F-015

Benford の法則とテキストマイニングを融合した大学入試統計データ信憑性分析
Text Mining and Benford's Law for Multidimensional Numerical Matrix - a Case Study on University Admission Statistics

戸﨑 祐輔† 鈴木 孝彦† 峯 恒憲† 廣川 佐千男†
Yusuke Tozaki Takahiko Suzuki Tsunenori Mine Sachio Hirokawa

1. はじめに

現代社会では多くの行動が，その根拠となるデータに基づいて決定されており，データの信憑性は重要である．小規模データであれば個々のアイテムについて，一つずつ信憑性を調べられる．しかし，規模が大きくなれば困難となる．

自然な数値データ集合について，上位 1 桁目の数字 i の出現確率が$log_{10}\left(1+\frac{1}{i}\right)$となることが，Benford の法則として知られている[1]．図 1 に Benford の法則に従う分布の例を示す．棒グラフは県や郡，市区町村などの行政区画ごとに集計された 2017 年 1 月 1 日時点での人口[2]の数値データ（2272 個）について，上位 1 桁目の数字の分布を示したものである．

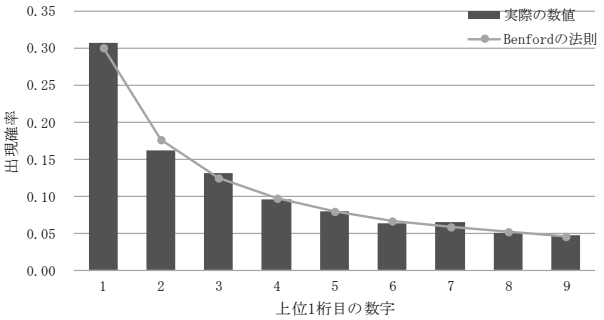

図 1　行政区画ごとの人口の上位 1 桁の分布

以下，Benford の法則に従わない分布を非 Benford 分布と呼称する．対象とする数値データ集合が非 Benford 分布となっていれば，なんらかの不整合があると考えられる．不整合の原因には，数値の改竄や，元の数値データそのものが不正であった場合などが考えられる．Nigrini ら[3]は会計データに対して，Benford の法則を適用し，不正を検出する方法を示した．一般に，Benford の法則を用いることで，数値データ集合の不整合を検出することができる．一つの数値データ集合について，Benford の法則に従うか否かの判定は容易にできる．しかし，そのデータが非 Benford 分布になっていることがわかっても，不整合箇所を特定することはできない．Suzuki ら[4]は，中国統計年鑑データについて，Benford の法則を用い，不整合箇所を特定することを試みた．

†九州大学

本研究では，統計データが 2 次元以上の表で構成される場合を想定し，数値を対象とする検索エンジンを構築した．いくつかの属性を制約条件として，数値データの部分集合を抽出し，その部分集合について検証することで，不整合箇所を特定する方法を提案する．日本の私立大学の志願者数，受験者数，合格者数を対象に，提案手法を適用し，不整合箇所があるかどうかを網羅的に調査した．

2. 信憑性分析

2.1 データの検索

表 1 は河合塾が提供した大学入試統計データ[5]の一部である．このデータは，各大学の公表資料をもとに作成されている．本研究では，このような表の数値データすべてを対象とする検索エンジンを構築した．大学，学部，学科，年度，データタイプなどを属性として，表の中の一つ一つの数値をデータベースに登録した．学部，学科は入試方式別に登録している．データタイプは，志願者数，受験者数，合格者数の 3 種類である．この表を含むページに現れるほかの単語もデータベースに登録する．検索条件を与えることで，その条件に合致する数値データの部分集合を抽出できる．

表 1　大学入試統計データの一部（2016 年度）

学部	学科	志願者	受験者	合格者
経済	経済前期 A 日程スタンダード	1618	1593	274
	経済前期 A 日程高得点	964	946	153
	経済前期 B 日程スタンダード	1216	1015	173
	経済前期 B 日程高得点	787	653	100

2.2 乖離度の検定

得られた数値データの部分集合について，Benford の法則に従うかどうかを検定する．標本分布が，理論分布と異なっているか否かの判定には，一般にカイ二乗検定が使われる．しかしながら，本研究で扱うような，比較的少数の数値データ集合に対して，Benford の分布からの乖離を検定する場合，カイ二乗検定よりも適しているとされる検定手法が提案されている[6,7]．本研究では，Cho らの d 値について，有意水準 5%の検定を採用した[8]．上位 1 桁目の数字を i，その出現頻度をP_iとし，d 値を以下の式で求める．

$$d=\sqrt{\sum_{i=1}^{9}\left(P_i-log_{10}\left(1+\frac{1}{i}\right)\right)^2}$$

$d \geq 1.33$（$P \leq 0.05$）であれば，非 Benford 分布と判定する．

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

3. 実験

3.1 実験方法

前述の大学入試統計データを使い，2016 年度及び 2017 年度における日本の 565 校の私立大学の志願者数，受験者数，合格者数について，2，3 個の属性の組合せを網羅的に検索し，得られる数値データの部分集合について Benford の法則に従うかを機械的に検証する．あまりに複雑な条件だと，抽出できるセルの個数が少なくなりすぎ，そもそも，Benford の法則に従うかどうかを検証する意味がなくなる．

3.2 実験結果

図 2 に分析事例を一つ示す．棒グラフは[5]の web ページにおいて，ある大学を選択したときに表示される様々な属性単位で集計された 2 年分の志願者数，受験者数，合格者数の数値データ（171 個）について，上位 1 桁目の数字の分布を示したものである．この大学の数値データ集合は，非 Benford 分布になっている．

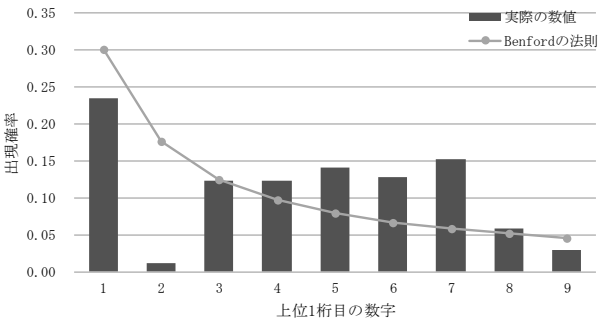

図 2　ある大学の志願者数などの上位 1 桁の分布

表 2 は検索条件ごとに，数値データ集合を抽出し，Benford の法則に従うかどうかを検定した結果である．2 行目は，図 1 の操作を大学ごとに行った結果である．収集した大学の数は 565 校だが，数値データの個数が 20 未満であった 84 校を除いた 481 校について調べ，その中の約 30%の 145 校について，非 Benford 分布と判定された．3 行目は，学部ごとの数値データ集合について，検定した結果である．20 個以上の数値データが得られた 1750 の学部について調べ，約 33%の学部が非 Benford 分布と判定された．各大学について，年度ごとに志願者数，受験者数，合格者数を絞り込んだデータについて，非 Benford 分布と判定されたものは 10%未満であった．

表 2　Benford の分布から乖離した集合の割合

抽出条件	全集合数	非 Benford 分布集合数	乖離率
大学	481	145	0.301
大学*学部	1750	583	0.333
大学*2016 年度*志願者数	223	16	0.072
大学*2016 年度*受験者数	202	11	0.054
大学*2016 年度*合格者数	199	9	0.045
大学*2017 年度*志願者数	217	17	0.078
大学*2017 年度*受験者数	199	16	0.080
大学*2017 年度*合格者数	198	4	0.020

4. まとめと今後の課題

本研究では，検索対象となる最小単位の数値データに対して多様な属性を付加し，いくつかの属性を制約条件として，分析対象データを絞り込み，Benford の法則に従うかどうかを機械的に判定した．この手法で非 Benford 分布の部分集合を特定し，単語や属性で示すことができる．このように，いろいろな切口でドリルダウンする分析は OLAP でもできる[9]．しかし，OLAP が対象にするのは定型的な構造化されたデータであり，データを規定するために使える次元は限定されている．

提案手法の問題点として，学科単位の募集定員の分散が小さければ，合格者数の上位 1 桁が非 Benford 分布になることが挙げられる．学科単位の定員が 30 名程度の募集であれば，上位 1 桁が 3 に偏る．このような偏りは，不整合と判断すべきではない．本研究では，信憑性分析において上位 1 桁目の数字しか使わなかった．Benford の法則は，上位 2 桁目の数字にも適用されるので，今後は上位 2 桁目の数字についても分析を行いたい．

5. おわりに

本研究では，大学入試統計データを使い，多次元の表として現れる数値データ集合の中から，不整合箇所を特定する方法を提案した．表の各数値について，行や列のタイトル，あるいは，表の説明文中の単語を対応づけることで，数値を対象とする検索エンジンを構築した．この検索エンジンを使うことで，複数の制約できまる数値データの部分集合を抽出でき，Benford の法則に従うかどうかも機械的に判定できる．

謝辞

本研究のきっかけとなったのは，北カロライナ州立大学 Stephen Porter 先生の DSIR2018 における講演であり，感謝します．

参考文献

[1] L. M. Leemis, B. W. Schmeiser, D. L. Evans, Survival Distributions Satisfying Benford's Law, The American Statistician, 54:4, pp. 236-241, 2000.
[2] 政府統計の総合窓口（e-Stat），【総計】市区町村別人口、人口動態及び世帯数, http://www.e-stat.go.jp/, アクセス 2018 年 6 月 25 日.
[3] M. J. Nigrini, Benford's Law Applications for Forensic Accounting, Auditing, and Fraud Detection, ISBN: 9781118152850, Wiley, 2012.
[4] T. Suzuki, T. Kamimasu, T. Nakatoh, S. Hirokawa, Identification of Unnatural Subsets in Statistical Data, Proc. AAI2018 (to appear).
[5] 河合塾，一般入試　入試結果（私立大学），http://www.keinet.ne.jp/dnj/result/ippan/s_index.html, アクセス 2017 年 11 月 6 日.
[6] W. K. T. Cho, B. J. Gaines, Breaking the (Benford) Law, The American Statistician, 61:3, pp. 218-223, 2007.
[7] C. A. Holz, The quality of China's GDP statistics, In China Economic Review, Volume 30, 2014, pp. 309-338, ISSN 1043-951X.
[8] J. Morrow, Benford's Law, Families of Distributions and a test basis, CEPDP1291, LSE Research Online, http://eprints.lse.ac.uk/60364/, 2010.
[9] Microsoft, Fraud analysis with SSAS: Benford's law test in OLAP Cubes, http://www.metrica-bi.de/fraud-analysis-with-ssas-benfords-law-test-in-olap- cubes/ Jun 19, 2015 (accessed Jan 12, 2017).

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.