最高桁の数

～ ベンフォード則と不正経理の捜査 ～

「数学で身に付ける柔らかい思考力」　ダイヤモンド社
ロブ・イースタウェイ／ジェレミー・ウィンダム 著　水谷淳 訳　より

1．新聞に載っている数字を調べよう

今日の新聞の第１面を見るとたくさんの数字が載っている。「１６兆円の景気対策・・・」「１０月入学を採用・・・」「８４年以降・・・」「３０万人余りと米国へ・・・」。

これらの数字はお互いに無関係であって、ほぼランダムにとってきた数字と思ってよい。本によると、実は、これらの数字はある傾向の分布をするというのだ。たとえば、１から始まる数字は、どれくらいあるだろうか。２から始まる数字はどうだろう。３から始まる数は？普通、１から９まで同じくらいの割合だと思うが、そうではないらしい。

実は、ランダムに取ってきた数が１で始まる確率は、他の数字に比べてずっと高く、約半分の数が１か２で始まるのだ。そして、９から始まる数は稀なのだ。

2．ベンフォード則

新聞に載っているたくさんの数字の統計をとると、次の表に近い結果を得られる。

数	1	2	3	4	5	6	7	8	9
割合（%）	30	18	12	10	8	7	6	5	4

Frank Albert. Benford, Jr. (1883-1948)

この分布は、１９３９年、ゼネラル・エレクトリック社の技術者フランク・ベンフォードが発見したもので、都市の人口、株価、川の長さ、スポーツの成績など様々な数に当てはまる。サンプル数が十分に多く、値の範囲が制限されていないものであれば、どのようなものにも当てはまる。この分布に従うことを発見者の名をとって、**ベンフォード則**に従うという。（値の範囲が制限されるもの、例えば、電話番号は桁数が限られておりベンフォード則には従わない。）

3．ベンフォード則と不正経理の捜査

ベンフォード則に纏わる面白い話題がある。１９９０年代初め、会計学校講師のマーク・ニグリニは、学生たちに企業収支のデータについて、あることを調べるように課題を出した。それは、「企業収支の各数値の最高桁の数字がベンフォード則に従った分布を示すかどうか確かめよ」というものだった。

ある学生が親戚の経営する金物屋の帳簿を調べたところ、その数字の分布はベンフォード則の分布とは全く違うものになった。それは、１から始まる数値はベンフォード則によれば３０％のはずなのに、この帳簿では９３％もあった。そして、残りはすべて８か９で始まる数値だった。つまり、この帳簿に何かおかしなところがあるということを示す結果となってしまったのだ。これをきっかけに、多くの会計士が不正経理を発見する方法としてベンフォード則を採用するようになったそうである。念のため、私の知る大学の会計の先生にこの話題について尋ねたところ「寡聞にして存じ上げません」ということだったので、日本では、それほどポピュラーではないのかもしれない。

４．ベンフォード則が成り立つ理由

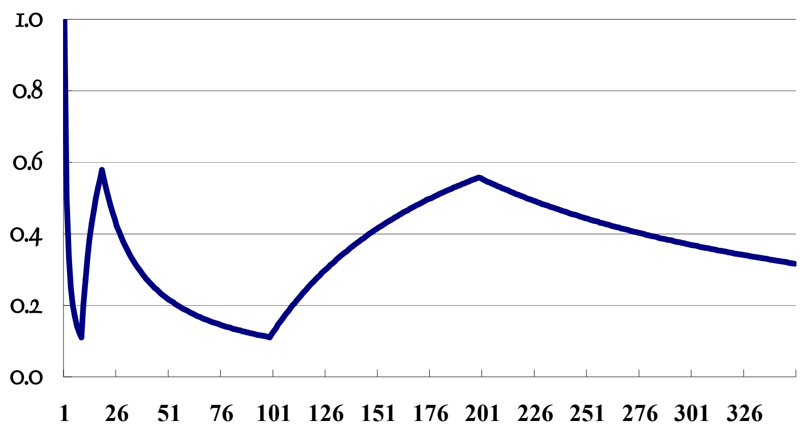
ベンフォード則の厳密な証明は難しいらしいので、ここでは感覚的に説明する。１桁の数１～４が書いてある４枚のカードから１のカードを引く確率は０．２５であるが、カードの枚数を増やしていくと１のカードを引く確率は下がり、１～９ではその確率は１÷９＝０．１１まで下がる。ところが、１～１０になると、1から始まるカードは１と１０の２枚になり、それらを引く確率は２÷１０＝０．２へと跳ね上がる。さらに、カードが１１，１２，１３，・・・と増えていくと、その確率は上がっていき１９枚のとき、１１÷１９＝０．５８にまでなる。

	1～4	～5	～6	～7	～8	～9	～10	～11	～12
1から始まる数の個数	1	1	1	1	1	1	2	3	4
1から始まる数の割合	0.25	0.20	0.17	0.14	0.13	0.11	0.20	0.27	0.33

～13	～14	～15	～16	～17	～18	～19	～20	～21	～22	～23
5	6	7	8	9	10	11	11	11	11	11
0.38	0.43	0.47	0.50	0.53	0.56	0.58	0.55	0.52	0.50	0.48

しかし、その先１０枚、２０枚と増やしていくと確率はまた下がり９９枚のとき、１から始まるカードを引く確率は１１÷９９＝０．１１となる。そして、１００枚以上になると確率は再び上がり、１９９枚になると1から始まるカードを引く確率は１１１÷１９９＝０．５６となる。この確率をグラフにしてみると以下のようになる。

1で始まる確率



カードの枚数を増やしていくと、確率は０．１１と０．５８の間を上下し続け、平均すると０．３くらいになる。

実は、この値はベンフォード則によって予測できて、数字Ｎから始まる番号のカードを引く確率は、正確には以下の式で求められる。

$$\log_{10}(N+1)-\log_{10} N$$

たとえば、Ｎ＝１とすると、$\log_{10} 2-\log_{10} 1=\log_{10} 2=0.301$　となる。

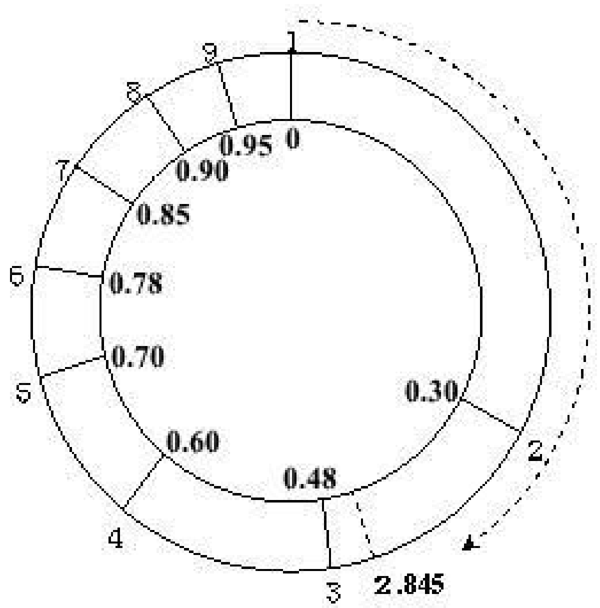
5．ポアンカレのルーレット定理による説明

10 倍するごとに１桁上がる構造を円周で考えよう。今、10 倍すると１周、さらに 10 倍すると２周という対応を考える。たとえば、１からスタートするならば、１は０周、10 倍して 10 は１周、さらに 10 倍して 100 は２周、という具合だ。すると、１や 10 や 100 は、対数関数 $\log_{10} x$ で周回数と対応させせることができて、$\log_{10} 1 = 0$（周），$\log_{10} 10 = 1$（周），$\log_{10} 100 = 2$（周）となる。そして、２～９の１桁の数は、０周から１周の間にあることになり、それらが何周に相当するかは、次のように求められる。

$\log_{10} 2 = 0.301$（周），$\log_{10} 3 = 0.477$（周），$\log_{10} 4 = 0.602$（周），$\log_{10} 5 = 0.699$（周），

$\log_{10} 6 = 0.778$（周），$\log_{10} 7 = 0.845$（周），$\log_{10} 8 = 0.903$（周），$\log_{10} 9 = 0.954$（周）

すなわち、２は 0.30 周、３は 0.48 周、４は 0.60 周、・・・・・、９は 0.95 周といった具合である。



ベンフォード則では、最高桁の数だけに着目しているので、それが何桁の数であるかは無視することになる。たとえば、284,572,341 という数は、2.84572341×10^8 と表せるから、８周を通り過ぎ、更に $\log_{10} 2 = 0.301$（周）よりもう少し通り過ぎたところにあるが、$\log_{10} 3 = 0.477$（周）よりは手前にあるということになる。このことは、最高桁が２の数は、それが何桁の数であっても、0.30 周から 0.48 周の間のどこかにあることを意味している。したがって、無作為に取ってきた数の最高桁が２である確率は、$\log_{10} 3 - \log_{10} 2 = 0.48 - 0.30 = 0.18$ という計算になる。