Vol.2019-MPS-123 No.33
Vol.2019-BIO-58 No.33
2019/6/18

ベンフォードの法則による障害者雇用状況集計結果の誤り箇所推定

戸﨑祐輔[†1]　鈴木孝彦[†1]　峯恒憲[†1]　廣川佐千男[†1]

概要：多くの数値データについて，最上位桁の数字の出現確率に法則性があり，ベンフォードの法則として知られている．この法則は統計データの不正検出に使われている．2018 年，厚生労働省が公表している障害者雇用状況の集計結果について誤りが判明し，修正が行われた．本論文では，まず，修正前後において，集計結果がベンフォードの法則に従うか否かを調べ，ベンフォードの法則の有用性を確かめる．さらに，複数の k 進法（k＝3, 4,…）上のベンフォードの法則を用いて，数値データの誤り箇所を推定する手法を提案する．障害者雇用状況の集計結果を用いて，推定性能を評価する．

キーワード：ベンフォードの法則，異常検知，データ改ざん，データマイニング，障害者雇用

AMultiple Views of Benford's Law for Detecting Numbers Corrected in Employment Statistics of Handicapped People

YUSUKE TOZAKI[†1]　TAKAHIKO SUZUKI[†1]
TSUNENORI MINE[†1]　SACHIO HIROKAWA[†1]

Abstract: For wide variety of numerical data, there is a rule in the distribution of the first significant digit, which is known as Benford's law. This rule is used to detect fraudulent statistics. In 2018, many errors were reported in "Employment statistics of disabled persons in Japan" published by Japan Ministry of Health, Labor and Welfare. After that, the revised statistics were published. In this paper, we firstly confirm whether the original and the revised statistics follow Benford's law or not. Following to that, we propose a new method that utilizes multiple views of Benford's law in k-adic system (k = 3, 4,...). The proposed method can detect the numbers which are candidates of correction in statistics beforehand. We evaluate the performance of the proposed method by using the original and the revised employment statistics.

Keywords: Benford's law, Anomaly detection, Data tampering, Data mining, Employment of handicapped people

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．多くの数値データについて，最上位桁の数字（FSD: First significant digit）である d（1〜9）の出現確率 $P_{FSD}(d)$が以下の式になることが，ベンフォードの法則として知られている[1].

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

図 1 にベンフォードの法則に従う分布の例を示す．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，FSD の分布を示したものである．数値の集合がベンフォードの法則に従わなければ，なんらかの不整合があると考えられ，この法則は統計データの不正検出に使われてきた．しかし，データがベンフォードの法則に従わないとき，データのどの部分に誤りがあるのかを推定することは難しい.

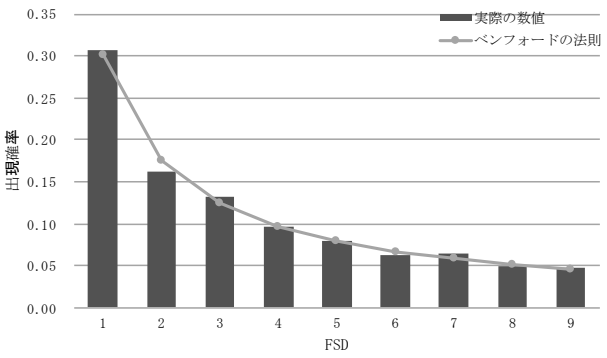

図 1　行政区画ごとの人口の FSD の分布

2018 年，厚生労働省が公表している障害者雇用状況の集計結果について誤りが判明し，修正が行われた．本論文では，まず，厚生労働省が公表している障害者雇用状況の集計結果がベンフォードの法則に従うか否かを検証する．

†1 九州大学
Kyushu University
tozaki.yusuke.713@s.kyushu-u.ac.jp
suzuki.takahiko.583@m.kyushu-u.ac.jp

Vol.2019-MPS-123 No.33
Vol.2019-BIO-58 No.33
2019/6/18

修正前と修正後の 2 つのデータを分析することで，ベンフォードの法則の有用性を確かめる．さらに，ベンフォードの法則を応用し，数値データの誤り箇所を推定する手法を提案する．修正前と修正後の障害者雇用状況の集計結果を用いて，提案手法の推定性能を評価する．

2. 関連研究

Nigrini ら[3]は，会計データの不正検出において，ベンフォードの法則が有用であることを示した．Rauch ら[4]は，ベンフォードの法則を用いて，EU 加盟国の経済データの信憑性を調査した．1999 年から 2009 年までの Eurostat のデータを分析し，ギリシャが報告したデータがベンフォードの法則から最も乖離していることを示した．当時，ギリシャの経済統計は複数回にわたり修正されており，信憑性が疑われていた．Berger ら[5]は，自然な数値データ集合において，10 以外の広い範囲の k 進数でベンフォードの法則が成り立つことを示した．

3. 障害者雇用状況の集計結果

厚生労働省は，毎年，障害者雇用状況の集計結果を公表している．しかし，2017 年 12 月 12 日に公表していた「平成 29 年 障害者雇用状況の集計結果」の数値に誤りがあることが判明し，再点検が行われた．2018 年 10 月 22 日に，最終的な数値を記載した報道発表資料[6, 7]を公表した．報道発表資料[6]には，国の機関の修正前と修正後の雇用障害者数が記載されている．報道発表資料[7]には，都道府県知事部局，その他の都道府県機関，都道府県教育委員会，独立行政法人等の修正前と修正後の雇用障害者数が記載されている．2 つの資料をあわせると，434 機関ある．

表 1 に障害者雇用状況の集計結果の抜粋を示す．短時間勤務職員は，1 人を 0.5 人としてカウントとしている．

表 1　障害者雇用状況の集計結果の抜粋

機関名	雇用障害者数		誤り
	修正前	修正後	
内閣官房	25.5	5.5	×
内閣法制局	2.0	2.0	
内閣府	56.0	29.0	×
宮内庁	22.5	10.0	×
公正取引委員会	18.0	17.0	×
警察庁	51.0	51.0	
金融庁	39.0	39.0	
消費者庁	10.0	0.5	×

内閣官房において，修正前の雇用障害者数は 25.5，修正後は 5.5 であり，修正が発生している．修正前，修正後ともにデータ数（雇用障害者数が未記載でなく，かつ 0 でない機関の数）は 422 である．また，修正前の 422 機関のうち，雇用障害者数が修正されたのは 167 機関である．

図 2，図 3 に修正前と修正後の雇用障害者数の FSD の分布を示す．

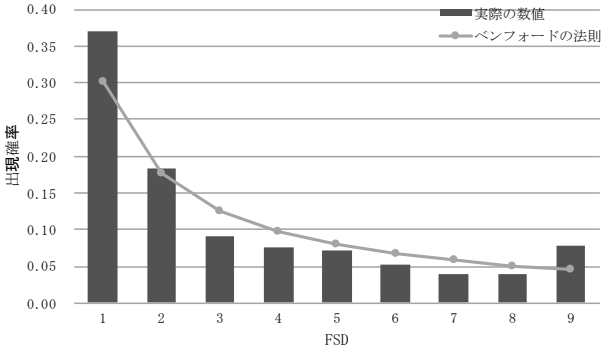

図 2　修正前の雇用障害者数の FSD の分布

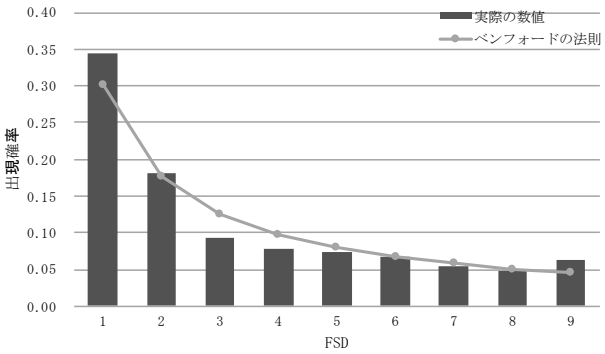

図 3　修正後の雇用障害者数の FSD の分布

カイ二乗検定によって，数値データのベンフォードの法則への適合度を見積もることができる．FSD である d の観測度数を O_d，ベンフォードの法則から予測される期待度数を E_d として検定統計量 χ^2 を以下の式で求める．

$$\chi^2 = \sum_{d=1}^{9} \frac{(O_d - E_d)^2}{E_d}$$

修正前のデータは $\chi^2 = 27.44$ ($p < 0.01$，自由度 8)，修正後のデータは $\chi^2 = 11.12$ (有意でない)となり，修正前のデータだけがベンフォードの法則に適合しないとみなされる．

4. 誤り箇所の推定

ベンフォードの法則を応用し，数値データの誤り箇所を推定する手法を提案する．提案手法は，自然な数値データの集合では 10 以外の基数の k 進法においても，ベンフォードの法則が成り立つという知見[5]に基づくものである．図 4，図 5 に 5 進法と 6 進法について，修正前の雇用障害者数の FSD の分布を示す．

Vol.2019-MPS-123 No.33
Vol.2019-BIO-58 No.33
2019/6/18

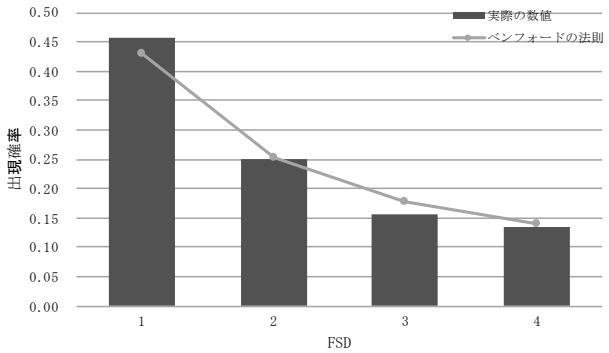

図 4　修正前の雇用障害者数の FSD の分布（5 進法）

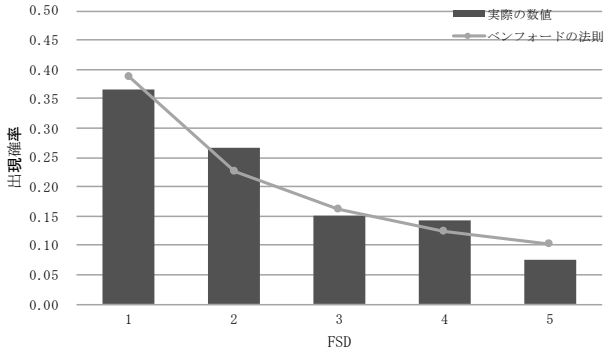

図 5　修正前の雇用障害者数の FSD の分布（6 進法）

以下に本論文中で使用する記法・用語について定義する．

・Pben(k, d)

基数 k のときの FSD である d のベンフォードの法則での出現確率

Pben(k, d) = $\log_k(1 + 1 / d)$

例）Pben(10, 1) = 0.301

・fsd(x, k)

数値 x を基数 k で記述したときの FSD である d を表す関数

例）fsd(123, 10) = 1

fsd(18, 3) = 2 ($18 = 200_{(3)}$)

・v(c)

機関 c における修正前の雇用障害者数

例）v(総務省) = 110

・r(c)

機関 c における修正後の雇用障害者数

例）r(総務省) = 40

・セル

統計データは複数の二次元の表として与えられることが多い．表の各要素をセルと呼ぶ．c を固定したとき，v(c)および r(c)はそれぞれ一つのセルに対応する．v(c) ≠ r(c)のとき，v(c)に対応する「セルが修正される」と表現する．

・C_{all}

修正前の雇用障害者数 v(c)が未記載でなく，かつ 0 でない機関 c の集合

C_{all} = {c_1, c_2, c_3,…} = {内閣官房, 内閣法制局, 内閣府,…}

C_{all} の要素数（本論文の分析対象となる機関の数）は，

$|C_{all}|$ = 422 である．|S|は集合 S の要素数を意味する．

・C_{kd}

v(c)を基数 k で表現したとき，FSD である d が同じになる機関 c の集合

C_{kd} = {c | fsd(v(c), k)) = d, c∈C_{all}}

例）$C_{10\,1}$ = {消費者庁, 総務省, 外務省,…}

$C_{10\,2}$ = {宮内庁, 財務省, 観光庁,…}

・over(c, k)

C_{kd} の要素数とベンフォードの法則から予測できる C_{kd} の要素数を比較し，0 と 1 を返す関数．実際の機関の数がベンフォードの法則から予測できる機関の数よりも大きければ 1 とし，そうでなければ 0 とする．

d = fsd(v(c), k)とし，

・$|C_{kd}| > |C_{all}|$・Pben(k, d)の場合

over(c, k) = 1

・それ以外

over(c, k) = 0

とする．

例）over(総務省, 10) = 1

over(金融庁, 10) = 0

・OverS(c)

ある範囲の基数 k に対する over(c, k)の総和．本論文では（3,…,16）の範囲を使う．

$$\mathrm{OverS(c)} = \sum_{k=3}^{16} \mathrm{over(c,k)}$$

例）OverS(外務省) = 14

OverS(徳島大学) = 0

OverS(c)が大きくなるほど，c が v(c) ≠ r(c)である可能性が高いとみなす．over(c, k)の計算において，

$|C_{kd}| > |C_{all}|$・Pben(k, d)の場合を 1 とした理由は，次の通りである．

「正しい」データがベンフォードの法則に従うと仮定するならば，$|C_{kd}| > |C_{all}|$・Pben(k, d)である C_{kd} の中に「誤った」データが含まれると推定できる(over(c, k) = 1)．逆に$|C_{kd}|$

Vol.2019-MPS-123 No.33
Vol.2019-BIO-58 No.33
2019/6/18

≦|C_{all}|・Pben(k, d)であるならば，C_{kd}の中に「誤った」データが存在するとは限らない（over(c, k) = 0）．

基数 k を 3 から 16 に変化させることによって，ある機関 c に対し，複数の C_{kd} の組み合わせについて，v(c) ≠ r(c)となる可能性を評価できる．v(c)が異なれば，C_{kd} の組み合わせも変化する．例えば，v(富山大学) = 45，v(造幣局) = 27であり，k = 5 では，’富山大学’, ’造幣局’∈C_{51} となるが，k = 6 では’富山大学’∈C_{61} (45 = $113_{(6)}$)，’造幣局’∈C_{64} (27 = $43_{(6)}$)となり，45 と 27 の誤りの可能性を相異なる複数の視点で評価できる．

5. 実験結果

5.1 提案手法

障害者雇用状況の集計結果を用いて，提案手法の推定性能を評価した．OverS(c) ≧ i の場合，v(c)を誤りと判定する．v(c) ≠ r(c)となる機関 c（c∈C_{all}）を正例として，しきい値 i を変化させた場合の Precision，Recall，F-measure，Accuracy を図 6 に示す．しきい値 i = 14, 13, 12, 11, 10 における結果を表 2 に抜粋する．OverS(c) ≧ 13 のとき，Precision が 0.783 となっている．このとき，誤りと判定された機関は 23（TP + FP）あり，18（TP）の機関で，修正前と修正後の雇用障害者数が異なっている．

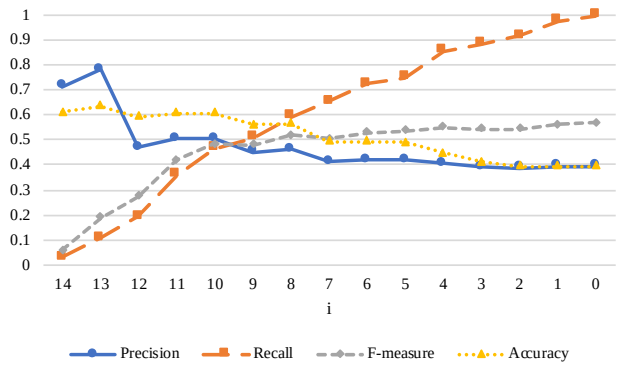

図 6　提案手法の推定性能

表 2　各手法の推定性能比較

			TP	FP	Precision	Recall	F-meas.	Accuracy
ゴールドスタンダード	最尤法	14	0	0		0.000		0.604
		13	18	5	0.783	0.108	0.189	0.635
		12	35	12	0.745	0.210	0.327	0.659
		11	56	30	0.651	0.335	0.443	0.666
		10	70	52	0.574	0.419	0.484	0.647
推定方式	ランダム判定		167	255	0.396	1.000	0.567	0.396
	提案方式	14	5	2	0.714	0.030	0.057	0.611
		13	18	5	0.783	0.108	0.189	0.635
		12	32	36	0.471	0.192	0.272	0.595
		11	60	59	0.504	0.359	0.420	0.607
		10	78	77	0.503	0.467	0.484	0.607

5.2 ランダム判定法

ベースラインとしてセル v(c)の誤りをランダムに判定する方法を考える．正しい判定となる確率は，

$$|\{c \mid v(c) \neq r(c), c \in C_{all}\}| / |C_{all}| = 0.395$$

であり，この値が Precision と等しい．すべてのセルを誤りと判定するとき Recall = 1 となる．すべてのセルを誤りと判定する方法をランダム判定法とする．比較のために，ランダム判定法の評価値を表 2 に示す．

5.3 10 進数のみの評価

|C_{10d}|, (d = 1, 2, 3,…,9)のベンフォードの法則からの乖離 max(|C_{10d}| - |C_{all}|・Pben(10, d), 0)を大きい順に並べ，上位 i 個の集合を誤りとみなし，評価した結果を表 3 に示す．なお，C_{101}，C_{109}，C_{102}，以外の乖離はすべて 0 であった．

表 3　10 進数のみの推定性能

Condition	TP	FP	Precision	Recall	F-measure	Accuracy
C_{101}	72	84	0.462	0.431	0.446	0.576
C_{101}∪ C_{109}	87	102	0.460	0.521	0.489	0.569
C_{101}∪ C_{102}∪ C_{109}	116	150	0.436	0.695	0.536	0.524
C_{all}	167	255	0.396	1.000	0.567	0.396

5.4 FSD と正解情報を使った評価（最尤法）

あるセルが修正されるかどうかを，そのセルの値のみを使って判定するものと仮定する．値が同じセルでも，修正されるものと修正されないものがあり得る．

この条件で理想的な判定は，同じ値のセルについて推定誤り（error rate）が最小になる判定である．あらかじめ修正の有無がわかっていれば，修正されている確率の高い値のセルを，修正されていると判定すればよい．

セルの値 u そのものではなく，u の fsd(u, k) (k = 3,…,16)だけを見て推定誤りを最小にする最尤法は以下のように定義できる．

・最尤法

C_{kd} が v(c) ≠ r(c)となる c を多く含むか否かの判定をベンフォードの法則ではなく，C_{kd} 中の v(c) ≠ r(c)の数から評価する．ここで，関数 over(c, k)，OverS(c)の代わりに，以下に定義する関数 wrong(c, k)，WrongS(c)を使っている．

・wrong(c, k)

d = fsd(v(c), k)とし，

・|{e | v(e) ≠ r(e), e∈C_{kd}|} / |C_{kd}| > 0.395
　　(0.395 = |{c | v(c) ≠ r(c), c∈C_{all}|} / |C_{all}|)
　の場合
　　wrong(c, k) = 1
・それ以外
　　wrong(c, k) = 0

とする．

Vol.2019-MPS-123 No.33
Vol.2019-BIO-58 No.33
2019/6/18

・**WrongS(c)**

$$\mathrm{WrongS}(c) = \sum_{k=3}^{16} \mathrm{wrong}(c,k)$$

最尤法は，修正内容を知った上で，fsd(v(c), k) (k = 3,…,16)の情報を用いた最善の判定法（ゴールドスタンダード）である．最尤法（WrongS(c) ≧ i, i = 1,…,14）の結果を図 7 に示す．また，しきい値 i = 14, 13, 12, 11, 10 の場合を表 2 に併記し，提案手法の結果と比較する．

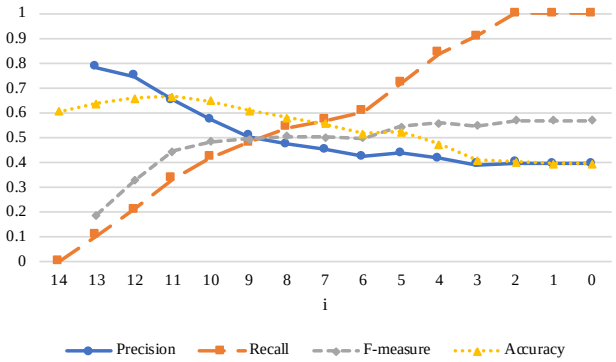

図 7　最尤法の推定性能

6. 考察

提案手法でしきい値 OverS(c) ≧ 13 とした場合，誤りと判定した 23 機関（TP + FP）のうち，18 機関（TP）において，実際にデータが修正されている（Precision = 0.783）．これは，ランダムに機関を選択した場合にデータが修正されている確率 0.395 より充分高い．全体の誤り発生状況が判明する前に，これら 23 機関を重点的に調査しておけば，誤りの発生を前もって指摘できた可能性がある．

6.1 ランダム判定法との比較

提案手法は，OverS(c)のしきい値 i = (4,…,14)の範囲で，ランダム判定法の Precision = 0.395 よりも高い Precision を示し（図 6），ベンフォードの法則は雇用障害者数の誤り識別に有用だと考えられる．

6.2 10 進法のみの評価との比較

10 進数で最上位桁がベンフォードの法則の期待値よりも頻繁に表れる集合 $C_{10\,1}$，$C_{10\,2}$，および $C_{10\,9}$ に含まれる c を誤りと判定する手法では，Precision，Recall ともに提案手法には及ばない（表 3）．

6.3 最尤法との比較

表 2 に示すように，しきい値 OverS(c) ≧ 13 および WrongS(c) ≧ 13 を用いた場合，提案手法と最尤法の Precision，Recall，F-measure，Accuracy は同じである．

これは，fsd(u, k) (k = 3,…,16)の情報とベンフォードの法則を使えば，あらかじめ答えを知っている場合と同等の優れた推定性能を達成できることを意味する．

しきい値を 12 以下にしたときは，最尤法の推定性能が提案手法を上回っている．その原因について以下に考察する．

6.3.1 最尤法が上回る原因の考察

C_{all} 中の v(c) = x の出現回数 count(x) = |{c | v(c) = x, c∈C_{all}}| を見ると，表 4 のようになっている．値 9 の出現回数が周囲の値と比べて突出しており，かつ x = v(c) = 9 について OverS(c) = 12 となっている．ベンフォードの法則は，基本的に数値の出現回数に依存しているため，9 の出現回数が非常に多い場合に本手法では誤りと判定する．

表 4　v(c) = x の出現回数 count(x)と OverS(c)（抜粋）

x	count(x)	OverS(c) for v(c) = x
7	8	6
7.5	2	6
8	10	9
8.5	1	9
9	17	12
9.5	1	12
10	11	11
10.5	0	
11	5	8

さらに，v(c) = 9 である機関を調べてみると，17 機関中 10 機関が’～県警察本部’となっている．これは，警察本部の総数が 46 件であることを考えると，警察本部特有の性質によって，9 の出現回数が多くなっているものと考えられる．本論文の手法では，このような現象と誤りを識別することができない．

6.4 基数の選択

本論文では，基数（3,…,16）までを選択し，ベンフォードの法則からの乖離を見積もった．提案手法は，原理的に 3 以上のいかなる基数にも対応できる．

6.4.1 最大基数の選択

本論文で基数 k の最大値を 16 としたのは，基数 k におけるベンフォードの法則が，数値データの最大値と最小値の桁が 2.6 以上異なる場合に限定したためである．

$$\max_c(\log_k \mathrm{v}(c)) - \min_c(\log_k \mathrm{v}(c)) > 2.6$$

しかし，しきい値を 2 まで引き下げれば，基数の最大値を 40 程度まで取ることができる．基数の最大値をあげることで本手法の推定性能が変化する可能性がある．

Vol.2019-MPS-123 No.33
Vol.2019-BIO-58 No.33
2019/6/18

6.4.2 小さな基数の選択的利用

9(＝3^2)進数でベンフォードの法則に従わない数値集合について，3 進法での評価を重ねて行うべきか疑問が残る．一般の数値集合について，

$C_{31} = \{C_{91} \cup C_{93} \cup C_{94} \cup C_{95}\}$,

$C_{32} = \{C_{92} \cup C_{96} \cup C_{97} \cup C_{98}\}$

の関係が成り立つ．3 進数での評価は 9 進数での評価を集成したものとみなせる．9 進数でベンフォードの法則に従わない数値集合が，3 進数ではベンフォードの法則に従うことも考えられる．

このような場合，OverS(c)から 3 進数での評価を取り除いた方が，推定性能は向上すると考えられる．同様な議論は 4 進数と 16 進数の組についても成り立つ．

7. おわりに

本論文では，まず，2018 年に判明した障害者雇用状況の集計結果の誤りについて，修正前のデータがベンフォードの法則を満たさず，修正後のデータはベンフォードの法則を満たすことを確認した．次に，基数 k (k＝3,…,16)でのベンフォードの法則を利用し，数値データの誤り箇所を推定する手法を提案した．さらに，修正前と修正後のデータを用いて, 提案手法の推定性能を評価した. 提案手法により，誤りが多い一部の箇所を推定することが可能になった．誤りが多いと推定される箇所の特徴を集中的に調べれば，前もって誤りの傾向を把握可能である．より広い範囲で誤り箇所を推定できるよう手法を改良することが今後の課題である．

参考文献

[1] Benford, F., The Law of Anomalous Numbers, Proceedings of the American Philosophical Society, vol. 78, pp. 551-572, 1938.
[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/, (参照 2018-06-25).
[3] Nigrini, M. J., *Benford's Law Applications for Forensic Accounting, Auditing, and Fraud Detection*, ISBN: 9781118152850, Wiley, 2012.
[4] Rauch, B., Goettsche, M., Braehler, G., Engel, S., Fact and Fiction in EU-Governmental Economic Data, German Economic Review, vol. 12(3), pp. 243-255, 2011.
[5] Berger, A., Hill, T. P., A basic theory of Benford's law, Probability Surveys, vol. 8(1), pp. 1-126, 2011.
[6] “平成 30 年 8 月 28 日に公表した「国の行政機関における平成 29 年 6 月 1 日現在の障害者の任免状況の再点検結果について」及び同年 9 月 7 日に公表した「立法機関及び司法機関における平成 29 年 6 月 1 日現在の障害者の任免状況の再点検結果について」の訂正について”. https://www.mhlw.go.jp/content/11704000/000369693.pdf, (参照 2018-10-31).
[7] “都道府県の機関，市町村の機関，都道府県等の教育委員会及び独立行政法人等における平成 29 年 6 月 1 日現在の障害者の任免状況等の再点検結果について”. https://www.mhlw.go.jp/content/11704000/000463282.pdf, (参照 2018-10-31).

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
タイトル(英)	AMultiple Views of Benford's Law for Detecting Numbers Corrected in Employment Statistics of Handicapped People	Multiple Views of Benford's Law for Detecting Numbers Corrected in Employment Statistics of Handicapped People