Applied Probability

School of Mathematics and Statistics, University of Sheffield, Sheffield S3 7RH, UK
Telephone: +44 114 222 3920
Fax: +44 114 272 9782
Http: www.appliedprobability.org

Journal of Applied Probability
Advances in Applied Probability
The Mathematical Scientist
Mathematical Spectrum

Dear Author,

Accompanying this are the proofs of your paper, which is scheduled to appear in *The Mathematical Scientist,* Vol. 31, No. 2 (December 2006). Please give them your immediate attention.

If any corrections are necessary, mark them clearly on the proof. Corrections in proof are expensive and time consuming, and should be kept to a minimum. Note that these are not 'page proofs', and so the final layout of your paper may differ.

As well as looking for typographic errors, pay particular attention to the following points:

- Ensure that the postal addresses at the foot of the first page are up-to-date. However, your affiliation (which follows your name in the title) should remain as it was when the paper was written.
- Check that the short title and abbreviated names chosen for the running heads are suitable.
- Look at the references again to check whether any 'preprints', 'unpublished manuscripts' etc. have since been published, or 'submitted' papers accepted.
- It is the house-style of the Trust to use:
  - e rather than (exponential) e, i rather than (imaginary) i and d rather than (differential) d;
  - bold italics for vectors and matrices, e.g. $\boldsymbol{v}$, $\boldsymbol{M}$, where they require to be distinguished from scalar values, and $^\top$ for the transpose (rather than $'$, ^{t}, ^{T} etc.);
  - $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$ and $\mathbb{N}$ for, respectively, the real, rational, integer and natural numbers;
  - P for probability statements (rather than P, $\mathbf{P}$, $\mathbb{P}$ etc.), E for expectation (rather than E, $\mathbf{E}$, $\mathbb{E}$ etc.) and $\mathbf{1}_A$ for the indicator function, or 1_A when $\mathbf{1}$ clashes with vector notation (rather than I_A, I_A, χ_A etc.);
  - and to avoid excessive use of abbreviations.

  Check that, where we have made such changes, we have done so completely, consistently and correctly.

Please sign and return one set of corrected proofs *within 3 working days of receipt* by airmail or express (to arrive within 1 week of dispatch) to:

Journals Production
Applied Probability Trust
School of Mathematics and Statistics
University of Sheffield
Sheffield S3 7RH
UNITED KINGDOM

Our production schedule allows for very little delay. We will incorporate authors' corrections if at all possible, but, if these are not received promptly, we reserve the right to proceed with only the corrections noted by our own proofreaders.

Please return a copy of your proofs even if you make no corrections

If you have any problems or queries, then please contact our Sub-Editor, Helen Mason (email address: h.m.mason@sheffield.ac.uk).

Thank you for your cooperation.

Math. Scientist **31,** 1–4 (2006)
Printed in England

PROBABILISTIC ANALYSIS OF NUMBERS OF NAMESAKES IN A LARGE POPULATION

QINGZHI YAO,* *Xinzhou Teachers University*
YUYUAN ZHAO,** *The University of Liverpool*

Author:
this has been re-phrased – OK?

Abstract

This short article describes a probabilistic model for analysing the numbers of namesakes (i.e. persons with identical names) in a population. The model is applied to the Chinese population, which has an extremely high number of namesakes. A practical measure to reduce the numbers of namesakes is proposed.

Keywords: Namesake; population; probability

2000 Mathematics Subject Classification: Primary 92D99
Secondary 60D05

1. Introduction

People are identified by their names. A name is normally composed of a surname or family name and one or more given names. While surnames are normally inherited from parents, given names are chosen by the parents. As a symbol of identity, a name should ideally be a unique combination of a surname and one or more given names. In a large population using the same language or having the same culture, however, it is unavoidable that many people should share identical names, or be namesakes, owing to the relatively limited number of common given names.

The existence of many namesakes is a common phenomenon in many societies. It is a particularly big problem in China, which is the most populous country in the world with a population exceeding 1.3 billion. Apart from the large population, the relatively rigid form of Chinese names also contributes to the problem. In comparison with Western societies, Chinese society has a relatively small number of surnames. The most common 100 surnames cover 87% of the population and the top 19 surnames cover 55.6% (see [5]). The most frequent surname, Li, accounts for 7.9% of the Chinese population (see [5]), which is nearly twice the population of the United Kingdom. The availability of choices for given names is not sufficient to offset the problem. The overwhelming majority of Chinese individuals have a given name consisting of either one or two Chinese characters. The number of Chinese characters suitable for use in names is limited. Modern Chinese names only use 3 356 characters. The most common 409 characters cover 90% of the names and the top six cover 10% (see [5]). As a consequence, a large number of individuals have the same popular names, which can pose many problems. The situation is further exacerbated when Chinese names appear in English in the

Received 1 November 2005; revision received 18 November 2005.
* Postal address: Department of Chinese Language and Literature, Xinzhou Teachers University, Xinzhou, 034000, P. R. China.
** Postal address: Department of Engineering, The University of Liverpool, Liverpool L69 3GH, UK.
Email address: y.y.zhao@liv.ac.uk

pinyin form where the Roman alphabet is used, because each pinyin can represent several dozen Chinese characters. It therefore seems worthwhile to analyse the frequency of namesakes in a population.

In this short article, we describe a probabilistic model for predicting the possible numbers of namesakes in a population, and discuss the practical measures to reduce the numbers of namesakes. For convenience, this article will focus on Chinese names. Each individual has one given name, which is composed of several Chinese characters. This approach, however, is applicable to other forms of names. For instance, a given name in English is analogous to a Chinese character and the permutation of all the given names of a person is analogous to a given name in Chinese.

2. The model

The random distribution of points over an area, which we use to represent a particular name, can be described by the Poisson distribution (see [7]). If the density of points on the area, i.e. the number of points per unit area, is λ, then the probability that there are k points falling on a unit area is

$$P_k = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \qquad k = 0, 1, \ldots.$$

The naming process can be regarded as following the Poisson distribution if the names are chosen randomly. Let us consider a population of m individuals who share a certain surname. Assuming that each individual has a given name consisting of x characters, each of which is chosen completely at random from a character set of a fixed number of characters, n, the total number of possible given names (or permutations of x characters) is n^x. The average number of individuals using a unique given name is analogous to the point density and is therefore

$$\lambda = \frac{m}{n^x}. \tag{1}$$

A given individual has no namesakes if either nobody or only one person uses the same name. With the assumption of random choices of characters, each given name is equally probable. The ratio P of individuals without namesakes in the population is thus equal to the probability that a given name is used either by nobody ($k = 0$) or by one person ($k = 1$) and can be expressed as follows:

$$P = P_0 + P_1 = (1 + \lambda)\mathrm{e}^{-\lambda} = \left(1 + \frac{m}{n^x}\right)\exp\left(-\frac{m}{n^x}\right) \tag{2}$$

3. Applications and discussion

The probabilistic model should not be applied directly to a population with mixed surnames, because surnames are normally predetermined. The model is applicable to a population with the same surname. Populations with different surnames must be treated separately. Fortunately, the relative proportions of the populations with different surnames do not vary much and statistical data is often readily available from censuses. This should not pose a problem for the analysis.

From a probabilistic point of view, the proportion of individuals without namesakes, P, is determined by the average frequency of the use of a given name, i.e. the point density, λ. It is shown in (2) that P decreases almost exponentially with λ. When $\lambda = 0.1$, 0.5, 1, 5, and 10, we have $P = 0.9953$, 0.9098, 0.7358, 0.0404, and 0.0005 respectively. The point density, λ, in turn is determined by the sample size of the population, m, the number of characters suitable

TMS31.2
1427

for names, n, and the number of characters customarily used in a person's given name, x. The effects of these three parameters are now discussed separately.

As shown in (1), the point density λ is directly proportional to the population considered, m. Popular surnames have big sample sizes and therefore large numbers of namesakes. For example, the population of Chinese with the surname 'Li' is estimated to be around 103 million (see [5]). Given that there are about 400 most frequently used Chinese characters for given names ($n = 400$) (see [5]) and that the given names most commonly use two characters ($x = 2$), the point density for the Li community is $\lambda \approx 51$. The probability that a person with the surname Li has no namesakes is $P \approx 4 \times 10^{-21}$. In other words, almost every Li will find that he or she has at least one namesake. When the sample size is reduced, for a less common surname and/or a smaller region considered, the number of namesakes is also reduced. In a typical Chinese city with a population of 1 million, for example, the point density for individuals with the surname Li is reduced to $\lambda \approx 0.5$. The probability that a person with the surname Li has no namesakes in the city becomes $P \approx 0.9$ (90%). In other words, 10% of the Lis will still find that they have at least one namesake in the city. In a small town of 10 000 people, the point density for the Li community is reduced again to $\lambda \approx 0.005$. The probability that an individual with the surname Li has no namesakes in the city is increased to $P \approx 0.995$ (99.5%). Only 0.5% of the Lis in the town, or four Lis, may find that they have a namesake.

The range of available choices of characters for names has a more marked effect on the number of namesakes, providing that they are chosen randomly. The higher the available number of characters, the lower the number of namesakes. Take a typical Chinese city, where the population of Lis is around 10 000 ($m = 10\,000$) and everybody is assumed to have a two-character given name, for example, $x = 2$. When the numbers of characters available for given names (i.e. values of n) are 10 000, 1 000, 100, or 10, the point density, λ, is approximately 0.0001, 0.01, 1, or 100, and the probability that an individual with the surname Li has no namesakes, P, is approximately 99.99%, 99.00%, 73.58%, or nearly 0 respectively. However, the choices of characters for names are limited in number and have varied popularity. Although the Chinese vocabulary has a large number of characters exceeding 56 000 (see [8]), the number of commonly used characters is much smaller. The Chinese character libraries for computer word processing in the Chinese National Standards GB18030 (see [4]), GB13000.1 (see [3]) and GB2312-80 (see [2]) contain 27 000, 20 092, and 6 763 characters respectively. Among these, only 3 000 characters are commonly used (see [1]). Personal names are concentrated on an even smaller number of characters which have commendatory or neutral meanings. As a consequence, the most frequently used 409 characters cover 90% of the names and the top six cover 10% (see [5]).

Author:
is this the
correct
reference?

The number of characters to form a given name has the most marked effect on the number of namesakes. The statistics of a population of 570 000 sampled in a Chinese national census showed that 8.7% of people have one-character given names and the majority of the rest have two-character given names (see [6]). Although the former is less than one tenth of the latter in number, the proportion of namesakes of the former is 67.7% while that of the latter is 32.4% (see [6]). Clearly, increasing the number of characters in a given name leads to a dramatic reduction in the number of namesakes. This can be demonstrated by the probabilistic model. Suppose that there are 400 characters that are suitable for use in names. For a population of 10 000 people with the same surname ($m = 10\,000$), the point densities, λ, for one- ($x = 1$), two- ($x = 2$), and four-character names ($x = 4$) are approximately 25, 0.0625, and 3.9×10^{-7} respectively. The probabilities, P, that individuals with one-, two-, and four-character given names in this population have no namesakes are approximately 0% (3.61×10^{-10}), 99.81%,

TMS31.2
1427

and 100% (0.999 999 6) respectively. If the population is increased to 100 million ($m = 10^8$), the probabilities, P, that the people with one-, two-, and four-character given names in this population have no namesakes are approximately 0% ($10^{-108\,569}$), 0% (10^{-268}), and 100% (0.999 992) respectively.

To sum up, the most effective measure to reduce the number of namesakes is to increase the number of characters in given names from one or two to four. This new measure would reduce the probability of the existence of namesakes to about eight in a million even for the most frequent surnames. This idea is not a new one. Traditionally, most well-educated Chinese had two given names, one given at birth by the parents and the other often chosen in adulthood by the individual. While the name chosen at birth could consist of one or two characters, the name chosen in adulthood almost invariably consisted of two characters. It is only in the last couple of centuries that the second given name was abandoned. Reviving the tradition is a feasible approach to solve the present problem.

4. Conclusion

The number of namesakes in a population can be analysed using the Poisson distribution. The proportion of namesakes in a population decreases with decreasing population size, an increase in the choices of given names and in the number of given names for a person. Adopting two two-character given names for each person would be an effective way of reducing the number of namesakes in China.

Author:
is [1] also by the Chinese National Standard Bureau?

References

[1] Chinese National Language Commission (1988). Table of Commonly Used Chinese Characters (in Chinese).
[2] Chinese National Standard Bureau (1981). Chinese Internal Code Specification, GB2312 (in Chinese).
[3] Chinese National Standard Bureau (1995). Chinese Internal Code Specification, GB13000.1 (in Chinese).
[4] Chinese National Standard Bureau (2000). Chinese Internal Code Specification, GB18030 (in Chinese).
[5] Du, R. and Yuan, Y. (1987). Statistical analysis of Chinese names. *People's Daily*, 3 May 1987, page 3 (in Chinese).
[6] Institute of Languages, Chinese National Language Commission (1988). Sampling statistics of names of the third national census (in Chinese).
[7] Kendall, M. G. and Moran, P. A. P. (1963). *Geometrical Probability*. Griffin, London.
[8] Xu, Z. (ed.) (1990). *Great Chinese Dictionary*. Hubei Dictionary Publisher, Wuhan (in Chinese).