Benford's Law の検証と活用

～ 勘定データを用いて ～

2007MI020 藤田一輝　　2007MI120 沓名貴規
指導教員: 尾崎俊治

1 はじめに

本論文では, 自然界に存在する数において, 数値の最初の桁の数を考えるとする. 一般的に桁数の多い値に関して 1 桁目の数値を分類した場合, 本論文の Benford's Law を知らない場合であればこれが 1 から 9 まで一様に分布していると考えてしまう人が多いであろう. だが自然界に存在する多くの(全てではない)数値の最初の桁の分布が一様ではなく特定の割合に限りなく近付くものになっているのである(文献[2]).

この論文では, まず Benford's Law との様々な関係性を考えていく. それに加え, 実際の勘定データを基に統計的解析や検定を用いることで, 不正会計の発見方法を考え考察を行う.

2 Benford's Law とは

Benford's Law は, 1881 年に数学者で天文学者のサイモン・ニューカムが対数表を見ることで最初の桁数において 1 が多いことに気づき, 最初の桁数が 1 から 9 である確率が全て 1/9 ではないことを発見した. そして 1938 年に物理学者であるフランク・ベンフォードがそれを基に 20,229 もの膨大な統計データを用いて関係性を証明し, 数値の分布を考えた(文献[2]).

また, これを数式に表すと, 最初の桁の数が x ($x = 1, 2, \cdots, 9$) であるときの確率を $f(x)$ を用いた式にすると (1) のようになる.

$$
\begin{aligned}
f(x) &= \log_{10}(x+1) - \log_{10} x \\
&= \log_{10}(\frac{x+1}{x}) \\
&= \log_{10}(1 + \frac{1}{x})
\end{aligned}
\tag{1}
$$

次の表 1 は, Benford's Lawを用いたとき, 最初の桁の値が 1 から 9 であるときの確率を示している(文献[1]).

表 1: 最初の桁が x ($x = 1, 2, \cdots, 9$)になる確率 (文献[1])

x	確率(%)
1	30.1
2	17.6
3	12.5
4	9.7
5	7.9
6	6.7
7	5.8
8	5.1
9	4.6

そして最初の桁数が x であるときの和は, 式(2)のようになる.

$$\sum_{x=1}^{9} f(x) = 1 \tag{2}$$

2.1 2 桁の Benford's Law

これまでは, 最初の 1 桁である数を対象に考えていた. ここでは, 最初の 2 桁の数を対象に Benford's Law を用いるものとする.

次に最初の桁が x , さらに第 2 の桁が y であるときの式は, 式(3)のようになる.

$$f(x, y) = \log_{10}(1 + \frac{1}{10x + y}) \tag{3}$$

次の表 2 では, 最初の 2 桁が 10 から 99 である確率を表している. そこで表 1, 2 は, 大きい数字になればなるほど最初の桁になる確率が少ないことが明らかである.

表 2: 最初の数値が $10x + y$ になる確率
($x, y = 0, 1, \cdots, 9$)

$y \mid x$	1	2	3	4	5	6	7	8	9
0	4.14	2.12	1.42	1.07	0.86	0.72	0.62	0.54	0.48
1	3.78	2.02	1.38	1.05	0.84	0.71	0.61	0.53	0.47
2	3.48	1.93	1.34	1.02	0.83	0.69	0.60	0.53	0.47
3	3.22	1.85	1.30	1.00	0.81	0.68	0.59	0.52	0.46
4	3.00	1.77	1.26	0.98	0.80	0.67	0.58	0.51	0.46
5	2.80	1.70	1.22	0.95	0.78	0.66	0.58	0.51	0.45
6	2.63	1.64	1.19	0.93	0.77	0.65	0.57	0.50	0.45
7	2.48	1.58	1.16	0.91	0.76	0.64	0.56	0.50	0.45
8	2.35	1.52	1.13	0.90	0.74	0.63	0.55	0.49	0.44
9	2.23	1.47	1.10	0.88	0.73	0.62	0.55	0.49	0.44

そして最初の桁が x , さらに第 2 の桁が y であるときの和は, 式(4)のようになる.

$$\sum_{x=1}^{9} \sum_{y=0}^{9} f(x, y) = 1 \tag{4}$$

3 Benford's Law の性質

ここでは, 基本的な Benford's Law の性質である最初の桁数の使用頻度以外の性質に注目し, 考察を行う.

3.1 最初の桁以外の桁に対する性質

次の表のように最初の桁以外の桁の値に注目した際, 0 から 9 になるにつれて数字の使用頻度は低くなるが, 最初の桁の使用頻度に比べ, 大きな差は存在しない(文献[5]).

つまり最初の桁に注目することが一番の性質ではあるが, この各 2, 3, 4 桁の使用頻度を考えると, 桁数を増やしていっても大きい数になることで頻度が下がることがわかる.

表 3: 最初の桁以外の数字の頻度 (文献[5])
(2 桁目から 4 桁目の確率)

数字	2番目の桁	3番目の桁	4番目の桁
0	11.968	10.178	10.018
1	11.389	10.138	10.014
2	10.882	10.097	10.010
3	10.433	10.057	10.006
4	10.031	10.018	10.002
5	9.668	9.979	9.998
6	9.337	9.940	9.994
7	9.035	9.902	9.990
8	8.757	9.864	9.986
9	8.500	9.827	9.982

4 Benford's Law に従う数列

4.1 等比数列との比較

次の図 1 は, 公比を 2 とした等比数列を 1,000 項, つまり 1,000 個のデータを用いてグラフにしたものである. これを見る限り, 1,000 項 までの等比数列による最初の桁の統計結果は, Benford's Law に近似している.

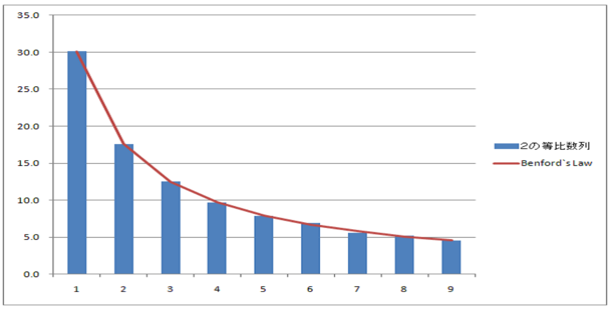

図 1: Benford's Lawと等比数列の比較

4.2 フィボナッチ数列との比較

まずフィボナッチ数列の定義を行う. ここでのフィボナッチ数列は, 次の数式により定義される(文献[4]).

$$F_1 = 1 \tag{5}$$

$$F_2 = 1 \tag{6}$$

$$F_{n+2} = F_n + F_{n+1} \qquad (n \geq 1) \tag{7}$$

この定義に従って, 次の図 2 の結果が得られる. フィボナッチ数列でも 1,000 個のデータを用いて統計を行ったところ, Benford's Law と近似する結果となった.

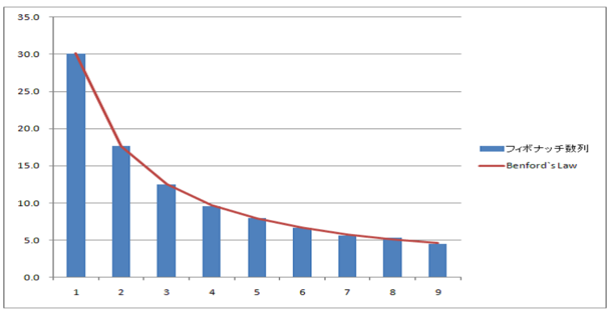

図 2: Benford's Lawとフィボナッチ数列の比較

4.3 等比数列とフィボナッチ数列の比較

今回 1,000 個のデータ数での等比数列やフィボナッチ数列に関して, 上の図 1 , 図 2 にもあるように Benford's Law に近似していることが分かった.

さらにこの結果を見ると少しではあるが等比数列の値がより Benford's Law に近似し, 近い値を示している. つまり Benford's Law に近似するデータに対しては, 等比数列にも近似し, そのデータは何らかの関係で比率的に増加や減少した値の数値となっているのではないかという結論に至った.

5 データによる Benford's Law の検証

ここでは, まず Benford's Law の検証を行う前に, 扱う実データの説明をし, そのデータを基に検証を行う中で考察を行う. そこで検証を行ったうえで Benford's Law との誤差を考え, Benford's Law の成り立ちを示す.

その際に, 統計結果からそのデータが Benford's Law に近似していると見なしてよいかをカイ二乗検定を用いて検証する.

5.1 実データによる Benford's Law の検証

5.1.1 グラフでの検証

ここでは, 自然界に存在するデータを用いて, Benford's Law の検証を行う. 次の図 3, 4 の x 軸は, 最初の桁の数値を示し, y 軸はその桁数の割合を示すものである.

まず図 3 では, Benford's Law の割合を棒グラフとし, 実データである「携帯電話の契約数」に折れ線グラフを用いて示したものである. このデータの最初の桁数を考えたとき, 図 3 の結果からこのデータは, Benford's Law の割合に近似していることが分かる(文献[6]).

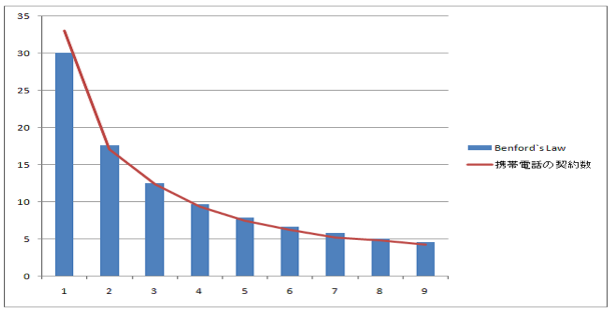

図 3: Benford's Law と実データの比較 (1) (文献[6])

次に図 4 は, 実データである「新車販売台数」と「日本の都市人口」を折れ線グラフとし, 示したものである.

この結果を見る限り, 図 3 のデータとは違い, データには明らかな誤差が発生している. つまりこのデータは, 成り立っていないのであろうか. しかし Benford's Law は, 自然界に存在するデータであれば例外なく成り立つので, このデータのデータ数が不足している可能性が大いに考えられる(文献[7-8]).

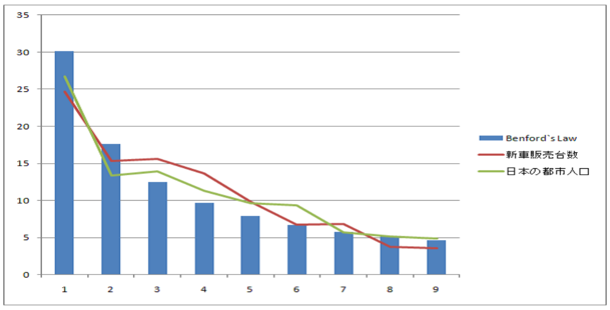

図 4: Benford's Law の割合と実データの比較 (2) (文献[7-8])

5.1.2　カイ二乗検定での検証

ここでは, 図 3, 4 で用いた 3 つのデータをカイ二乗検定を用いて検定し, Benford's Law に近似するものなのか否かを検証する(文献[3]).

この論文で用いる検定において, 帰無仮説を「Benford's Law に近似するデータである」と考えるので,

$$H_0 : p_i = p_{i0} \tag{8}$$
$$(p_{i0}\text{は, Benford's Lawの値とする})$$

表 4: 各データのカイ二乗値

項目	カイ二乗値
携帯電話	7.293922
販売台数	46.33765
都市人口	31.83359

ここで用いたデータの自由度は 9 であり, 有意水準 5 %の検定を考えると,

$$\chi^2_9(0.05) = 16.92 \tag{9}$$
$$\text{棄却域は}, R = \{\chi^2 > \chi^2_9(0.05)\}$$

この値から上のデータを検証すると, 「車の販売台数」と「日本の都市人口」に対しては棄却され, 「携帯電話の契約数」に対してのみ対立仮説が有意となった.

今回, グラフを比較しただけでも Benford's Law に従うか否かを見極めることが可能だが、カイ二乗検定を行うことで, 正確な検証ができることがわかった.

5.2　勘定データによる 2 桁の Benford's Law

ここでは, あらゆる勘定データを集め, 約 10,000 個のデータ数を用いて 2 桁の Benford's Law を検証し, その結果を述べ, 考察を行う(文献[3]).

さらに, ここでもカイ二乗検定を用いた検証を行うことで 2 桁の Benford's Law による近似度がどの程度かどうかを考える.

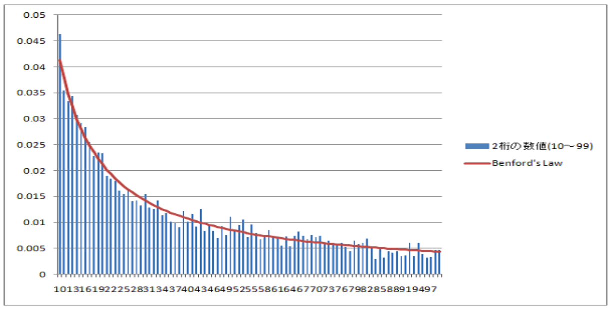

図 5: 2 桁の Benford's Law と勘定データの比較

表 5: 勘定データの 2 桁を用いたカイ二乗値

項目	カイ二乗値
勘定データ	110.0642

ここで用いたデータの自由度は 89 であり, 有意水準 5 %の検定を考えると,

$$\chi^2_{89}(0.05) = 112.02 \tag{10}$$
$$\text{棄却域は}, R = \{\chi^2 > \chi^2_{89}(0.05)\}$$

2 桁の Benford's Law の検証を行ううえで, 最初の 1 桁のデータ数では不十分であるため, 今回は約 10,000 個の勘定データを用いて Benford's Law と比較するために図 5 のグラフとした. 結果を見ると, やはり誤差は生じているが, このデータにおける最初の 2 桁を考えたとき, 2 桁の Benford's Law に近似してきていることがわかる.

そこでカイ二乗検定を行った結果, 棄却されるので, Benford's Law に近似するといえる.

今回のデータは, あらゆる種類の勘定データを含めていることから, 半信半疑で実行を行ったが, 近似することが分かり, 良い検証となったと考えている. そしてこの 2 桁の Benford's Law に関してもデータ数をもっと増やしていけばより近似した結果が期待できると考える.

6　Benford's Law と不正会計

6.1　不正会計とは

不正会計は, 一般に認められている会計の基準に基づいていないものであり, 不正な処理を行ったうえで財務諸表を最終的に歪めて作成することである. つまり, すべての結果となる財務諸表の誤差が問題点であり, 不正という点では, その行為に悪意があるものであったのかが問題となります.

6.2　Benford's Law による不正会計

ここでは, Benford's Law を用いることで, この不正会計を発見する方法を見つめると共に検証を行う. そこでまず, 不正会計を考えるために, 財務諸表などの勘定データを用いる.

勘定データは, 実際にそこで行われた取引などに対し存

在しているデータであるので, 乱数とは違う自然界に存在する数に値する. つまり勘定データは, Benford's Law に従うデータなのである. よってこの勘定データに不正が行われている場合に, この Benford's Law に従っていない結果が導かれるため, データの不正を発見することができる(文献[5]).

6.3 勘定データによる不正会計

ここでは実データを用いて, そのデータを基に意図的に不正データを作成し, グラフの作成やカイ二乗検定を用いることで比較を行い, 不正会計について考察を行う.

実際に求める上で, 図 5 のデータを用いて, 最初の 1 桁目について, 1 から 9 までの割り合いを図 6 に示す.

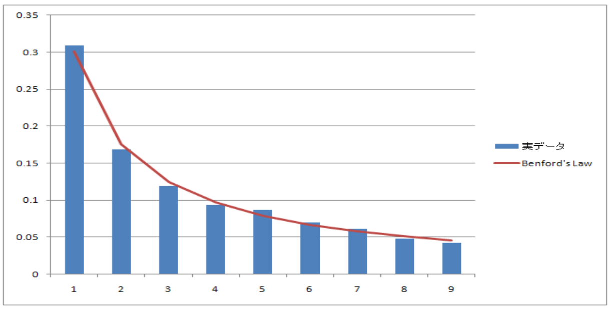

図 6: 1 桁の Benford's Law と勘定データの比較

次に, この図 6 のデータを意図的に 2 桁目の値を四捨五入することで不正データを作成し, そのデータを次の図 7 に示す.

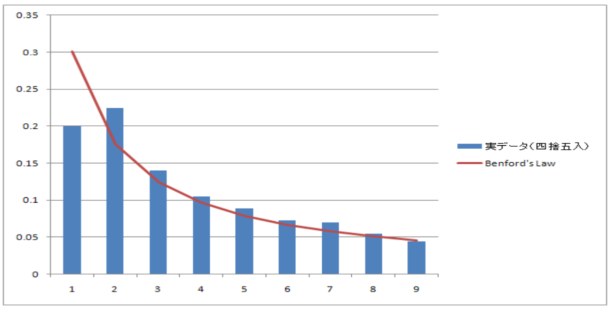

図 7: Benford's Law と不正データの比較

上の図 6, 7 を見て比較を行っても明らかであるが, ここでこの 2 つのデータについてのカイ二乗検定を行う(文献[3]).

表 6: 不正データに対するカイ二乗検定

項目	カイ二乗値
図 6 のデータ	16.48127
図 7 のデータ	386.2407

今回グラフを見た限りでも明らかである結果をカイ二乗検定を用いて, 数学的に検証した. その結果, (8) の帰無仮説を用いて, 棄却域を (9) としたとき, 実際のデータである図 6 に関しては棄却されず, 明らかに不正としてつくったデータである図 7 に対し, カイ二乗値が高くなり棄却された.

ここで扱った不正データに対しては, グラフにするだけでも明らかに Benford's Law に近似していない結果となっていたが, 不正データを発見する際にグラフとして集計を行うこと自体が困難であり, それができたと考えてもカイ二乗検定での検証を行うことが発見につながるという結論に達した.

7 おわりに

本研究では, Benford's Law を用いた実データの検証を行うことで, 何の統一性も無いデータのほとんどが Benford's Law に近似していることを証明した.

さらに, この Benford's Law の性質から不正会計にまで進め, 検証を行った. その結果, 入手できるデータ数での検証は困難であり, より多くのデータ数が必要であることが分かった.

ここでは, 統計的検証の他にカイ二乗検定を用いて考えた. ここでのグラフの差だけでは充分に結論を出せず, カイ二乗検定を用いることでより正確にそのデータが正しいか否かを検証することができた.

今回の研究で用いたデータはまだ一部であるが, データ数の少なさにより良い結果が出ないことが多かったので, これからも多くのデータを扱う必要がある.

また, 今後はこの不正会計を課題とし, より効率の良い検証方法を検討することが必要である.

参考文献

[1] ベンフォードの法則…数字の不思議な法則: http://www.rd.mmtr.or.jp/~bunryu/benford.shtml

[2] ベンフォードの法則をめぐって: http://www.geocities.jp/ikuro_kotaro/koramu/872_b.html

[3] 白旗慎吾:『統計解析入門』, 共立出版株式会社 (1992).

[4] フィボナッチ数列の性質: http://web2.incl.ne.jp/yaoki/kinou.htm

[5] 事例から学ぶ『有効な不正対策』デジタル分析入門: http://www.tabisland.ne.jp/acfe/fraud/fraud_008.htm

[6] 携帯電話・PHS契約数 社団法人 電気通信事業者協会: http://www.tca.or.jp/database/index.html

[7] 自販連のホームページ/統計データ http://www.jada.or.jp/contents/data/type/index12.php

[8] 統計局ホームページ -日本の統計- 第2章 人口・世帯 - 都市別人口: http://www.stat.go.jp/data/nihon/02.htm