Zipf's Lawの適合性に関する研究

2003MM076 中野 義之　　2003MM098 鈴木 佑昂
指導教員　尾崎 俊治

1　はじめに

本論文では Zipf's Law の適合性・信頼性を調べるため，わたしたちの身の回りのあらゆるデータを用い，そのデータが Zipf's Law に基づいているのかを調べ，またそこからどのようなことがわかり，どのようなことに役立てることができるのかを研究する．

1.1　Zipf's Law とは

ジップの法則 Zipf's Law とは，世界中のあらゆる事象について，規模の順位が x 番目のものの規模は $1/x$ に比例するという法則であり，もとはアメリカの言語学者ジョージ・キングスリー・ジップ George Kingsley Zipf が英単語の使用頻度，およびその順位に関して発見した言語学の法則である．

数学的に Zipf's Law は

$$f(x;s,N)=\frac{\frac{1}{x^s}}{\sum_{n=1}^{N}\frac{1}{n^s}}$$

x ：順位
N : 全要素の数
と，一般に表される．ここで本来の Zipf's Law では $s=1$ であり，N を無限大にすると分母は収束しない．そのため，Zipf's Law では N を有限にしなければならない．

ただし，s が 1 より少しでも大きい実数ならば，N を無限大にしても分母は収束し，x の値を無限にとりうる分布関数とすることができる．これはゼータ関数 $\zeta(s)$ に等しい．

2　Zipf's Law の例

主な例として，日本の都市の人口が上げられる．

表 1　日本の各都市の人口 (2006/8/1)

順位 (x)	都市名	人口 (y)	$x\times y$
1	東京	8,396,594	8,396,594
2	横浜	3,559,867	7,119,734
3	大阪	2,633,819	7,901,457
4	名古屋	2,204,496	8,817,984
5	札幌	1,870,170	9,350,850
6	神戸	1,521,362	9,128,172
7	京都	1,463,941	10,247,587
8	福岡	1,393,659	11,149,272
9	川崎	1,306,992	11,762,928
10	広島	1,146,413	11,464,130

このように，順位 x とその規模の大きさ y とすると，それらをかけたもの，つまり、

$$x\times y$$

の値がいずれも近いものになっているのがわかる．このような結果が求められたのなら，そのデータは Zipf's Law に従っているということが言える．

3　都市の人口

日本の都市に続き，アメリカの各都市の人口のデータを調べ，考察する．

表 2　アメリカの各都市の人口 (2006)

順位 (x)	都市名	人口 (y)	$x\times y$
1	ニューヨーク	8,143,197	8,143,197
2	ロサンゼルス	3,845,541	7,691,082
3	シカゴ	2,842,518	8,527,554
4	ヒューストン	2,016,582	8,066,344
5	フィラデルフィア	1,463,281	7,316,405
6	フェニックス	1,461,575	8,769,450
7	サンアントニオ	1,256,509	8,795,563
8	サンディエゴ	1,255,540	10,044,320
9	ダラス	1,213,825	10,924,425
10	サンノゼ	912,332	9,123,320

日本に続き，アメリカの各都市の人口もおおよそ Zipf's Law に当てはまることがわかった．ここで両データの数を増やすことにより，それが Zipf's Law にどう影響を及ぼすのかを考えてみたい．

3.1　下位データ

表 3　日本の各都市の人口 下位データ (2006/8/1)

順位 (x)	都市名	人口 (y)	$x\times y$
660	美祢	18,638	12,301,080
661	土佐清水	18,512	12,236,432
662	豊後高田	18,506	12,250,972
663	牛深	18,284	12,122,292
664	竹田	17,489	11,612,696
665	日光	17,428	11,589,620
666	両津	17,394	11,584,404
667	赤平	15,753	10,507,251
668	夕張	14,791	9,880,388
669	三笠	13,561	9,072,309
670	山田	11,686	7,829,620
671	歌志内	5,941	3,986,411

下位データについて，表 3 のように日本では Zipf's Law

に適するものとなった．次にアメリカのデータを表 4 に表す．

表 4　アメリカの各都市の人口 下位データ (2006)

順位 (x)	都市名	人口 (y)	$x \times y$
91	ラボック	207,852	18,914,532
92	モデスト	206,769	19,022,748
93	オーランド	205,648	19,125,264
94	チュラビスタ	204,879	19,258,626
95	ラレド	203,212	19,305,140
96	フレモント	202,373	19,427,808
97	ダーラム	201,726	19,567,422
98	グレンデール	201,326	19,729,948
99	モンゴメリー	200,983	19,897,218
100	シュレベポート	198,675	19,867,500
298	トレントン	85,403	25,450,094
299	ブルーミントン	85,172	25,466,428
300	シトラスハイツ	85,071	25,521,300

アメリカではデータの量を増やすにつれて，Zipf's Law とはかけ離れる結果となってしまう．これは日本とアメリカでは何が違い，それがどのようにこのような結果を生み出すのか．

日本に比べアメリカは所得格差が大きい，ということに目をつけ，日本とジニ係数の近いロシア，アメリカとジニ係数の近い中国についても，それぞれ人口のデータを取る．

表 5　ロシアの各都市の人口 (2005/10)

(x)	都市名	人口 (y)	$x \times y$
1	モスクワ	10,126,424	10,126,424
2	ペテルブルク	4,661,219	9,334,438
3	ノボシビルスク	1,425,508	4,275,524
4	ニジニノヴゴロド	1,311,252	5,245,008
5	ニカテリンブルク	1,293,537	6,467,685
996	リャンガソヴォ	13,021	12,968,916
997	ナズィバエフスク	13,012	12,972,964
998	レニンケント	12,995	12,969,010
999	ツィヴィリスク	12,967	12,954,033
1000	アク＝ドヴラク	12,965	12,965,000

表 5 のように，ロシアはデータが増えるにつれ，Zipf's Law に準ずる結果となり，逆に表 6 より中国はアメリカと同じくデータを増やすことにより，Zipf's Law との違いが生じる結果となった．このことより各国の都市人口分布と，所得格差には関係があるのではないかということが言える．

つまり，所得格差の小さい国は，Zipf's Law に従い，都市の人口差が激しい都市集中型．逆に所得格差の大きい国は，Zipf's Law に従わず，都市の人口差が小さい，ということが考えられる．このことから人口を Zipf's Law にあてはめることにより，その国の経済面のデータをも推測できる．

表 6　中国の各都市の人口 (単位:千人 2005)

順位 (x)	都市名	人口 (y)	$x \times y$
1	上海 (シャンハイ)	14,349	14,349
2	北京 (ペキン)	11,510	23,020
3	重慶 (チョンチン)	9,692	29,076
4	広州 (クワンチョウ)	8,525	34,100
5	武漢 (ウーハン)	8,313	41,565
28	南海 (ナンハイ)	2,134	59,752
29	福州 (フーチョウ)	2,124	61,596
30	長沙 (チャンシャー)	2,123	63,690
31	蘭州 (ランチョウ)	2,088	64,728

4　日本の県別の人口

つぎに，日本の 47 の都道府県について，それぞれ各都市の人口のデータを取る。

表 7　例 愛知県の都市別人口 (2006/10/1)

順位 (x)	都市名	人口 (y)	理想人口
1	名古屋市	2,222,907	2,222,907
2	豊田市	415,706	1,111,453
3	豊橋市	373,927	740,969
32	岩倉市	47,983	69,465
33	弥富市	42,698	67,360
34	高浜市	42,131	65,379

ここで，表 7 ように，実際の各都市の人口と， Zipf's Law に的確にあてはまる理想値との相関係数を各県について調べる．

表 8 に示したように，各県における都市の人口という，数少ないデータであったが，日本の都市の人口は，県別にみても Zipf's Law に適合するということがいえる結果となった．つまり日本は，国として見ても，県別にして見ても，まさに都市集中型国家となっているということがわかる．

5　ゴルフの賞金額

ゴルフの試合での順位別に出される賞金は Zipf's Law に当てはまるのか，様々な大会の賞金額を調べることとする．

5.1　定義

まず，同順位の出場者が多数いたため，同順位の中で順位をつけ，賞金金額を決める．決め方は以下の通りである．

Xi : 順位 i の賞金額

n : 同順位の人数

表 8　相関係数

県名	係数	県名	係数	県名	係数
北海道	0.9681	石川	0.9719	岡山	0.9545
青森	0.9349	福井	0.9803	広島	0.9953
岩手	0.9820	山梨	0.9774	山口	0.9121
宮城	0.9406	長野	0.9909	徳島	0.9731
秋田	0.9754	岐阜	0.9903	香川	0.9656
山形	0.9889	静岡	0.9428	愛媛	0.9858
福島	0.8695	愛知	0.9499	高知	0.9318
茨城	0.9070	三重	0.8927	福岡	0.9618
栃木	0.9811	滋賀	0.9741	佐賀	0.9858
群馬	0.8964	京都	0.9215	長崎	0.9934
埼玉	0.9907	大阪	0.9775	熊本	0.9489
千葉	0.9553	兵庫	0.9867	大分	0.9688
東京	0.9797	奈良	0.9816	宮崎	0.9964
神奈川	0.9839	和歌山	0.9578	鹿児島	0.9565
新潟	0.9841	鳥取	0.9254	沖縄	0.9869
富山	0.9925	島根	0.9623		

$\sum_{j=1}^{n} j$: 順位の和

$Yi = \frac{Xi \times n}{\sum_{j=1}^{n} j} \times$ 順位を逆から並べたもの

とし，Yi が順位ごとの賞金額とする．特に決められたやりかたではないが，色々な手法を試してみた結果一番 Zipf's Law に沿う形となったので用いることにする．同順位の賞金額を区別したのが下の表になる．

5.2　結果

このデータをカイ 2 乗検定により，帰無仮説として Zipf's Law に従わないと仮説を立て分析してみる．何個かの試合結果を分析してみたが，いずれも上位，下位の誤差が大きいため有意水準 95 % で帰無仮説は棄却されなかった．以下の表が検定した値となる．

ただ，どの大会の賞金額においても中盤では Zipf's Law に従う結果となり，また相関係数も 0.9 以上になることから非常に関係の深いものであることがわかる．そこで中盤の誤差の少ない部分である 10 位から 50 位を抜粋してカイ 2 乗検定をしてみる．調べたすべての大会において有意水準 95 %で帰無仮説が棄却され，対立仮説の Zipf's Law に従うが採択される．以上のことから条件をつければ Zipf's Law に従うことがわかる．以下の表が条件付でのカイ 2 乗検定値となる．次に，違う大会どうしの賞金額において相関係数を調べ，大会どうしの賞金額の関わりが深いか調べてみる．以下の表の通り，相関係数の高いものや低いものがあり，大会どうしの金額に深い関わりがないことが分かる．

表 9　アデランスウェルネスオープン賞金額 (単位:千円 2006)

順位 (x)	賞金額 (y)	理想賞金額	$x \times y$
1	13000.0	13000.0	13000
2	5700.0	6500.0	11400
3	4050.0	4333.3	12150
4	2859.6	3250.0	11438
5	2383.0	2600.0	11915
6	1906.4	2166.7	11438
7	1674.0	1857.1	11718
8	1488.0	1625.0	11904
9	1458.6	1444.4	13127
10	1326.0	1300.0	13260
11	1193.4	1181.8	13127
12	1180.3	1083.3	14164
13	1096.0	1000.0	14248
14	1011.7	928.6	14164
15	956.9	866.7	14354
16	897.1	812.5	14354
17	836.0	764.7	14212
18	792.0	722.2	14256
19	748.0	684.2	14212
20	701.0	650.0	14022
21	673.1	619.0	14135
22	645.0	590.9	14190
23	617.0	565.2	14191
24	588.9	541.7	14134
25	560.9	520.0	14023
26	516.0	500.0	13416
27	481.6	481.5	13003
28	466.6	464.3	13065
29	451.5	448.3	13094
30	436.5	433.3	13095
31	421.4	419.4	13063
32	406.4	406.3	13005
33	392.3	393.9	12946
34	381.4	382.4	12968
35	370.6	371.4	12971
36	359.7	361.1	12949
37	351.0	351.4	12987
38	342.0	342.1	12996
39	333.0	333.3	12987
40	325.0	325.0	13000

表 10　大会別賞金額のカイ 2 乗検定

大会名	カイ 2 乗値
ウェルネスオープン	2.88
シニアコマツカップ	5.19
ボーネストオープン	0.074

表 11　大会別賞金額 10 位〜50 位までのカイ 2 乗検定

大会名	カイ 2 乗値 (10 位〜 50 位)
ウェルネスオープン	$2.58 \times 10^{-}7$
シニアコマツカップ	$1.81 \times 10^{-}9$
ボーネストオープン	$5.27 \times 10^{-}11$

表 12　大会どうしの相関係数

大会名	大会名	相関係数
ウェルネスオープン	シニアコマツカップ	0.88
ウェルネスオープン	ボーネストオープン	0.31
ボーネストオープン	シニアコマツカップ	0.30

表 13　男女別 20 位〜80 位までのカイ 2 乗検定

性別	カイ 2 乗値
男性	$1.16 \times 10^{-}7$
女性	$1.08 \times 10^{-}8$

5.3　考察

ゴルフの賞金額は従来の Zipf's Law とは異なり，上位と下位において誤差が生じた．特に上位においては，非常に大きな誤差があった．これは上位の賞金金額が見栄えなどからきりのいい数字にするため，人為的な作為があるからだと考えられる．またどの試合も Zipf's Law との相関係数が高くなったことから，完全に Zipf's Law に従いはしないが，Zipf's Law とゴルフの賞金額には深く関わりがあるといえる．

6　ゴルフの国内賞金ランキング

男女別の国内年間賞金ランキングが Zipf's Law に従うか分析する．

6.1　賞金データの分析

まず，男性と女性の上位 100 位までの賞金額を Zipf's Law にあてはめる．

6.2　カイ 2 乗検定による分析結果

このデータをカイ 2 乗検定により，帰無仮説として Zipf's Law に従わないと仮説を立て分析してみる．分析結果は上位，下位の誤差が大きいため有意水準 95 %で帰無仮説が棄却されなかった．ここでも，中盤において値が一定になることから，20 位から 80 位を抜粋してカイ 2 乗検定を先と同じ帰無仮説を立て分析する．表より中盤においては有意水準 95 % で帰無仮説が棄却され対立仮説の Zipf's Law に従うが採択される．以上のことから条件をつければ Zipf's Law に従うことがわかる．

6.3　考察

前章と同じく，上位と下位に誤差が生じたため，全体のカイ 2 乗検定で帰無仮説を棄却しなかった．中盤も前章と同じでカイ 2 乗検定で帰無仮説を棄却する結果となった．これはゴルフが実力通りの順位になりにくいことや，すべての人がすべての試合に出場していないからだと考えられる．ここでも相関係数は 0.9 以上となったので，完全に Zipf's Law に従いはしないが，Zipf's Law の法則とゴルフの賞金額には深く関わりがあるといえる．

7　おわりに

Zipf's Law は自然現象や社会現象などあらゆる現象において，成り立つものとされているが，本研究での結果はそれとは多少違うものとなった．今回私達が研究した事象はすでに Zipf's Law に成り立つものとされている都市人口の分布と，成り立つかどうかが未知であるゴルフの賞金額において，ともにあらゆる視点からの分析・考察を試みた．

どちらの分析も上位及び下位に人為的作為が見られ，データ全体が Zipf's Law に適合するということはなかった．しかし，データ中盤においては人為的作為が加わることが少なく，順位と規模とが反比例し，Zipf's Law に適合する場合が多いことがわかった．

今回，中盤において Zipf's Law に従うデータを主に紹介してきたが，経済データなど人為的作為が多く働くデータにおいては Zipf's Law が適合しないということがわかっている．このように人為的作為の度合いなどにより Zipf's Law の適合度に差が出るということが今回わかった．これを元に，今後も Zipf's Law の研究が進めていければと考えている．

参考文献

[1] http://www.neo-luna.cside.tv/population/(都 市 人口ランキング)

[2] http://www.jaist.ac.jp/kshirai/lec/i223/10.pdf(言語処理理論統計的文語分析)

[3] http://s-nobu-web.hp.infoseek.co.jp/tiri.html(地 理に関するランキング)

[4] http://www.bloomberg.co.jp/markets/currencies/(bloomberg)

[5] http://www.city.yokosuka.kanagawa.jp/data/t-k-syo/16080

[6] http://www.jma.go.jp/jma/(気象庁)

[7] フリー百科事典『ウィキペディア (Wikipedia)』

[8] 世界統計白書 木本書店

[9] http://www.h-yamaguchi.net/(H-Yamaguchi.net)

[10] http://www.pga.or.jp/(日本プロゴルフ協会)

[11] http://www.lpga.or.jp/(日本女子プロゴルフ協会)

[12] 尾崎俊治：確率モデル入門，朝倉書店 (1996)

[13] 長畑秀和：統計学へのステップ，共立出版 (2000)