Open Science Repository Mathematics

doi: 10.7392/openaccess.23050440

Fitting a Distribution for Length of Last names in a mixed African community in Ghana

Osborn A. Y. Jackson, Dominic Buer Boyetey, Richard K. Bemile, and Edward Acheampong

Department of Mathematic & Statistics, Methodist University College Ghana (MUCG)
July 2013

Abstract

The uniqueness of surnames makes it curious for continuous studies to be made on them. In Ghana surnames are dominantly inherited from the male side except some few ethnic groups that acquire the surname of the mother. The pronunciation and spelling of these surnames are complex because of their length. It is by this reason that designers of computer database programmes may allocate far in excess of needed scarce computer memory for the storage of a surname. The study postulates a probability model that nearly fits the number of characters contained in a surname for over 3263 sampled students from Methodist University College Ghana. The distributions are the Loggamma and the Dagum distributions. Their descriptive properties are very close to the actual values of the length of surnames in the study area.

Keywords: surname, character, distribution,

Introduction

Identification of all things in the universe is by name. There is no single item or object, or person or occurrence or place that has no name. Without a name nothing exists. All inventions, discoveries and natural occurrences have names of some kind. Rudolf Christian Diesel (1858-1913) invented the diesel engine and it was named after him. Sir Isaac Newton (1642-1727) discovered the laws of motion, thereafter known as the Newton's Laws of Motion. Recent occurrences of natural disasters such as hurricanes, tornados, typhoons were equally named perhaps for historical reference. We take note of a few; Hurricane Sandy (2012) in the USA, the Indian Ocean tsunami in 2004, the Rio de Janerio floods and mudslide (2011) in Brazil. Naming is such an important act that it found its practice right at the beginning of creation. Religious studies established that God created the heavens and the earth and all that is contained in and named every single constituent in the creation. The first man was named Adam (Genesis 2: 19, KJV). This practice has continued to this present time and uniquely every society has its own way of naming a new born person, place and unusual occurrences.

In the human society, a persons' name is usually defined by the religious background of the parental ancestors, the ethnic factor, race and the geographical origin of the person. Modern computerized forms demand three names, the first name, middle name and the surname. In the African traditional set up the surname and the middle name tells of the ancestral home of the person. The name tells the day the person was born and some peculiar occurrences surrounding the birth both before and after. Until recently most European names were identified by the first name which is usually of a Christian religious background.

In Spain, and recently adopted by people of African origins, Blanden et al (2005) noted that people have two surnames. The first one is the surname of the father and the second one is the surname of the mother. The two surnames are usually separated with a hyphen. Studies have found something

special about surnames. Collado and Ortuno-Ortin (2011) established an association between the status of individuals in a population and the status of their corresponding descendants based on the use of information contained in the surnames. They found surnames useful in their study because their distribution and the distribution of socioeconomic characteristics among people are not independent. In a study of social mobility in England, Clark (2009, 2010) findings showed that socioeconomic bias in the distribution of surnames disappears with time.

The uniqueness of surnames makes it curious for continuous studies on them. Lafuerza and Toral (2011) established a differential equation model for the evolution of surnames distribution under gender-equality measures. The model was to address the assignment of children surnames if the parents disagree to the surnames their ward will inherit. Healy (1968) fitted the length of surnames using the lognormal distribution. His study was critical in the allocation of a fixed number of alphabetic characters to each name for a computer file. Also, Fox and Lasher (1983) in a different approach fitted a discrete Pareto distribution for the occurrence of surnames in an area of Reading, United Kingdom.

The usefulness of surnames distribution cuts across many fields. In Population Genetics, information contained in surnames is used to analyse geographical mobility and mating structure in a population. In the Health Sciences, surnames are used to study the relationship between levels of inbreeding and the prevalence of certain types of tumor and other diseases of genetic origin (Holloway and Soafer (1992)). In economics, the distribution of surnames is applied in the analysis of every specific discrimination and social integration problems (Einau and Yariv (2004)).

Unfortunately in Ghana and perhaps Africa not much study can be found on the distribution of surnames. Surnames in this part of the world are dominantly inherited from the male side. Their pronunciations are complex because of the length of the name. Designers of computer database programmes allocate far in excess of needed scarce computer memory for the storage of a field (surname, first name, etc). The excess memory allocated for the field could have been used for other fields or left unassigned to improve the computer's performance. The extra space may prevent performance hit to applications (Larson, (2007). Teal (2008) also emphasised the need for a computer to have enough free space to be able to comfortably write out large amount of memory all at once. In practice, Dubrava (2008) recommends the keeping of at least 20% of a hard drive capacity free. Thus, just storage space for the length of a field name will enable achieving this 20% free space or even better.

In this regard the study set the stage for the study of the distribution of surnames in a university community in Ghana which have students from about twelve different African countries. The study formulates and estimates a probability model that fits the number of characters contained in a surname for over 3263 sampled students from Methodist University College Ghana.

Methodology

Data

The studied community has a population of about 5,300 students as at the end of the 2012/2013 academic year. Students of Ghanaian origin form about 90% of the selected sample and non Ghanaians coming from 13 different African countries make up about 10%. A total of 3,263 names were obtained from academic files of students from the registry. An excel formula was then used to extract the number of characters that makes up a name.

Table 1.0 Frequency Distribution of number of characters in a surname

observations	Frequency	Percent	Cumulative Percent
3	49	1.5	1.5
4	257	7.9	9.4
5	762	23.4	32.7
6	838	25.7	58.4
7	556	17	75.5
8	314	9.6	85.1
9	137	4.2	89.3
10	102	3.1	92.4
11	84	2.6	95.0
12	72	2.2	97.2
13	41	1.3	98.4
14	29	0.9	99.3
15	15	0.5	99.8
16	5	0.2	99.9
19	2	0.1	100
Total	3263	100	

Characters such as hyphens, apostrophes and spaces found in names were omitted and not counted. Example: Esono-Obeng Mokey is considered to have 15 characters as a single surname instead of 17. Healy (1968) adopted such a convention in a paper on the distribution of surnames in the United Kingdom. The summary of the number of characters forming a surname is displayed in the frequency table above. A cursory view of the valid percent and the number of characters making up a surname depicts a positively skewed distribution.

Fitting a Probability Distribution to the data

It is a normal practice in statistical methodology to use the graphical plots of observations to judge whether a gathered data came from any specific distribution if of course we do not know the distribution the data is coming from. This may be done for its own sake, or as a preliminary to applying a statistical method which assumes a certain underlying distribution for the observations. As can be observed from the plot in figure 1.0, a number of distributions nearly fit the data. Few among those distributions are the Lognormal distribution, Gamma distribution, Log Gamma distribution, the Dagum distribution and the Burr distribution.

The study compares which of these distributions fits very well to the data under consideration by estimating their respective parameters (mean, variance, etc) that gives the highest probability of producing the observed data. Appendix I shows the definitions of the probability density function (p.d.f), mean and variance for each of the distributions mentioned above.

Goodness of fit Tests

Beside fitting a number of distributions to data and judging from the visual patterns thereof, it is also important to consider whether the properties of the fitted distribution (example, the range and skewness) are appropriate. Above all, three very popular goodness of fit tests can be conducted to select the distribution that best fits the data. These are the Chi-square statistic, Kolmogorov – Smirnoff and Anderson – Darling statistics.

The Chi-square statistic measures how well the expected frequency of the fitted distribution compares with the observed frequency of a histogram of the observed data. The conclusion is drawn in such a way that if the computed test statistic is large, then the observed and expected values are not close and the distribution is a poor fit to the data. The Chi-square statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{N} \frac{\{O(i) - E(i)\}^2}{E(i)}$$

Under the null hypothesis H_0, that the data are governed by the assumed distribution, against the alternative that the data are not drawn from the assumed distribution.

Where $O(i)$ is the observed frequency of the i[th] histogram class or bar and $E(i)$ is the expected frequency from the fitted distribution of x-values falling within the x-range of the i[th] histogram bar. $E(i)$ is calculated as $E(i) = \{F(i_{\max}) - F(i_{\min})\} \cdot n$, where $F(x)$ = distribution function of the fitted distribution. $(i_{\max})$ = the x-value upper bound of the i[th] histogram bar and $(i_{\min})$ = the x-value lower bound of the i[th] histogram bar (Vose (2010)).

The Anderson-Darling (A-D) statistic is a more useful measure of fit than the Kolmogorov – Smirnoff (K-S) statistic. It addresses the lack of fit problem the K-S statistic falls short of. The K-S statistic D_n is defined as:

$$D_n = \max\left[|F_n(x) - F(x)|\right]$$

Where D_n is known as the K-S distance, n =total number of data points, $F(x)$ = distribution function of the fitted distribution, $F_n(x) = i/n$ and i = the cumulative rank of the data point. The A-D statistic is defined as:

$$A_n^2 = \int_{-\infty}^{\infty} |F_n(x) - F(X)|^2 \, \Psi(x) f(x) dx \, , \; \Psi(x) = \frac{n}{F(x)\{1-F(x)\}}$$

Where n = total number of data points, $F(x)$ = distribution function of the fitted distribution, $f(x)$ = density function of the fitted distribution, $F_n(x) = i/n$ and i = the cumulative rank of the data point (Vose (2010)).

The K-S statistic and the A-D statistic were used to test the null hypothesis that the data sampled came from the i^{th} distributions under consideration. H_o is accepted if the p-value of their respective test statistic is greater than 0.05 (the significance level in this study). The values of the most of the statistics obtained were estimated from the easy fit statistical package.

Results and Discussions

The average length of a surname in the community was found to be approximately seven characters with a standard deviation of 2.2206. The result showed that the length of surnames is positively skewed. The skewness is substantial (1.4433) meaning that the distribution of length of surnames is far from being symmetrical. Also the shape of the distribution of the length of surnames is quite different from the normal distribution. The distribution of this data has a high peak compared to the normal distribution hence its kurtosis value of 2.5628.

Table 2.0 Descriptive properties of the data and the fitted distributions

	Properties	Mean	Standard deviation	Coefficient of variation	Skewness	Kurtosis
*	Actual	6.6626	2.2206	0.3333	1.4433	2.5628
Fitted Distributions						
1	Lognormal	6.6479	2.0625	0.31024	0.96059	1.6544
2	Gamma	6.6626	2.2206	0.3333	0.6666	0.66653
3	Loggamma	6.6585	2.1813	0.32759	1.4681	4.3468
4	Dagum	6.6691	2.3778	0.35654	3.0623	37.836
5	Burr	6.6943	2.5775	0.38503	4.5596	191.21

The shape and other statistics computed from the raw data of length of surnames enabled five other distributions to be fitted to it. With these statistics one will easily be convinced that the data of length of surnames may be coming from distributions that have similar form and shape as the actual. However all the five distributions fitted to the length of surnames data have their properties different besides having approximately the same mean and standard deviation as the actual. The Lognormal and the Gamma distributions failed the skewness test when compared to the actual skewness of the data. When fitted to the Gamma distribution its kurtosis is close to the Guasian distribution. The loggamma, Dagum and the Burr distributions, even though are seldom used, their estimated properties are so close to the actual observations. Except the Burr distribution which has much higher

peak than the actual. Table 2.0 shows the descriptive properties of the data and the fitted distributions.

Table 3.0 Result of the Model Parameters

Distributions	Mean (μ)	Std dev(σ)	Shape (α)	Scale (β)	Shape (k)
Lognormal	1.8483	0.30314	-	-	-
Gamma	-	-	9.0019	0.74013	-
Loggamma	-	-	37.165	0.04973	-
Dagum	-	-	4.9313	5.1149	2.1012
Burr	-	-	7.8346	5.455	0.54316

The study further showed that the Loggamma and the Dagum distributions nearly fit the data for the length of surnames with the following model parameters Loggamma (37.165, 0.04973) and Dagum (4.9313, 5.1149, 2.1012). Their probability density functions (p.d.f) when plotted against the length of a surname on a histogram of the data are shown in figure 1.0.

The study was unable to conclude that the length of surnames in the study community was coming from any of the identified distributions since the goodness of fit test rejected the null hypothesis in all the five fitted distributions.

Table 4.0 Results of Goodness of fit test

	Distributions	K-S Statistic	p-value	A-D Statistic	p-value	Chi-Sq Statistic	p-value
1	Lognormal	0.15016	0.000	65.026	0.000	276.81	0.000
2	Gamma	0.16204	0.000	84.471	0.000	681.64	0.000
3	Loggamma	0.13731	0.000	56.125	0.000	276.51	0.000
4	Dagum	0.12948	0.000	53.257	0.000	263.83	0.000
5	Bur	0.13263	0.000	53.743	0.000	260.89	0.000

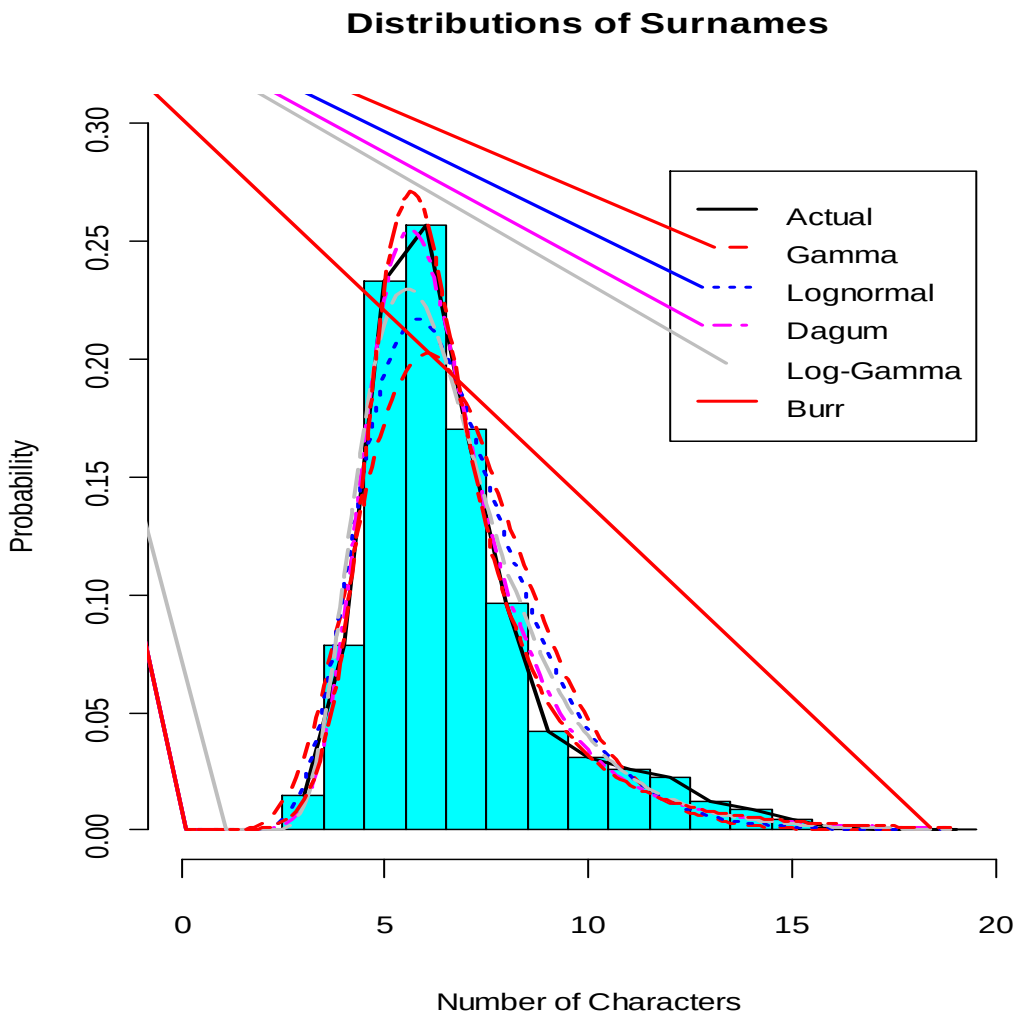

Figure 1.0 Histogram with fitted plots of the distributions

Conclusion and recommendation

The study is of the view that length of surnames on their own may not be coming from any particular family of statistical distribution but may nearly fit on some particular distributions such as the Loggamma and the Dagum distribution. Their descriptive properties are very close to the actual values of the length of surnames in the study area. Designers of data base systems must therefore be guided by these distributions in defining the number of characters for the surname field in their programming. This we believe may reduce the huge amount of space allocated for the surname field. More research must be done to identify the distribution that surnames in Ghana and the rest of Africa may be coming from.

References

1. Blanden J, Gregg P, Macmillan L (2005). Intergenerational mobility in Europe and North America: A report supported by the Sutton Trust. Centre for Economic Performance.
2. Clark G (2009). The indicated and the wealthy: Surnames reproductive success, genetic selection and social class in pre-industrial England, Mimeo.
3. Clark G (2010). Regression to mediocrity? Surnames and social mobility in England, 1200 - 2009. University of California.
4. Collardo D. M, Ortuño-Ortín (2011). Long-run intergenerational social mobility and distribution of surnames. Universidad de Alicante, Spain.
5. Dubrava D (2008). The importance of free Hard Drive space. http://www.helium.com/items/815514-the-importance-of-free-hard-drive-space. Accessed: 5/06/2013
6. Einav L, Yariv L (2004). What's in a surname? The effects of surname initials on Academic success. Mimeo, Stanford University.
7. Fox W. R, Lasker G. W (1983). The distribution of surname frequencies. Department of Computing Management Science, Mathematics and Statistics, City of London Polytechnic. London EC3N, 2EY, UK. International Statistical Review. Vol. 51, No. 1, pp. 81-87. http://www.jstor.org/stable/1402733. Accessed: 23/07/2013.
8. Healey, M. J. R (1968). The Lengths of surname. The journal of the royal Statistical Society series A (General), volume 131, part 4, pp. 567-568.
9. Holloway S. M, Soafer J. A (1992). Coefficients of Relationship by Isonomy among Registrations for five common cancers in Scottish males. Journal of Epidemiology and community health. Vol. 46, pp. 368-372.
10. Lafuerza L. F, Toral R (2011). Evolution of surname distribution under gender equality measures. University of Zaragoza, Spain
11. Larsen G. A (2007). Disk space usage and SQL server performance. http://www.databasejournal.com/features/mssql/article.php/3718066/Disk-Space-Usage-And-SQL-Server-Performance.htm Accessed: 10/07/2013
12. Teal A (2008). The importance of free Hard Drive space. http://www.helium.com/items/811920-the-importance-of-free-hard-drive-space, Accessed: 3/04/2013.
13. Vose D (2010). Fitting distributions to data. http://www.vosesoftware.com/vosesoftware/whitepapers/Fitting%20distributions%20to%20data.pdf, Accessed: 8/04/2013

Appendix I

List of definitions of Probability Distributions and their moments

1. The continuous random variable X has a Lognormal distribution if the random variable $Y = In(X)$ has a normal distribution with mean μ and standard deviation σ. Its probability density function is:

$$f(x;\mu,\sigma)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma x}e^{\frac{-1}{2\sigma^2}[In(x)-\mu]^2}, & x\geq 0\\ 0 & x<0.\end{cases}$$

The mean and variance of the log normal distribution are:

$$\mu=e^{\mu+\frac{\sigma^2}{2}} \quad \text{and} \quad \sigma^2=e^{2\mu+\sigma^2}(e^{\sigma 2}-1)$$

2. The continuous random variable X has a Gamma distribution with parameters α and β. Its probability density function is:

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{\frac{-x}{\beta}}, & x>0\\ 0 & ,elsewhere.\end{cases}$$

The mean and variance of the gamma distribution are:

$$\mu=\alpha\beta \quad \text{and} \quad \sigma^2=\alpha\beta^2$$

3. The continuous random variable X has a Log gamma distribution if $Y=In(X)$ has a gamma distribution with parameters α and β. Its probability density function is:

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{(In(x))^{\alpha-1}}{x\beta^{\alpha}\Gamma(\alpha)}e^{\left(\frac{-In(x)}{\beta}\right)}, & x>0\\ 0 & ,elsewhere.\end{cases}$$

The mean and variance of the log gamma distribution are:

4. The probability density function of the Dagum distribution is given by;

$$f(x;\alpha,\beta,p)=\frac{\alpha\beta}{x}\left(\frac{\left(\frac{x}{\beta}\right)^{\alpha p}}{\left(\left(\frac{x}{\beta}\right)^{\alpha}+1\right)^{p+1}}\right), \quad \text{for} \quad x>0, \text{ where } \alpha,\beta,p>0$$

The mean and variance of the Dagum distribution are:

$$\mu=\begin{cases}\dfrac{-\beta}{\alpha}\dfrac{\Gamma\left(\frac{-1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}+p\right)}{\Gamma(p)}, & \text{if } \alpha>1\\ \text{indeterminate}, & \text{otherwise}\end{cases}$$

$$\sigma^2=\begin{cases}\dfrac{-\beta^2}{\alpha^2}\left(2\alpha\dfrac{\Gamma\left(\frac{-2}{\alpha}\right)\Gamma\left(\frac{2}{\alpha}+p\right)}{\Gamma(p)}+\left(\dfrac{\Gamma\left(\frac{-1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}+p\right)}{\Gamma(p)}\right)^2\right), & \alpha>2\\ \text{indeterminate}, & otherwise\end{cases}$$

5. The probability function of the Burr distribution is given by:

$$f(x;c,k) = ck\frac{x^{c-1}}{\left(1+x^{c}\right)^{k+1}}, \quad c>0, k>0, x>0.$$

The mean is given as:

$$\mu = kB\left(\frac{k-1}{c}, 1+\frac{1}{c}\right), \quad \text{where } B(\ ) \text{ is the beta function.}$$

Appendix II

R program of the distributions under consideration

```
require(stats)
require(MASS)
require(VGAM)


surn<-read.delim(file="E:\\Backup\\Data\\surname.txt",head=TRUE);



x_a<-3:19
x<-seq(0,19,0.1)


f<-unname(table(surn$surname))/3263


y1<-c(f[1:14],0,0,f[15])
y2<-dgamma(x, 10.54589699, 1.58285658)
y3<-dlnorm(x, 1.848346477, 0.303144580)
y4<-ddagum(x, 4.9313, 5.1149, 2.1012)
y5<-(((log(x, base = exp(1)))^36.165)*exp(-log(x, base =
exp(1))/0.04973))/(x*(0.04973^37.165)*gamma(37.165))
y6<-(7.8346*0.54316*(x/5.455)^6.8346)/(5.455*(1+(x/5.455)^7.8346)^1.54316)


hist(surn$surname, col="cyan", breaks = 1.5:19.5, prob = T, xlim = c(0,20), ylim = c(0,0.3), main =
"Distributions of Surnames", xlab = "Number of Characters", ylab = "Probability")
lines(x_a,y1,type="l",lty=1,col=1,lwd = 2)
lines(x,y2,type="l",lty=2,col=2,lwd = 2)
lines(x,y3,type="l",lty=3,col=4,lwd = 2)
lines(x,y4,type="l",lty=4,col=6,lwd = 2)
lines(x,y5,type="l",lty=5,col=8,lwd = 2)
lines(x,y6,type="l",lty=6,col=10,lwd = 2)
legend (12,0.28, c("Actual","Gamma", "Lognormal","Dagum","Log-Gamma","Burr"),
lty=c(1,2,3,4,5),col=c(1,2,4,6,8,10),lwd = c(2,2,2,2,2,2))
```

Jackson, O. A. Y., Boyetey, D. B., Bemile, R. K., & Acheampong, E. (2013). Fitting a Distribution for Length of Last Names in a Mixed African Community in Ghana. Open Science Repository Mathematics, Online(open-access), e23050440. doi:10.7392/openaccess.23050440