The Statistical Analysis of Family Names of Donators For WenChuan

Liu Yufan[1], Chen Qinghua[2]
[1]Shijjiazhuang University of Economics
Shijiazhuang, 050031
[2]School of Management
Beijing Normal University
Beijing, 100875

Journal of Digital Information Management

ABSTRACT: *It is analyzed that the family names of personal donators who donated through China Construction Bank Corporation to Chinese Red Cross Foundation for the WenChuan earthquake. The distribution of family names, the first 100 family names and their shares, the probability of the same family names as well as the Gini coefficient are all given in this paper. A heavy disproportion is showed in the distribution, that is to say a small part of family names occupy the absolutely large proportion while most of them are rare. This character is similar to power-law that has been observed in many demonstrations about family names. However, the Kolmogorov-Smirnov test fails to support the hypothesis. Instead, part of the data shows an exponential characteristic. As an innovative sampling method, the discussion may help to know the family name's distribution of China.*

Categories and Subject Descriptors:
I.2.7 **[Natural Language Processing]** I.7 **[DOCUMENT AND TEXT PROCESSING]** *Languages*

General Terms:
Name Analysis, Natural Language Processing

Keywords: The Family Name, Distribution, Exponential, Power-law

Received: 5 February 2013, Revised 17 March 2013, Accepted 25 March 2013

1. Introduction

General, last name, passed down from his father, so you can put it as a biological marker for population genetics research, surname, as research for the specific model of the *Y* chromosome has a long history [1-3]. Surname as a special kind of cultural symbol, it is not only used for groups or individual identification of the members of society, and carries a national language features, the history, geography, culture, religion, status and class division in its internal information. Chinese surname for at least 4,000 years of history, its evolution with the evolutionary history of the Chinese nation is synchronized surname distribution and dynamics of the process has a very important significance. Statistical analysis as an important method for human studies of complex systems has been widely used in the study of names. The statistical properties of the surname may reflect some of the structure of human populations [1], provides the necessary basis for further discussion about the mechanism of the evolution of this complex system.

There are a lot of empirical work demonstrating the surname distribution law of some countries or regions, and even the dynamic behavior of the differences between the regions. S. Miyazima et al studied Japan's five different regional communities, they found that these areas surname distribution has the characteristic: With the changes in the population, the total number of names have also changed, and they have a power-law relationship; Surname distributed as a power-law distribution, have more individual surname species; 3 from the population, the surname ranking corresponding number of Zipf law, in approximately a negative slope of the straight line in the double logarithmic coordinates [4]. DH Zanette and others to discuss the actual data of the nation's cities and Berlin, and found that the size of the surname (which

includes the number of individuals) is power-law distribution, and both have a similar index [5]. WJ Reed mentioned in his article, Taiwanese distribution of surnames in line with the 1.9 power-law distribution of index [6]. BJ Kim and SM Park, a Korean surname has a power-law distribution characteristics, and the index is approximately 1 [7]. C. Scapoli et al detailed examination of the 2094 cities in 125 regions of the eight countries in Western Europe, found although each city is different, but the overall surname distribution for the power-law distribution [8].

Some Chinese surnames object statistics, Yuan Yida, a major work in this regard [3, 9-11]. They found that there are two kinds of state, Chinese surname common surnames and uncommon surnames. Only accounted for less than 5% of the total surnames amount of 100 common surnames concentration of more than 85% of the population, while more than 95% of the amount of the total surnames uncommon surname representatives of less than 15% of the population. Some of the work is given to the Song, Ming and the surname distribution curve, but not the power-law distribution, but fitting as follows [9,10]:

$$S = 0.185 \ln [(r + 1) / r]$$

Yuan Yida also pointed out that the distribution of surnames exist stability will be no major changes in long-term [11]. Chinese Academy of Sciences, was released in January 2006, chaired by Yuan Yida Chinese surnames Statistics "project findings, the range of data, including China's 31 municipalities, provincial capitals, the sample of the total population of 202 million people, were found in 4100 surname. Ranked the top three Li, Wang, Zhang, accounted for 7.4%, 7.2% and 6.8% of the total population in China. Tube Limin Ningxia region as a whole Han, Hui and other ethnic groups as well as the entire region are listed in the top 100 surnames based on a sample survey of 1% of the country in 2005, [12]. The article also details the Yinchuan, Wuzhong City, Guyuan City and Zhongwei City "*Family Names*".

These works are based on statistical sampling method, randomly selected from a population of sufficient sample to infer the nature of the overall, by discussing the nature of the sample. The sample data are generally from the census or the actual long-term investigation. Wenchuan earthquake personal donor information provides us with a simple and easy sampling because such contributions is spontaneous from across the country can guarantee the basic sampling randomness, and more than two hundred thousand of the amount of data for surnames roughly statistics is sufficient. In this paper, the statistical analysis of earthquake donors name surname, calculate the rate of the same surname and Gini coefficient, and its distribution characteristics. The statistical results show that our very uneven distribution of surnames, the surname of a small part accounted for the vast majority of the population share. Discuss the results do not support the surname obey a power law distribution assumptions.

2. Data Sources and Processing

May 12, 2008, Wenchuan County, Sichuan Province, China occurred in the vicinity of the strong earthquake of 8.0 Richter. Community to act quickly to meet basic livelihood security of the people in the disaster areas and support post-disaster reconstruction, spontaneously donated money and goods. The Red Cross Society of China and the Chinese Red Cross Foundation announced a donation bank account and opened a variety of contributions to the path. The Chinese Red Cross Foundation has received a large number of China Construction Bank, Bank of China, Industrial and Commercial Bank of China into the personal and corporate donations.

In order to facilitate the donor's verification donation information as well as to meet the needs of the social supervision, the Chinese Red Cross Foundation will be part of the donor information on the Home http://www.crcf.org.cn/ for download. We visit the site on June 10, 2008, the businesses and individuals between May 13, 2008 to June 4, Construction Bank remittance to the Chinese Red Cross Foundation account information, including the name of the donor amount and donate time, a total of 250,500 of the original records. Name information extracted from these data are not personal real name be addressed, in order to meet the needs of statistical analysis.

2.1 Processed as follows:

(1) Remove all spaces in the record, including the name before and after, and which. "*Zhu Yan*", Jia GuangRui, "*Wong Wing Kin*".

(2) Delete the records of all the name is longer than four characters or character. The name of the Chinese people is generally two words or three words [13], while the name of the enterprise or unit, such as "*Zhenxiong County, Yunnan Province Chamber of Commerce and Industry. Even some companies use referred Guangshan driving school*" "*Haiyuan sound, but often not less than four words, so we first operation is hope leaving only part of personal names, and delete all non-name record. But in fact, this operation will cause the part of the individual's name was mistakenly deleted, for example, "Huangfu Yongjin" Zhang Qin civilization Dequan, Zhang Xi* 1101031971 ******** ", "*Zhang Haizhou* 1,351,107 ** *like*. ** " Containing the phone number or ID number of records or hyphenated words we will recover in the following steps.

(3) Restore the name, phone number or ID number. Find from the original name of the length of more than 13 records, including numbers and Chinese characters, hand-processed, the delete name phone number and ID number, the name re-added to the current database.

(4) Restore the name hyphenated long name. To determine length of 4 and above, whether the name of the hyphenated the beginning, if this name restored to the current library. We consider the hyphenated "*Sima*", "*Shangguan*,"

Ouyang "," *Jaap de Hoop*, "*Zhuge*," *Huangfu*" 95.

(5) Remove not the real name of the record. This process requires data statistics and artificial row seized combination repeated. Ultimately, we are determined not name names excluded, for example: "*Anonymous*", "*love*", "*unknown*", "*peace*", "*Gansu*", "*Yichang*", "*student*", "*rescue*", "*Christian* "," *lady* "," *good people* "," *dust* "," *family* ".

After the above processing, we end up with 221,800 valid records, 44 of which only the last name does not have a name.

3. Surname Statistical Distribution

Currently, on the Internet there is a lot of information about Chinese surnames, the pages collected a total of 1,057 surnames http://tieba.baidu.com/f? kz = 229923872 95 hyphenated. We will get the names of these database match, the first match longer hyphenated name, and then match the short hyphenated name, followed by single surname. "*Ouyang Jian*" We think that is the surname "*Ouyang*" name "*health*", "*European*" also can be used alone as the surname. Surnames collection is not complete, some names cannot match, and we will be the first word in its name as the last name of the donor. So we actually added some surnames, for example, "*pay*", "*Lo*", "*g*", "*Huan*", "*slag*", "*Luo*", a total of 1,077 surnames.

According to statistics, the sample surname "*king*" of the number of up to 16,867 people, accounting for about 7.61% of the total sample, "*Lee*" and "*Zhang*", followed by frequency for 15591 and 15509, respectively. 100 surnames and their frequency, as shown in Table 1. Seen by the statistical knowledge about the range of the multiplier indicators estimated for not repeat sampling conditions

$$P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}\sqrt{\frac{N-n}{N-1}} \approx P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

So that we can get the entire Chinese surnames 95% confidence interval, such as the "*king*" surname (0.0751, 0.0770), the surname "*Lee*" (0.0694, 0.0712), "" surname Zhang (0.0690, 0.0708).

Statistics, the surname distribution severe imbalance. More types of surnames arranged by size (number of samples), a rapid decline in the number of its size with the rankings, and small-scale surname. 6 people surnames such as similar to "*guarantee*" 31, while the surname "*Huyan, caught only a*" *dry* ", only a sample of nearly 300 surnames. The initial grasp us through some simple macroscopic quantity of this nature, for example, the same surname rate and the Gini coefficient. The same surname rate of concept first raised by JF Crow et al and said random encounters within a group of two people have the same surname probability [14].

Our data is calculated as 0.0264, similar to the results in the literature [10] 0.0276. In other words, arbitrarily selected two namesake probabilities close to 3%. This probability is quite large, the reason is the uneven distribution of the surnames. Gini coefficient [15] can also be used to measure the uneven distribution of surnames in Figure 1, we will all samples by last name into different components, these components in descending order of population, the cumulative population share and surname share relationship. Gini coefficient is defined as the area of the region 2 in FIG ratio:

$$G = \frac{S_A}{(S_A + S_B)} = 0.91$$

This means that the distribution of the surname serious inequality The above analysis shows that the majority of the population does focus on a small number of surnames, the surname is rare, very few people with these surnames. This nature is very similar to a power-law distribution, and this distribution has been found to exist in the distribution of surnames in different countries [4-8]. But the data we are discussing whether to obey the power law distribution strict estimation and hypothesis testing, the currently used method of maximum likelihood parameter estimates and the Kolmogorov-Smirnov test [16]. The results show that our data cannot pass the Kolmogorov-Smirnov test, so cannot accept the assumptions of the surnames of these donors with power-law distribution characteristics. Figure 2 is the diagram of the relationship between the rank and the number of different coordinate more in line with a power-law behavior at the bottom, while the middle segment (100-400 members) is more in line with the exponential distribution, this distribution pattern is not a simple function form will be described.

The 100 surnames ranked scale relationship is shown in Figure 3. Consider the equation Yuan Yida et al [12], the actual data can be roughly fitted:

$$S = 0.156 \ln[(r+1)/r]$$

The coefficient of determination $R^2 = 0.88$ However, the figure can be clearly found ranked 20-100 surname share a strong exponential relationship, the functional form of the blue line in Figure:

$$S = e^{-[(r+227.287)/50.747]}$$

The goodness of fit of 0.97. The saying this exponential distribution characteristics and the literature [1].

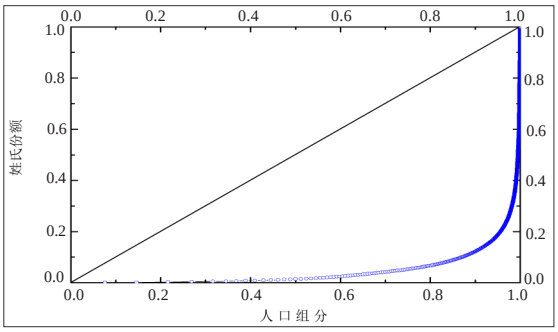

Figure 1. Gini coefficient calculated of surname

Ranking	Surname	Frequency%	Ranking	Surname	Frequency%	Ranking	Surname	Frequency%	Ranking	Surname	Frequency%	Ranking	Surname	Frequency%
1	Wang	7.61	21	Song	0.78	41	Wei	0.49	61	Jia	0.34	81	Duan	0.22
2	Li	7.03	22	Luo	0.77	42	Tian	0.49	62	Zou	0.34	82	Wu	0.22
3	Zhang	6.99	23	Liang	0.76	43	Pan	0.49	63	Meng	0.34	83	Wan	0.21
4	Liu	5.28	24	Xie	0.69	44	Lv	0.48	64	Fang	0.33	84	Qian	0.21
5	Chen	4.74	25	Xu	0.69	45	Du	0.48	65	Shi	0.33	85	Shao	0.21
6	Yang	2.96	26	Han	0.67	46	Jin	0.48	66	Lu	0.31	86	Tao	0.21
7	Huang	2.25	27	Yu	0.63	47	Ding	0.46	67	Fu	0.30	87	Hao	0.21
8	Zhao	2.10	28	Tang	0.63	48	Shen	0.45	68	Xue	0.30	88	Hong	0.20
9	Wu	2.05	29	Feng	0.61	49	Yu	0.44	69	Bai	0.29	89	He	0.20
10	Zhou	1.92	30	Cao	0.58	50	Su	0.44	70	Jiang	0.29	90	Mao	0.19
11	Sun	1.63	31	Cai	0.56	51	Jiang	0.44	71	Liao	0.28	91	Tang	0.19
12	Xu	1.62	32	Deng	0.56	52	Cui	0.44	72	Qiu	0.28	92	Gong	0.19
13	Zhu	1.35	33	Dong	0.55	53	Yao	0.42	73	Qin	0.28	93	Lei	0.18
14	Lin	1.34	34	Xiao	0.54	54	Ren	0.41	74	Xiong	0.28	94	Yan	0.18
15	Guo	1.22	35	Ye	0.52	55	Lu	0.40	75	Yi	0.27	95	Shi	0.18
16	Hu	1.20	36	Cheng	0.52	56	Wang	0.39	76	Hou	0.26	96	Li	0.18
17	Ma	1.16	37	Yuan	0.51	57	Fan	0.39	77	Shi	0.25	97	Chang	0.17
18	Gao	1.09	38	Peng	0.51	58	Zhong	0.37	78	Yan	0.25	98	Wen	0.17
19	Zheng	1.08	39	Zeng	0.50	59	Xia	0.36	79	Dai	0.24	99	Kong	0.17
20	He	0.96	40	Jiang	0.50	60	Tan	0.36	80	Gu	0.23	100	Guan	0.16

Table 1. 100 surnames and their frequency

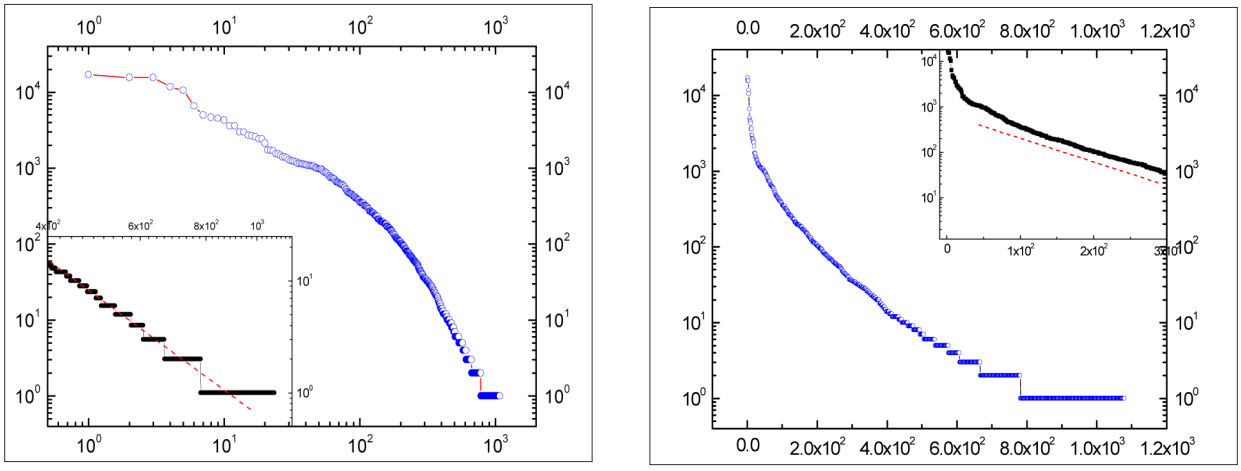

Figure 2. Relationship of surnames ranking and the scale

4. Conclusions

The individual donors based on the Wenchuan earthquake surname were analyzed, the surname distribution characteristics of this group. Frequency comparison, the calculation of the same surname rate surname Gini coefficient, we get the conclusion of the group distribution of surnames serious imbalance, the Kolmogorov-Smirnov test results do not support the distribution of surnames to obey a power law distribution of the original assumptions, but has a more complex form of the curve . As a statistical sampling by the surname distribution characteristics of a certain understanding of the surname. This article is only a superficial discussion of some related theoretical measure of inequality and the distribution characteristics of the groups surname statistics on the application, there

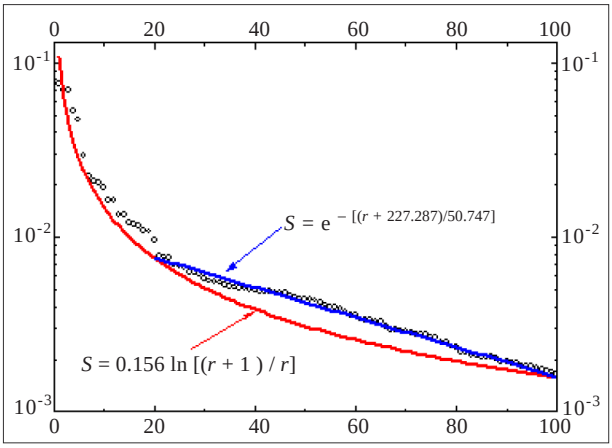

Figure 3 Relationship of "*Surnames*" ranking and scale

are a lot of issues need to be careful analysis and expand, for example, how to accurately portray the surname distribution characteristics and behind it in the end there is what kind of mechanism.

5. Acknowledgment

The work is supported by the Natural Science Fund of China under Grant Nos. 70601002, 60534080.

References

[1] Baek, S. K., Kiet, H. A. T., Kim, B, J. (2007). Family name distributions: Master equation approach. Phys Rev E. 76, 046113

[2] Lasker, G. W. (1985). Surnames and genetic structure [M]. Cambridge: Cambridge University Press.

[3] Yuan Yida, Zhang Cheng, Ma Qiuyun. (2000). Chinese surnames population genetic: I Surname Frequency Distribution and Genetic differentiation. *Genetics*, 27 (6) 471-476

[4] Miyazima, S., Lee, Y., Nagamine, T. et al. (2000). Power-law distribution of family names in Japanese societies. *Physica A*, 278, 282-288.

[5] Zanette, D. H., Marunbia, S. C. (2001). Vertical transmission of culture and the distribution of family names. *Physica A*, 295, 1-8.

[6] Reed, W. J., Hughes, B. D. (2003). On the distribution of family names. *Physica A*, 319, 579-590

[7] Kim, B. J., Park, S. M. (2005). Distribution of Korean family names. *Physica A*, 347, 683-694

[8] Scapoli, C., Mamolini, E., Carrieri, A. et al. (2007). Surnames in Western Europe: A comparison of the subcontinental populations through isonymy. *Theoretical Population Biology*, 71, 37-48.

[9] Yuan Yida, Zhang Cheng, Yang Huanming. (2000). Chinese population genetic surnames: the stability and geographical II Surnames Consanguinity of Population Frequency Distribution and Genetic differentiation. *Genetics*, 27 (7) 565 -572.

[10] Yuan Yida, Zhang Cheng. (2002). Chinese surnames: population genetic and demographic distribution. Shanghai: East China Normal University Press.

[11] Yuan Yida. (2003). three surnames in China is how the statistics out of. *Statistics of China*. (07) 32-34.

[12] Guan Limin. (2006). 1% of the population survey reveal: Ningxia surname "*Ma*" headed by. Statistics and Economic, (1) 63-64.

[13] Zhang YangSen, Xu Bo, Cao Yuan, etc. (2003). Chinese name based on surname driven automatic identification method *Computer Engineering and Applications*, 39 (4) 62.

[14] Crow, J. F., Mange, A. P. (1965). Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugenics Quarterly*, 12 (4) 199-203.

[15] Lambert, P. J., Aronson, J. R. (1993). Inequality Decomposition Analysis and the Gini Coefficient Revisited. *The Economic Journal*, 103 (420)1221-1227.

[16] Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46 (29) 323-351

[17] Vani, M. P. (2011). Computer Aided Interactive Process of Teaching Statistics Methodology - II, *IEIT Journal of Adaptive & Dynamic Computing*, (3) :18-21, Jul, DOI=10.5813/ www.ieit-web.org/IJADC/2011.3.4.

[18] Vani, M. P. (2011). Computer Aided Interactive Process of Teaching Stayistics Methodology – III Evaluation Questionnaire for LearnersThrough statistical display using Bar chart, *IEIT Journal of Adaptive & Dynamic Computing*, (4) 9-14, Oct. DOI=10.5813/www.ieit-web.org/IJADC/20114.2.

[19] Sumathi, P., Vani, M. P. (2011). Computer Aided Interactive Process of Teaching Statistics Methodology - I, *IEIT Journal of Adaptive & Dynamic Computing*, (2) 1-6, Apr. DOI=10.5813/www.ieit-web.org/IJADC/2011.2.1.