benford 分布，即我们的防控措施是在逐渐升级、逐渐生效的。

4.2 英国—群体免疫

为了对比，我们选取了英国在疫情防控上的几个时间节点，由于我们没能找到英国的疫苗接种数据的准确来源，我们选取了英国第一次封城时的每日新增数据，挖掘英国在封城防疫措施下的统计量变化。

2020 年 3 月 12 日，英国英国约翰逊政府宣布要采取“群体免疫”策略应对疫情，导致疫情形势迅速恶化，2020 年 3 月 22 日，英国被迫全国封城。在这期间，让我们来看看英国每日新增对 Benford 定律的拟合程度，结果如图 8所示：

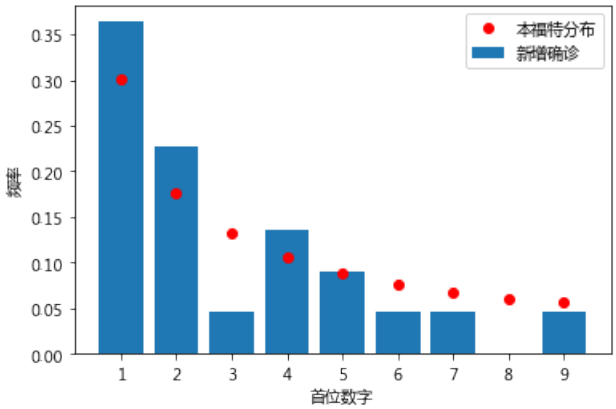

图 8: 英国宣布群体免疫到封国期间的首位数字分布情况，$V = 0.55821$

由于样本量较小，从图像上看并不明显，但 Kuiper 检验量仅为 0.55，从统计的角度上说，我们有 99% 以上的把握说在英国宣布群体免疫到封城这一段期间，会符合 Benford 分布。

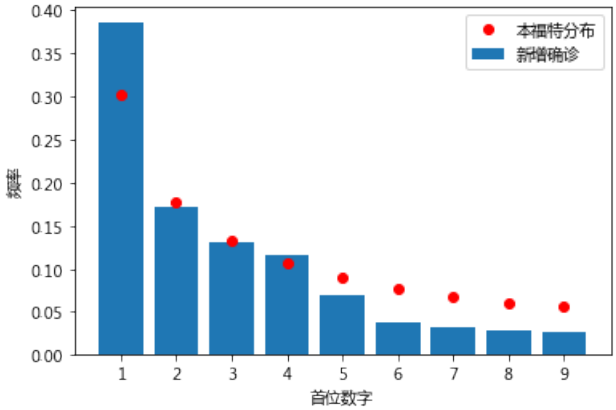

图 9: 英国由封国到 6 月 15 日，$V = 1.98095$

采取了封城的政策后，英国的 Kuiper 检验统计量明显上升了，如图 9 所示，已经接近不拟合的边缘。考虑到民众对封国的反感以及封国对经济的破坏力，英国在第一次封国期间并没有十分严格的措施，得出这个结果也很正常的。

4.3 总结

从两国对比可以看出，中国在 3 月 23 日时就已经偏移了 Benford 分布，而英国直到 6 月 15 日仍符合 Benford 分布，这表明中国在 1 月到 3 月短短两个月的时间内就比较有效地控制住了疫情，而英国的“群体免疫”导致了更多的感染者，直到 6 月新增确诊人数仍在呈上升趋势。这表明了我国人民在疫情的初期抗击疫情的决心，但是，最近上海疫情的严重程度，已经超越了疫情爆发初期，这无疑是对我国疫情防控敲响了警钟。我们仍需要保持像疫情初期那般对疫情的警惕，切不可放松防控政策，要通过有效防控措施的落实及时控制疫情、减少疫情对人们生命健康的威胁。对于防控的基本措施一定要有，比如要在人群密集的地方戴口罩；各景区、商场和餐厅也要控制好人群密度，让人们之间保持一

定的社交距离。

5 结尾

新冠疫情的爆发，让我们与外界的接触更加困难。入境“7+14”的政策无疑是外国友人来华的一道门槛，当下严格的入境政策会造成我国大量经济损失，随着时间的推移，逐渐改变入境政策是世界趋势的必然选择，根据各国疫情数据、趋势对各国入境人口进行动态调控，实现“动态清零”才是真正以人为本。没有一个冬天不会过去，没有一个春天不会到来，坚持“动态清零”总方针不犹豫不动摇，坚持人民至上、生命至上不动摇，从严从实把党中央关于疫情防控的决策部署落实到每一个环节，才能同心协力尽快打赢这场大仗硬仗。

附录

A 数据来源

- 三个经济指标的数据来源：https://data.worldbank.org/
- 人口数据来源：https://www.worldometers.info/
- 国家级的新冠数据来源：https://www.owid.de/
- 地区级的新冠数据来源：https://github.com/CSSEGISandData/COVID-19

B Kuiper 检验

Kuiper 检验是一个改进版的 Kolomogorov-Smirnov 检验，它更适合用于小样本的拟合优度检验。

$$V_{\text{Kuiper}} = (D^+ + D^-)[\sqrt{N} + 0.155 + \frac{0.24}{\sqrt{N}}] \tag{2}$$

其中，

$$D^+ = \sup_{-\infty<x<+\infty}[F_o(x) - F_e(x)], D^- = \sup_{-\infty<x<+\infty}[F_e(x) - F_o(x)], \tag{3}$$

其中 $F_o(x)$ 表示经验分布的累积密度函数，$F_e(x)$ 表示 Benford 分布的累积密度函数，N 为样本大小。

表 2: Kuiper 检验表

α	0.10	0.05	0.01
V	2.00	1.75	1.65

C 多元线性回归模型及其参数的最小二乘估计方法

在回归分析中，一种现象常常是与多个因素相联系的，由多个自变量的最优组合共同来预测或估计因变量，比只用一个自变量进行估计更有效、更符合实际。我们的回归模型如下：

$$V = \beta_0 + \beta_1 I + \beta_2 P + \beta_3 D + \varepsilon \tag{4}$$

其中 V 是被解释变量，I、P、D 是解释变量，分别代表着发展指数，人口的自然对数，非零感染的天数；$(\beta_0, \cdots, \beta_3)$ 是相关系数，常数项 ε 是与 V 相关但没有添加到回归模型的变量。

最小二乘法[17](Ordinary Least Square, OLS)：给定 $m \times 1$ 数据向量 $\boldsymbol{b}$ 和 $m \times n$ 矩阵 $\boldsymbol{A}$，当矩阵方程超定，即 $m > n$ 时，我们可以使用这种求解准则：使误差的平方和

$$\phi = \|\Delta \boldsymbol{b}\|^2 = (\Delta \boldsymbol{b})^{\mathrm{T}} \Delta \boldsymbol{b} = (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^{\mathrm{T}} (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$$

最小。这样得到的解 $\boldsymbol{x}$ 称为最小二乘解。最小二乘方法等价于

$$\text{在条件} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} + \Delta \boldsymbol{b} \text{约束下，} \min_{\boldsymbol{x}, \Delta \boldsymbol{b}} \|\Delta \boldsymbol{b}\|$$

当 $\operatorname{rank}(\boldsymbol{A}) = n$ 时，方程有唯一解

$$x = \left(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{b}$$

参考文献

[1] KOCH C, OKAMURA K. Benford’s law and COVID-19 reporting[J]. Economics letters, 2020, 196: 109573.

[2] ADAM A, TSARSITALIDOU S. Data misreporting during the COVID19 crisis: The role of political institutions[J]. Economics Letters, 2022, 213: 110348.

[3] BUNKER D. Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic[J/OL]. International Journal of Information Management, 2020, 55: 102201. https://www.sciencedirect.com/science/article/pii/S0268401220311555. DOI: https://doi.org/10.1016/j.ijinfomgt.2020.102201.

[4] FARHADI N. Can we rely on COVID-19 data? An assessment of data from over 200 countries worldwide[J]. Science Progress, 2021, 104(2): 00368504211021232.

[5] IDROVO A J, MANRIQUE-HERNÁNDEZ E F. Data Quality of Chinese Surveillance of COVID-19: Objective Analysis Based on WHO’s Situation Reports[J]. Asia Pacific Journal of Public Health, 2020, 32(4): 165-167.

[6] BENFORD F. The law of anomalous numbers[J]. Proceedings of the American philosophical society, 1938: 551-572.

[7] KUIPER N H. Tests concerning random points on a circle[C]//Nederl. Akad. Wetensch. Proc. Ser. A: vol. 63: 1. [S.l. : s.n.], 1960: 38-47.

[8] HORTON J, KRISHNAKUMAR D, WOOD A. Detecting academic fraud in accounting research: The case of Professor James Hunton[J]. Available at SSRN, 2018, 3164961.

[9] JUDGE G, SCHECHTER L. Detecting problems in survey data using Benford’s Law[J]. Journal of Human Resources, 2009, 44(1): 1-24.

[10] DIEKMANN A. Not the First Digit! Using Benford€s Law to Detect Fraudulent Scientific Data[J]. Others, 2005.

[11] NYE J, MOUL C. The political economy of numbers: on the application of Benford's law to international macroeconomic statistics[J]. The BE Journal of Macroeconomics, 2007, 7(1).

[12] GONZALEZ-GARCIA J, PASTOR G. Benford's Law and Macroeconomic Data Quality[J]. Social Science Electronic Publishing,

[13] PINILLA J, LÓPEZ-VALCÁRCEL B G, GONZÁLEZ-MARTEL C, et al. Pinocchio testing in the forensic analysis of waiting lists: using public waiting list data from Finland and Spain for testing Newcomb-Benford' s Law[J]. BMJ open, 2018, 8(5): e022079.

[14] NIGRINI M J. A taxpayer compliance application of Benford's law[J]. The Journal of the American Taxation Association, 1996, 18(1): 72.

[15] IDROVO A, FERNÁNDEZ-NIÑO J, BOJÓRQUEZ-CHAPELA I, et al. Performance of public health surveillance systems during the influenza A (H1N1) pandemic in the Americas: testing a new method based on Benford's Law[J]. Epidemiology & Infection, 2011, 139(12): 1827-1834.

[16] BALASHOV V S, YAN Y, ZHU X. Using the Newcomb–Benford law to study the association between a country' s COVID-19 reporting accuracy and its development[J]. Scientific reports, 2021, 11(1): 1-11.

[17] 张贤达. 矩阵分析与应用[M]. 北京: 清华大学出版社, 2013: 403-404.