新技術説明会
New Technology Presentation Meetings!

統計データの中の個々の数値の怪しさランキング法

九州大学 情報基盤研究開発センター
学術情報研究部門
教授　廣川佐千男

KYUSHU UNIVERSITY

どの数値が怪しいか？

25.5 2 56 22.5 18 51 39 10 110 802 37 150 264.5 1411.5 51
1442 364 93 14 153.5 65.5 890 2 112 4 5 46 27 516 36 15
32.5 33 2 25.5 2 22 23 40 405 173 331 90 107 155 91 143
144 132 121 126 194 199 712 239.5 172 89 106 91.5 82 133
125.5 155.5 244.5 138.5 92 106 288.5 167.5 93 93 102 94
103 140.5 112 77 95 99 103.5 269.5 74.5 99 115.5 107.5
109 108.5 108.5 2 2 2 37 18 11 3 73 8 1 2 16 10 9 6 14 5 14
14.5 14.5 7 12.5 12.5 10 30 4 2 33.5 4 28 29 3 1 34 5 3 2 61
78 35.5 124.5 36.5 26 2 50.5 2 53 16 10 7 8 3 10 12 15.5 17
19 13 24.5 8 1 27 5 8 9 9 2 20.5 56.5 6 57 26 9 7 10 15 8 13
7.5 15.5 15 14 2 10 9 9 24 9 9 9 27.5 7.5 3 33 12.5 12 1 14
9.5 2 14 9 10 13.5 9 1 29 9 637 181.5 194.5 233.5 174 165.5
245 347 266.5 284 571 503.5 960.5 518 265.5 136.5 137
124 133.5 243.5 252 349.5 698.5 260.5 185.5 183.5 640.5
483.5 151 122 111 136 239 227.5 181 117 142 197 130 329
142 189 176.5 163.5 170 262 221.5 6 47 9 27.5 16 4 7 20 53
38 29 26.5 22 15 116 24 16 33.5 23.5 18 13 80 114 27 8.5
94 26.5 32 9 26 5 3 9 4 3 2 227 3 43 17 5.5 115 6 133 3.5 9
17.5 1 6 1236 19 28 12.5 21 1 4 11 14 27 4 2 502 24.5 8
40.5 16 93 4 16 7 17 24.5 1 5 2 15.5 5 36.5 418 6 2 104 15
4 3 6 30 5 39.5 17 155 9 38 45.5 12 16.5 93.5 28 13 46 14
67.5 224 51 10 18 14 13 42.5 6 13 13 14 25 60 8 4 45 65 41
41 69 45 23.5 30 113 11 12.5 8 49 8 28 148.5 7 11 144 17 6
87.5 6 8 14 50 45 75 71 50 56 7 47 55 43 11 127 17 47 67.5
47 24 42 63.5 5 50 3 7 10 24 25 18 10 25 36 6 14 196.5 3 16
591 147


朝日新聞　２０１８年１０月２３日　朝刊　１ページ　東京本社

障害者雇用 退職者を算入

省庁、「うつ状態」「裸眼0.1以下」も

第三者委報告

中央省庁の障害者雇用数水増し問題で、原因を検証してきた第三者委員会（委員長＝松井巌・元福岡高検検事長）は22日、多くの行政機関で健常者の職員を恣意的な解釈で「障害者」と見なしてきたとする報告書を公表した。また、政府は全国の自治体で計３８０９・５人の不適切な障害者雇用数の算入があったとの再調査結果を発表。障害者雇用を牽引すべき行政機関で水増しが横行していた実態が改めて鮮明になった。

▼2面＝低い規範意識、35面＝怒りの声

自治体は3800人水増し

ともに、この日あった関係府省連絡会議に示された。障害者雇用率の算出方法に基づくと、中央省庁では昨年6月1日時点で３４４５・５人の水増しがあったが、第三者委の報告書によると、実数ベースでは28機関で計３７００人が不適切に算入されていた。

具体的な事例では、退職者など在籍していない職員が7機関で計91人算入されていた。このうち国土交通省では退職者74人、出向者7人を不適切に算入していた。死亡した人や約10年前に退職していた人も含まれていた。長年引き継がれた障害者職員の名簿の人について、在籍の有無を確認していなかったという。

最多の１１０３人を不適切に算入した国税庁は、

■報告書が指摘した水増しの手法

- 診断書や人事調書に「うつ状態」「適応障害の一歩手前」「不安障害」と記載があることを根拠に算入（国税庁）
- 障害者職員の引き継ぎ名簿に名前があるとして、退職者を算入（国土交通省）
- 障害者数に計上しながら、法定雇用率の分母となる職員数に算入せず（法務省）
- 眼鏡の使用やしぐさなどから視力が悪そうな者に裸眼視力を聴いて算入（農林水産省）
- うつ状態で病気休暇に入ったと診断書で確認できた人を算入（財務省）
- 採用時の健康診断で裸眼視力0.1以下の人を算入（総務省）

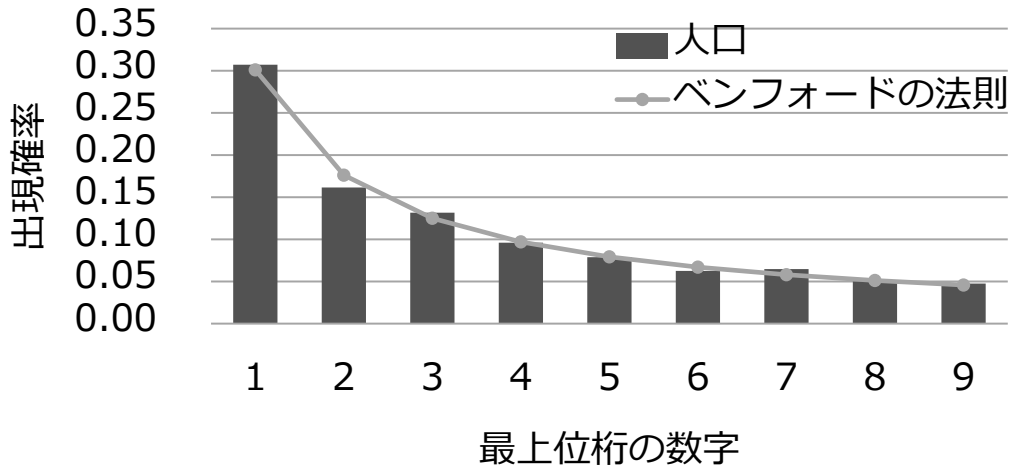
全体が怪しい → ベンフォード分布からの乖離（従来手法）

どの数値が怪しいか？ → 本研究

背景	統計データの信憑性は社会の基盤
問題	Q1 数値集合が怪しいかわからない。 → 信憑性概略評価：ベンフォードの分布と比較（先行研究）
	Q2 数値集合が怪しいとわかっても、その中でどの数値が怪しいかわからない。 **→ 信憑性詳細評価：k = 3, …, 16進数で考える（本研究）**
問題の例	障害者雇用水増し問題
提案手法	基数k=3,…,16で複数のベンフォード分布を使い、各数値の怪しさを定量化し、Q2を解決する。

ベンフォード則による全体評価（従来の手法）

自然な数値の集合はベンフォードの分布に従う



13
114
1159
1265
2
235
9897

最上位桁の数字dの出現確率は、

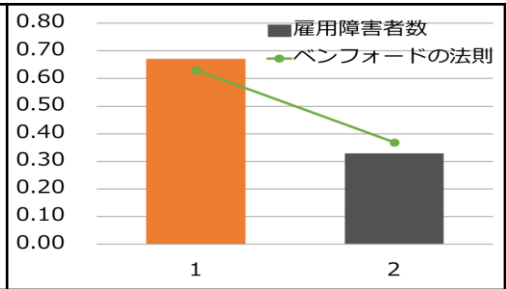
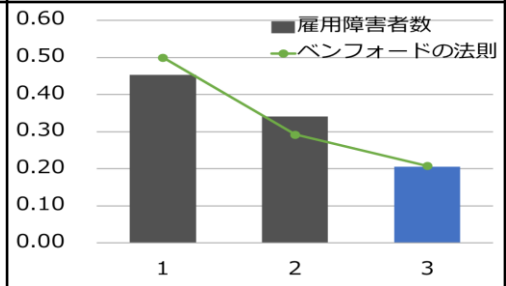
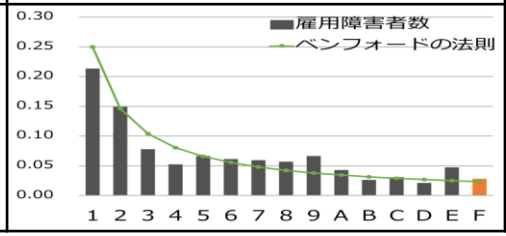
$$log_{10}\left(1+\frac{1}{d}\right)$$

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

ベンフォードの分布から乖離していれば、信憑性に疑問

245(＊県教育委員会）の怪しさ

Ｋ進法での評価を統合（本研究）

3進数		$100002_{(3)}$	怪しい
4進数		$3311_{(4)}$	怪しくない
⋮	⋮	⋮	⋮
16進数		$F5_{(16)}$	怪しい
		怪しさの合計	9

提案する怪しさランキングの妥当性評価

id	修正有	人数	怪しい/総基数	score	rank	precision@N
379	1	144.0	14/14	1.0000	1	0.7143
422	0	147.0	14/14	1.0000	1	0.7143
12	1	150.0	14/14	1.0000	1	0.7143
376	0				1	0.7143
20	1				1	0.7143
225	1				1	0.7143
48	1				1	0.7143
233	1				8	0.7826
49	1				8	0.7826
47	1				8	0.7826
334	1				8	0.7826
61	1				8	0.7826
45	1	155.0	13/14	0.9286	8	0.7826
237	0	142.0	13/14	0.9286	8	0.7826
⋮	⋮	⋮	⋮	⋮	⋮	⋮

推定でランク１位の数値の７割が実際に修正されたものだった

新技術の特徴・従来技術との比較

10進数で表示したときの最上位の数字の出現頻度はベンフォード分布になることが知られており、統計データの不自然さ評価に使われている。

しかし、全体データの中で違和感がある箇所があった場合、具体的にどの値がその違和感部分に関係しているかを判別することは困難である。

本技術は、個々の数値の不自然さを評価することができる。

想定される用途

- 統計データ検証(公的機関の集計データなど)
- 会計情報の検証
- センサーデータの異常検出と要因抽出

実用化に向けた課題

（1）リアルデータでの有効性評価

本技術をシステムとして構築しているが、理論的妥当性能は検証しているが、リアルデータでの有効性評価は、事例を積み上げていく必要がある。リアルデータのこの部分が改ざんされたものという実際のデータは、入手することができない。公的機関のデータの改ざんについては、社会的に問題となった結果公開され、我々の実験でも利用することができた。このように、実際にどの部分が改ざんされたかという情報が分っているデータでの評価実験を積んでいきたいが、そのようなデータは入手困難。

（2）データの“偏り“への対応

公表された大学入学受験者数のデータで、どの大学、どの学部、どの学科の数値の信憑性評価をやったことがあります。ところが、ある学科では定員が30数名に限定されているので、分布として偏ったものになり、我々の手法で、誤って「怪しい」という判定になったことがあります。このように、実データでは、改ざんとは違う自然な原因で統計的に正常な分布から乖離する場合があります。このような状況を例外としてどう扱うかは、現在検討中の課題です。

企業への期待

（1）リアル数値データの提供と評価実験への協力

企業様が保有する具体的な「数値データの集合」を提供してもらい本技術の有効性を検証したい。数値改ざん有無、数値がどんな意味を持っているかという情報は不要。データの規模（数値の件数）100件でも数百万件でも構わない。提供頂いたデータセットについて、本システムでの検証結果をフィードバックします。検証結果を企業様で評価頂き、正答率がどれだけだったかを、教えていただきたい。

（2）ビジネスモデルの検討

対象データや対象分野毎の有効性評価が今後の大きな課題であるが、本技術をビジネスとしてどのような分野で展開できるか、一緒にアイディアをだし検討したい。

本技術に関する知的財産権

- 発明の名称：METHOD AND APPARATUS FOR DETECTING CORRECTION IN A SET OF NUMBERS
- 出 願 番 号：米国仮出願 62/861111
- 出　願　人：国立大学法人九州大学
- 発　明　者：廣川　佐千男、戸﨑　祐輔、鈴木　孝彦

お問い合わせ先

九州大学学術研究・産学官連携本部
知的財産グループ
ＴＥＬ　092-802-5137
ＦＡＸ　092-802-5145
e-mail　transfer@airimaq.kyushu-u.ac.jp