Anthroponymical Comparative Variability

Josep M. Albaigès

DOI: 10.2436/15.8040.01.28

Abstract

It is very often necessary to establish a comparison between two populations relating their anthroponymical variability. In this case we do not have mathematical and statistical tools to make this comparison from objective points of view. It is not enough to know the number of persons in the population and the number of different names to define their degrees of variability. It is necessary to establish a homogeneous system to make this comparison using new accurate parameters. This article will try to define them and create an objective basis for defining the degree of variability of two or more populations. To do this it will be very useful to improve tools and statistics. As a final result we will define an objective index of variability through a "standard population" to which the object of study will be referred.

*****

1. Problem Statement

A given population of P individuals may for onomastical purposes be divided into subgroups (e.g. male and female population, native and immigrant population, population by age group, etc.).

Let us establish, for each subgroup, the corresponding list of names taken by their components, and let us classify them in descending order of frequency, every one characterized by its ordinal number *x* obtained with this criterion. We call F(*x*) the number of individuals for each name *x*, and FR(*x*) the same figure as a percentage of the total population subgroup. If *P* is the total population, then:

$$FR = f = \frac{F}{P}$$

Let us illustrate this with an example. These are the frequencies of appearance of the male names of the sailors who participated with Columbus in the discovery of America in 1492 (see details in Annex 1). We shall call them the "Columbian population".

x	NAME	F	FR = F/P
0		0	0
1	Juan	19	21.35%
2	Pedro	12	13.48%
3	Diego	7	7.87%
4	Rodrigo	7	7.87%
5	Alonso	4	4.49%
6	Bartolomé	4	4.49%
7	Cristóbal	4	4.49%
8	Francisco	4	4.49%
9	Martín	4	4.49%
10	Andrés	2	2.25%
11	Domingo	2	2.25%
12	Fernando	2	2.25%
13	García	2	2.25%
14	Sancho	2	2.25%
15	Álvaro	1	1.12%

16	Antón	1	1.12%
17	Antonio	1	1.12%
18	Bernal	1	1.12%
19	Chachu	1	1.12%
20	Gil	1	1.12%
21	Gómez	1	1.12%
22	Gonzalo	1	1.12%
23	Jácome	1	1.12%
24	Lope	1	1.12%
25	Luis	1	1.12%
26	Miguel	1	1.12%
27	Ruy	1	1.12%
28	Vicente	1	1.12%
	TOTAL	**89**	**100.00%**

Column *F* contains the total number of persons bearing the name of the column corresponding to *x*, and *FR* or *f* (to use the usual notation in statistics) the same percentage of the total figure, which we depict as a decimal fraction or a percentage. This value also coincides with the statistical probability that an individual in the population taken at random bears that name. For example, number 8 is the name Francisco and is borne by 4 persons, accounting for 4.49% of the population. So we write:

$$FR(8) = f(8) = 4.49\% = 0.0449$$

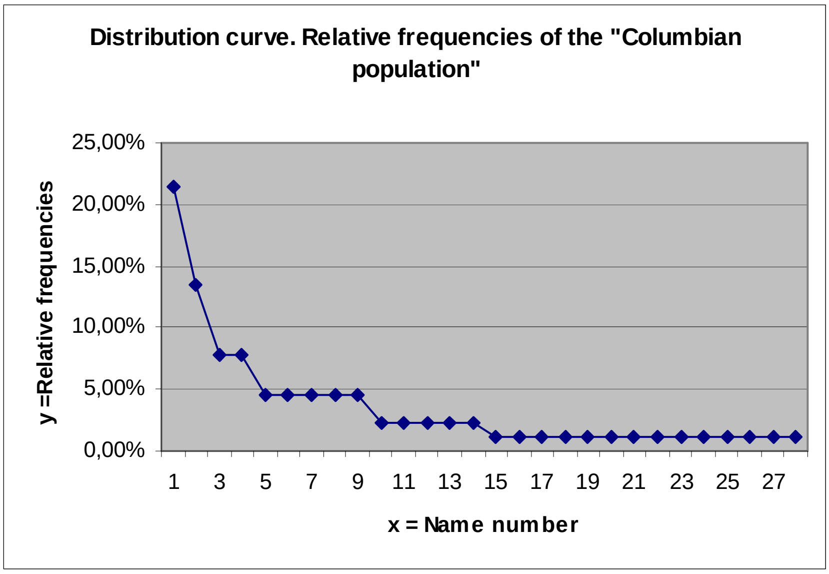
Similarly, the probability that an individual chosen at random has the name of Francisco is:

$$p(8) = 0.0449$$

There is a total population of 28 different Columbian names, borne by 89 individuals. *We shall call "v parameter" the number of individuals divided by the number of different names.*

This, in our case, is v = 89/28 = 3.18 individuals/name. That shows that each name is borne, on average, by 3.18 people. But the particular values differ greatly from that average.

We can better represent the above values of this population graphically in what we call its "distribution curve":



Any distribution of the names of a community is usually characterized by its irregularity, which we appreciate in the graph: there are a small number of “favorite or fashionable names” which alone accumulate an important percentage of the total population, and a mass of individuals with unusual names. In the example above, we can see a clear preference for the names number 1 (Juan) and 2 (Pedro), which, by the way, were the most common ones in Spain in the Middle Ages. The two together amount to more than one third of the total (34.83%), while less than half of the registered names (from 15 to 28, or from Alvaro to Vicente) only come to just over 15% of the population.

The giving of these names corresponds to certain onomastic habits of the population. *It is obvious that new names appear as the group increases, but it is also expected that the proportions remain approximately the same. In fact, statistical laws predict they would*.

The first and fundamental question to be asked is: Is there some kind of mathematical law mastering the distribution of names in a given population? And, if there is, to what extent does it resemble that of other populations?

No easy answer can be given to these questions. But it is possible from the analysis of many populations to discover regularities in them, as evidenced by analyzing the corresponding distribution curves.

Summarizing our findings, we concluded that any distribution curve of the population can be divided into three sections:

Section 1. Formed by a small number of names, 20 at most, for which there is no law. Not only can it not be found, but their rates vary in an irregular and fluctuating way. In general, it represents a significant percentage of about 10 to 25%.

Section 2. The middle of the distribution curve is usually adjusted to a potential law, in other words, that frequencies take the progression form called harmonic. This curve corresponds to the mathematical equation:

$$f(x) = ax^n$$

Section 3. At some point, say between 100 and 500 names, frequencies are higher than would be expected according to the formula above, and maintained in order of values from 1 to 2 bearers of each name. However, other laws can be found such as:

$$f(x) = \frac{1}{\alpha x + \beta x^2}$$

Let us examine each one applied to a large population: boys born in Barcelona in 2010.

2. Onomastic study of boys born in Barcelona (Spain) in 2010

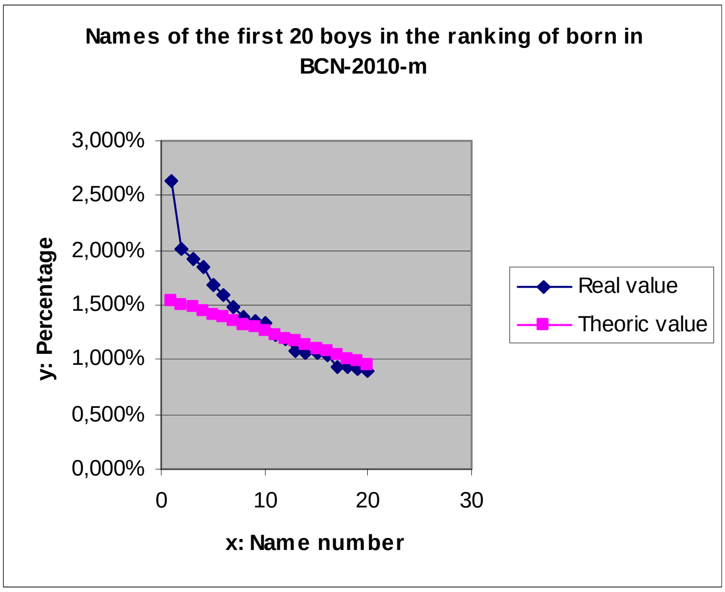
We shall call this population BCN-10-m. These are their main onomastic features:

- Individuals who make up the total population: 7384.
- Total different names: 2328.
- Every name is borne by: v = 7384/2328 = 3.17 children/name

Section 1.
We have reduced this first group conventionally to 20 names, including the "favorite names or fashion names". These account for 27.59% of the total, and the rates of representation of each name are:

N	MALE NAMES	F	F/N
1	MARC	194	2.627%
2	POL	148	2.004%
3	ÀLEX	142	1.923%
4	PAU	137	1.855%
5	MARTÍ	124	1.679%
6	ERIC	117	1.585%
7	ARNAU	110	1.490%
8	DANIEL	103	1.395%
9	GERARD	100	1.354%
10	JAN	98	1.327%
11	DAVID	90	1.219%
12	BIEL	88	1.192%
13	ALEJANDRO	80	1.083%
14	HUGO	79	1.070%
15	PABLO	78	1.056%
16	VÍCTOR	77	1.043%
17	ADRIÀ	69	0.934%
18	NIL	69	0.934%
19	BRUNO	68	0.921%
20	GUILLEM	66	0.894%

The whole group added up to 2037 individuals (a total of 27.59%). Here, every name is borne by an average of 10.2 persons per name, i.e. a high index *v*. This is the group's distribution curve, obtained by the method of minimum squares adjustment (see Annex 2). It has emphasized the individual points that comprise:



Note the strong irregularity of the whole. When looking for any known mathematical law, the one that most closely fits has been drawn by the pink line, and its equation is:

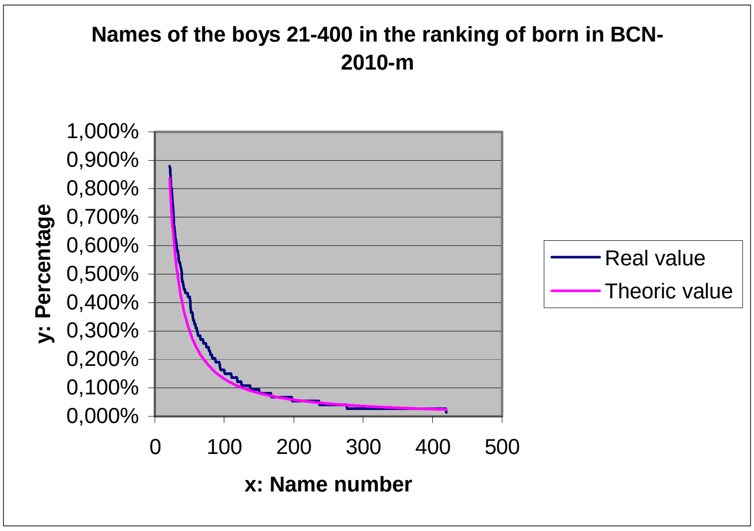
$$y = -0.00031x + 0.0157$$

Section 2. From x = 20, the graph shows some kind of regularity, and remains so until x = 420, approximately. This section fits nicely to a curve of potential type with a fractional exponent.

The approximate equation has been obtained through the minimum least-squares procedure (see Annex 2), and it results from the following law:

$$y = 0.309x^{-1.185}$$

The values fit quite well, as shown in the graph.



Section 3. From name 421 to the last one (2328), each individual name has only one bearer. From 5476, names are all single person ones. Note that this is the case for 1909 individuals, i.e. 25.8% of the children. It is to note that this big list of single person names is due largely to imperfect spelling, names incorrectly written, and even to capricious or aberrant forms.

In other words, for x > 5475, holds:

$$F(x) = 1$$

$$FR(x) = \frac{1}{7384} = 0.000135 = 0.0135\% = 1.35\,{}^{00}\!/_{00}$$

We have expressed here, as usual, the probability in percentages (%) or even in per-mil (‰).

Of course, w = 1, minimum possible value.

3. Comparison between the onomastic diversity of different populations.

It is an obvious fact that some populations are more diverse than others in personal names. Speaking about the Spanish one, until recently onomastic variety was very small. Among men, 25% were named José; among women, María was the name borne by… 50% of women! Such onomastic monotony almost inevitably forced a second name to complement the first one. Among men frequent complementary names were José Manuel, José María, José Ramón, and so on. Among women, María became so widespread that this name meant almost nothing, and the word *maría* was interpreted as a simple equivalent of "woman", which occurred even in words like *maritornes*, *marizápalos*, *maritates*, *marimacho* ('tomboy'), and so on. Indeed, it is said that Spain is "the land of Holy Mary".

"Onomastic diversity" is part of the onomastic habits of a population, but it is also a fashion and can vary with the time, as shown in Spain. The idea of "onomastic degree of diversity" of a population is evident, but it is difficult to express quantitatively and even to compare two populations to establish which of them has more richness in this field.

Some examples will help to explain this. Let us call P the number of individuals of a population and N the number of different names used at any one moment. Let us suppose these two populations are as follows:

- Population 1: $P1 = 2000$ persons, $N1 = 400$ names
- Population 2: $P2 = 500$ persons, $N2 = 100$ names

If we want to set a parameter defining the onomastic range of a population, we often proceed as we have seen before by simply finding the average number of individuals bearing each name for each community, which we call value *v*. In our case:

- Population 1: $v_1 = P1/N1 = 2000/400 = 5$ persons/name
- Population 2: $v_2 = P2/N2 = 400/100 = 4$ persons/name

It seems obvious that the greater the onomastic variation, the less a name is repeated, so that at first glance we might conclude from the above figures that population 2 is more varied than 1 in onomastic terms. We have found that many studies in this area have been made with this simple stating.

However, further analysis shows that this method may be incorrect. In fact, it could happen in this case that population 1 was more varied. Indeed, as population increases without changing onomastic habits, many of the existing names will be repeated, while new ones will appear, because most of the new members will adopt names already used. In other words, the number of persons per name increases, but it does so at lower rates than the increase in population. Therefore, the ratio *P/N* decreases. In the former case, it might occur that population 2 in reaching 2000 individuals (same as 1) only records 360 different names as a consequence of the impoverishment noted before, so the new value of *v* would be:

- $v'2 = 2000/360 = 5.56$ persons/name.

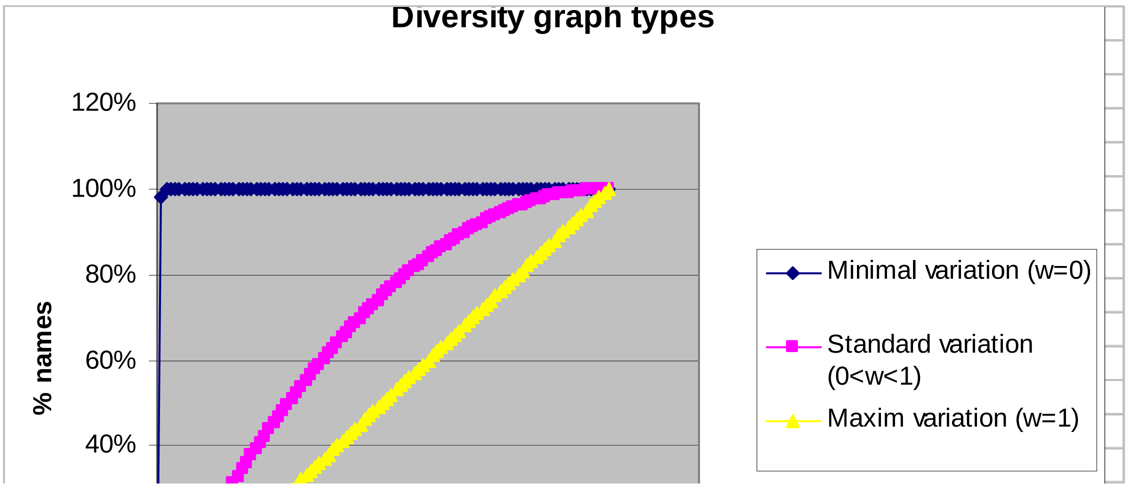
In other words, the new population 2 would be poorer than population 1 in onomastics as a result of increasing *without changing onomastic habits*. But it could also occur, for example, that in reaching a sample population of 2000, the number of different names could be $N' = 550$, which would lead us to an index $v''_2 = 2000/550 = 3.64$ persons/name.

This brings us to the core of the problem: *it is only possible to draw reliable conclusions about the degree of onomastic wealth of two populations when they have the same number of individuals.*

The "diversity" parameter

Now we will try to express the degree of onomastic diversity of a population through a number called "parameter *w*". We will set it so that the value *w* = 0 will represent the extreme case of minimal variability (all people bear the same name!), and the value *w* = 1 will apply to a population in which everybody has different names.

We will make the comparison by setting the abscissa graph of the population (as a percentage of the total) and ordering the names in accumulative relative frequency for each one. The graph may have any of these forms:



The dark blue colored line (parameter *w* = 0) indicates a null onomastic variation (all members of the population have the same name). Therefore the value of *F/N* rises suddenly to 100%, and stays there in a horizontal line with an equation *y* = 1.

The yellow line (*w* = 1) represents the opposite case: all persons have different names (or every name has the same number of bearers) and the rate of change is maximum. The curve is now a line that rises uniformly as the bisector of the first quadrant and its equation is *y* = *x*.

A real curve will stay in an intermediate position, such as the pink one. Here the index *w* has a value between 0 and 1. The higher this value (important onomastic diversity, *w* close to 1) the closer it will stick to the yellow line, the poorer the distribution (high onomastic uniformity, *w* small) the closer it will stick to the dark blue line. The proposed example of the pink line corresponds to a value about *w* = 2/3.

We can derive the area between the curve and the *x*-axis through integral calculus, integration by parts or simple computation.

If this value is *S*, the area above the curve will be 1 - *S*, and this area, relative to the total area of the blue-yellow triangle, will finally give the parameter *w*:

$$w = \frac{1 - S}{1/2} = 2(1 - S)$$

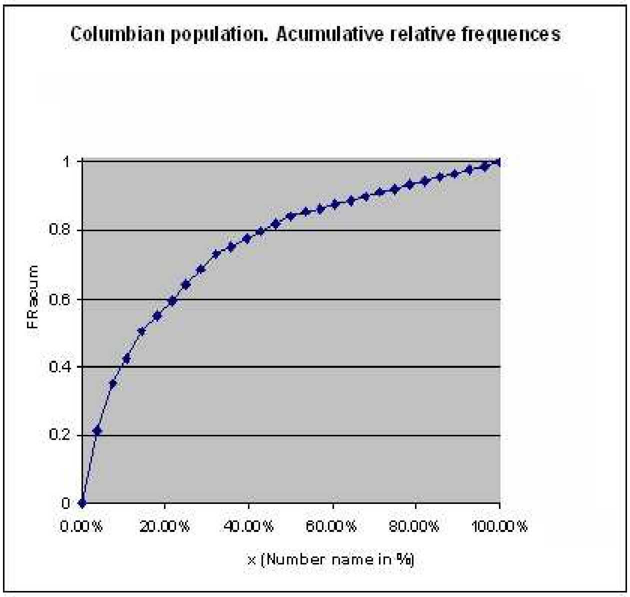
Comparison between onomastic diversities of different populations

We are now ready to compare the onomastic differences between two different populations. Starting with the "Columbian population", the new table will contain new required parameters:

COLUMBIAN POPULATION					
x	x % accum	NAME	F	F/N	F/N Accum
0	0.00%		0	0	0
1	3.57%	Juan	19	21.35%	21.35%
2	7.14%	Pedro	12	13.48%	34.83%
3	10.71%	Diego	7	7.87%	42.70%
4	14.29%	Rodrigo	7	7.87%	50.56%
5	17.86%	Alonso	4	4.49%	55.06%
6	21.43%	Bartolomé	4	4.49%	59.55%
7	25.00%	Cristóbal	4	4.49%	64.04%
8	28.57%	Francisco	4	4.49%	68.54%
9	32.14%	Martín	4	4.49%	73.03%
10	35.71%	Andrés	2	2.25%	75.28%
11	39.29%	Domingo	2	2.25%	77.53%
12	42.86%	Fernando	2	2.25%	79.78%
13	46.43%	García	2	2.25%	82.02%
14	50.00%	Sancho	2	2.25%	84.27%
15	53.57%	Álvaro	1	1.12%	85.39%
16	57.14%	Antón	1	1.12%	86.52%
17	60.71%	Antonio	1	1.12%	87.64%
18	64.29%	Bernal	1	1.12%	88.76%
19	67.86%	Chachu	1	1.12%	89.89%
20	71.43%	Gil	1	1.12%	91.01%
21	75.00%	Gómez	1	1.12%	92.13%
22	78.57%	Gonzalo	1	1.12%	93.26%
23	82.14%	Jácome	1	1.12%	94.38%
24	85.71%	Lope	1	1.12%	95.51%
25	89.29%	Luis	1	1.12%	96.63%
26	92.86%	Miguel	1	1.12%	97.75%
27	96.43%	Ruy	1	1.12%	98.88%
28	100.00%	Vicente	1	1.12%	100.00%
		TOTAL	**89**	**100.00%**	

The first column contains the values *x* (number of name). The second, the same values as a percentage, accumulated to origin. The last one, also added, contains the accumulative percentages of relative frequencies.

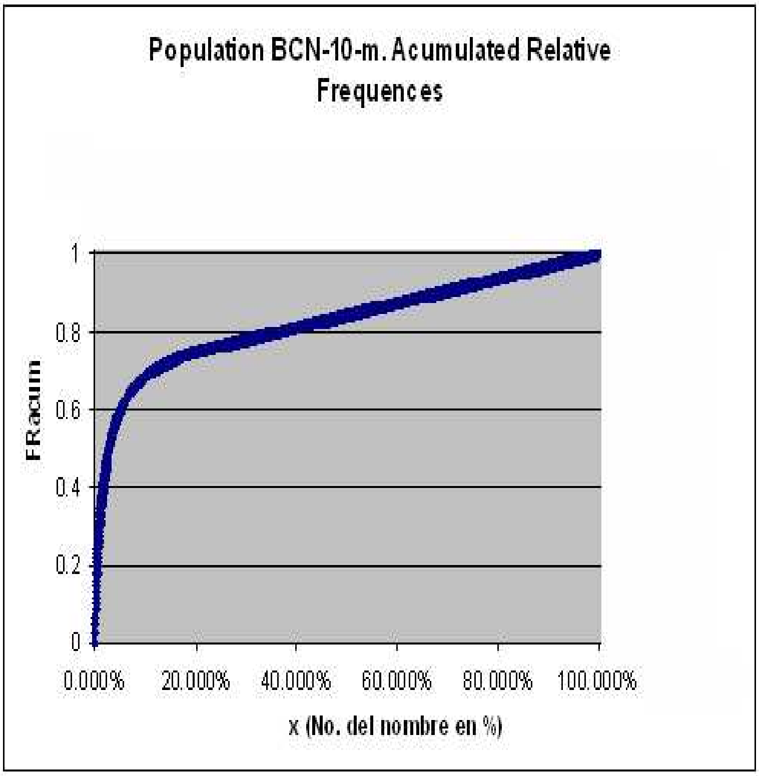
The following graph is built from the above values, taking as abscissa the second column and as ordinates the last one:



In the case of the Columbian distribution, we have calculated the area subtended between the curve and the horizontal axis (Annex 1). It is S = 0.752. Therefore, the parameter w equals:

$$w = 2(1-0.752) = 2 \cdot 0.248 = 0.496$$

In the case of the "Population BCN-10-m", we have built a similar table, which we do not include here due to its excessive size (it is shown in Annex 3). The curve is now:



The calculation of the area between the curve and the horizontal axis, done with integral calculus, gives the result S = 0.824. It gives:

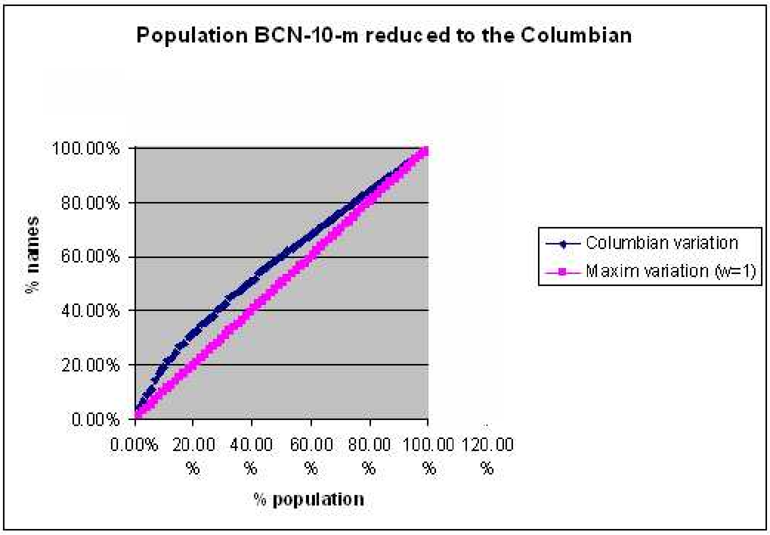
$$w = 2(1 - 0.824) = 2 \cdot 0.126 = 0.352$$

This value is lower than the Columbian distribution, and this could indicate a poorer onomastic wealth. However, to be sure it will be necessary to carry out a further calculation.

Testing by the Monte-Carlo method

To do this, we will make use of 89 randomly selected persons (equal number of individuals to the Columbian population) among the "Population 2328 BCN-10-m". We can do this with a similar procedure to that shown in Annex 4.

After the calculations (Annex 4) we have the following graph:



We now have $S = 0.585$.

Therefore, $w = 2\ (1\text{-}.585) = .830$

So we see that the value obtained is much higher than the "Columbian population", which, as we have seen, had a $w = 0.352$. *The population "BCN-10-m" has a much greater variation than the "Columbian".*

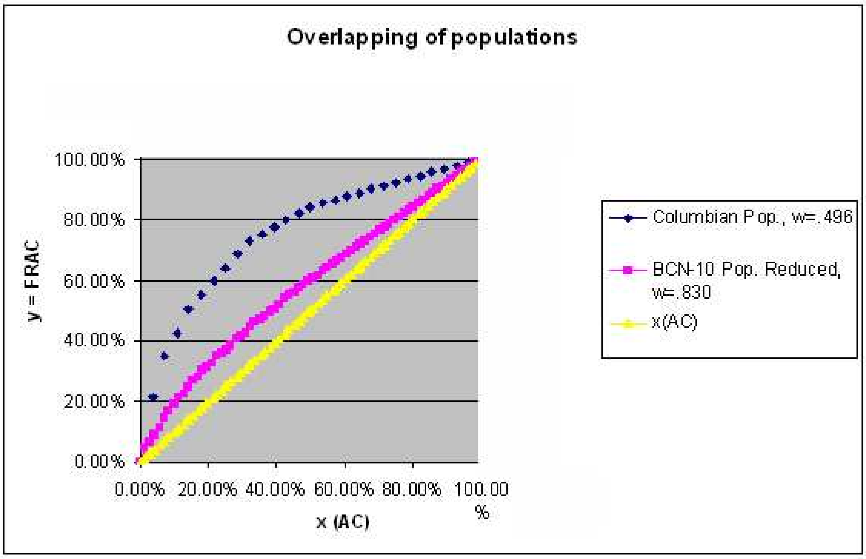
4. Conclusions

1. The best curves to represent names distribution are potential curves, of the type $y = ax^n$, with exponents somewhat higher than 1.

2. Given two different communities, to establish their onomastic wealth we need to reduce them to the same number of individuals.

This could be done by increasing the larger community through systems that we will here leave for another study, but the simplest way is to take a number of individuals at random from the larger community as numerous as in the smaller one, noting their names. This new "small community" can be compared to the other by the usual parameters v (individuals/name) or the "variation index" w (between 0 and 1). The variability of the population will be greater as this index grows.

In the case we have studied, the respective curves obtained graphically express the difference in variability between the Columbian and BCN-10-m populations:



Josep M. Albaigès i Olivart († 2014)
Societat d'Onomàstica
Barcelona

Annex 1

Names of members of the first trip to America.

In the article *The Great Voyage of Columbus* (*History* magazine, no. 198, October 1992) the list of the sailors who accompanied Columbus on his first voyage compiled by historian Alice B. Gould was published. She wrote: "For those who do not understand how difficult it is for the historian to locate many people during this time, it should be known that the spelling was very anarchic, the same name often varying from one scribe to another, there were no generally accepted standards and sometimes an individual appeared referred to by his Christian name alone or the surname. Moreover, sometimes he was referred to by place of origin and even only by his nickname. With these four different ways, added to the spelling, it may be that that documents refer to the same person in different ways."

The list is as follows:

Alonso, grumete.
Alonso, Maestre, físico.
Alonso Clavija, criminal.
Alonso de Morales.
Alvaro, marinero
Andrés de Huelva, grumete.
Andrés de Yevenes, grumete.
Antón Catabres, marinero.
Amonio de Cuéllar.
Bartolomé Bives, marinero.
Bartolomé García, contramaestre.
Bartolomé Roldan, marinero.
Bartolomé de Torres, criminal.
Bernal, grumete,
Cristóbal Caro, platero, grumete.
Cristóbal Colón, cap. general
Cristóbal Quinteto, marinero, uno de los dueños de la *Pinta*).
Cristóbal García Sarmiento, piloto de la *Pinta*.
Chachu, contramaestre.
Diego, Maestre,
Diego de Arana, alguacil de la flota.
Diego Bermúdez.
Diego Leal, grumete.
Diego Lorenzo, alguacil.
Diego Pérez, pintor.
Diego Martín Pinzón.
Domingo, tonelero.
Domingo de Lequeitio.
Fernando Medel, grumete.
Fernando de Triana, grumete.
Francisco de Huelva.
Francisco Medel, grumete.
Francisco Niño.
Martín Pinzón, maestre de la *Pinta*.
Francisco García Vallejos, marinero.

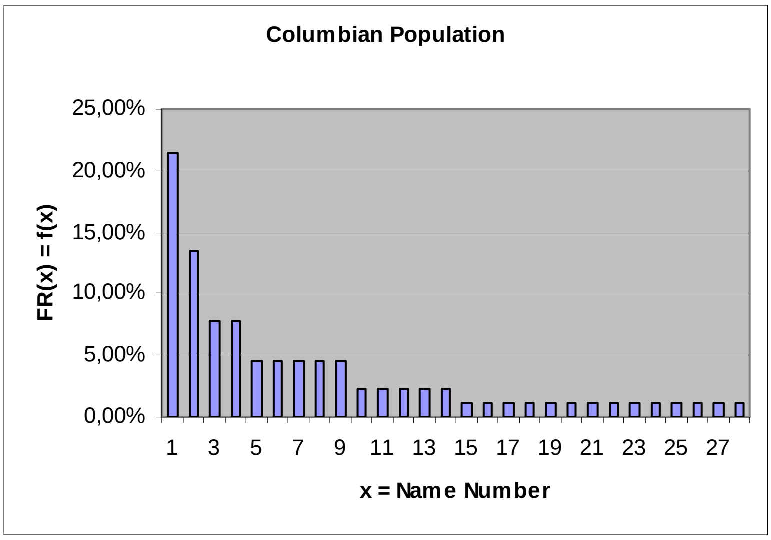
Juan Arráez, marinero.
Juan de la Cosa, maestre y dueño de la *Santa María*.
Juan Martínez de Azogue, marinero.
Juan de Medina, sastre, marinero.
Juan de Moquer, criminal, marinero.
Juan Niño, maestre de la *Niña*.
Juan de la Plaza, marinero.
Juan Quadrado, grumete.
Juan Quintero de Algruta, contramaestre de la *Pinta*.
Juan Reynal, marinero.
Juan Rodríguez Bermejo, marinero de la *Pinta*.
Rodrigo de Triana.
Juan Romero, marinero.
Juan Ruiz de la Peña, marinero.
Juan Verde de Triana, marinero,
Juan Vezano, marinero.
Juan de Xeres, marinero.
Lope, calafate.
Luis de Torres.
Martín Alonso Pinzón, capitán de la *Pinta*.
Martin de Urtubia.
Miguel de Sofía, grumete.
Pedro de Arcos, de la Pinta.
Pedro Arráez, marinero.
Pedro Gutiérrez, repostero del Rey.
Pedro de Lepe.
Pedro Alonso Niño, piloto.
Pedro de Salcedo, paje de Colón.
Pedro de Soria.
Pedro Tegero, grumete.
Pedro de Terreros, maestresala de Colón.
Pedro de Villa, marinero.
Pedro Yzquierdo, criminal.
Pedro Sánchez de Montilla, marinero.
Rodrigo de Escobedo, escribano de la

García Alonso.
García Hernández, marinero, despensero de la *Pinta*.
Gil Pérez, marinero.
Gonzalo Franco.
Gómez Rascón, marinero, uno de los dueños de la *Pinta*.
Jácome el Rico, genovés.
Juan, grumete.
Juan, Maestre, cirujano.
Juan Arias, grumete.

armada.
Rodrigo Gallego, grumete.
Rodrigo de Triana = Juan Rodríguez Bermejo.
Rodrigo Monge.
Rodrigo de Xerez.
Rodrigo Sánchez de Segovia, veedor en la *Santa María*.
Rui García, marinero.
Sancho de Rama, marinero.
Sancho Ruiz de Gama, piloto.
Vicente Yáñez Pinzón, capitán de la Niña.

We will use this list for creating statistics of current names at that time. This is the list, with the names ordered in descending order of their frequencies of appearance:

NAME	F	F/N
	0	0
Juan	19	21.35%
Pedro	12	13.48%
Diego	7	7.87%
Rodrigo	7	7.87%
Alonso	4	4.49%
Bartolomé	4	4.49%
Cristóbal	4	4.49%
Francisco	4	4.49%
Martín	4	4.49%
Andrés	2	2.25%
Domingo	2	2.25%
Fernando	2	2.25%
García	2	2.25%
Sancho	2	2.25%
Álvaro	1	1.12%
Antón	1	1.12%
Antonio	1	1.12%
Bernal	1	1.12%
Chachu	1	1.12%
Gil	1	1.12%
Gómez	1	1.12%
Gonzalo	1	1.12%
Jácome	1	1.12%
Lope	1	1.12%
Luis	1	1.12%
Miguel	1	1.12%
Ruy	1	1.12%
Vicente	1	1.12%
TOTAL	**89**	**100.00%**



Main consequences:

- As usual in the Middle Ages, Pedro ranks in one of the first places. Juan is the first one, representing 21% of the crew, which puts this name in a situation similar to José until recently.
- The 4 first names (Juan, Pedro, Diego and Rodrigo) monopolize 50% of the total. That is, monotony in the giving of names was worse than today.
- 89 people bear 28 different names. New sample of monotony.

If we express the same histogram with a curve, it can be used to calculate the area subtended between it and the horizontal axis, using the approximate integration formula:

$$S = \left[\sum_{1}^{n} y_i - \frac{y_1 + y_n}{2} \right] s$$

Where *S* is the area, y_i the ordinates and *s* the respective amplitude of each interval.

Annex 2

A-2.1. Calculation of the regression line

Let us assume that two correlated variables (*x*,*y*) are linked by a relationship such as:

$$y = px + q$$

We have *n* data pairs $\{x_i, y_i\}$.

The best way to find the most suitable coefficients (*p*,*q*) is the least squares adjustment. That is, if we assume that for each point the difference between actual and predicted values is given by the formula:

$$\varepsilon_i = y_i - px_i - q$$

The sum of squares over all the points of the set $\{x_i, y_i\}$ will be:

$$S = \sum_1^n \varepsilon_i^2 = \sum_1^n [y_i - px_i - q] = \sum y_i^2 + p^2 \sum x_i^2 + \sum q^2 - 2p\sum x_i y_i - 2q\sum y_i + 2pq\sum x_i$$

Here we have suppressed, for convenience, sub-and superscripts of summation. Appropriate values of *p* and *q* should be the ones to make this amount minimum. To find them, we derive the expression above:

$$\frac{\partial S}{\partial p} = 2p\sum x_i^2 - 2\sum x_i y_i + 2q\sum x_i$$
$$\frac{\partial S}{\partial q} = 2\sum q - 2\sum y_i + 2p\sum x_i$$

These expressions will be equalized to zero, leading to a system of two linear equations:

$$p\sum x_i^2 + q\sum x_i = \sum x_i y_i$$
$$p\sum x_i + nq = \sum y_i$$

They can be easily solved by the Cramer method:

$$p = \frac{\begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - [\sum x_i]^2}$$

$$q = \frac{\begin{vmatrix} \sum x_i^2 & \sum x_i y_i \\ \sum x_i & \sum y_i \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left[\sum x_i\right]^2}$$

These values, despite their complex appearance, can easily be obtained by computer using an Excel sheet or BASIC program.

For example, you may find the following useful:

PROGRAM REGNOMS1.BAS

```
10 REM ***** Programa per al càlcul de les correlacions de primer grau en “Variacions de noms en una comunitat”
20 REM ***** Josep M. Albaigès i Olivart, albaiges@ciccp.es
30 REM ***** © 09.09.11
40 REM ***** L’autor agrairà qualsevol comentari, suggerència o millora en el programa
50 INPUT “N”;N
60 OPEN “\x.txt” FOR INPUT AS 1
70 OPEN “\y.txt” FOR INPUT AS 2
80 SX=0:SX2=0:SY=0:SXY=0
90 INPUT#1,X
100 INPUT#2,Y
110 SX=SX+X:SX2=SX2+X^2:SY=SY+Y:SXY=SXY+X*Y
120 IF NOT EOF(1) THEN GOTO 90
130 CLOSE
140 PRINT SX,SX2,SY,SXY
150 DELTA=N*SX2-SX^2
160 PRINT “delta”
170 PRINT DELTA
180 P=(N*SXY-SX*SY)/DELTA
190 Q=(SX2*SY-SX*SXY)/DELTA
200 PRINT “p”,”q”
210 PRINT P,Q
```

A-2.2. Regression curve of second degree

Let us now suppose that two correlated variables (x,y) are linked by a relationship such as:

$$y = \alpha x + \beta x^2$$

We have, as before, n pairs of data $\{x_{\mathrm{i}},y_{\mathrm{i}}\}$.

We will make use as before of the method of minimal squares. That is, if we assume that for each point the difference between actual and predicted values is given by the formula:

$$\varepsilon_i = y_i - \alpha x_i - \beta x_i^2$$

The sum of squares over all the points of the set $\{x_{\mathrm{i}},y_{\mathrm{i}}\}$ will now be:

$$S=\sum_{1}^{n}\varepsilon_i^2=\sum_{1}^{n}\left[y_i-\alpha x_i-\beta x_i^2\right]=\sum y_i^2+\alpha^2\sum x_i^2+\beta^2\sum x_i^4-2\alpha\sum x_i y_i-2\beta\sum x_i^2 y_i+2\alpha\beta\sum x_i^3$$

We have also suppressed here, for convenience, sub-and superscripts of summation. As before, we derive first the sum S:

$$\frac{\partial S}{\partial\alpha}=2\alpha\sum x_i^2-2\sum x_i y_i+2\beta\sum x_i^3$$

$$\frac{\partial S}{\partial\beta}=2\beta\sum x_i^4-2\sum x_i^2 y_i+2\alpha\sum x_i^3$$

These expressions equal to zero, leading to a system of two linear equations:

$$\alpha\sum x_i^2+\beta\sum x_i^3=\sum x_i y_i$$

$$\alpha\sum x_i^3+\beta\sum x_i^4=\sum x_i^2 y_i$$

We will solve them through the Cramer method:

$$\alpha=\frac{\begin{vmatrix}\sum x_i y_i & \sum x_i^3\\ \sum x_i^2 y_i & \sum x_i^4\end{vmatrix}}{\begin{vmatrix}\sum x_i^2 & \sum x_i^3\\ \sum x_i^3 & \sum x_i^4\end{vmatrix}}=\frac{\sum x_i y_i\sum x^4-\sum x_i^2 y_i\sum x_i^3}{\sum x_i^2\sum x_i^4-\left(\sum x_i^3\right)^2}$$

$$\beta=\frac{\begin{vmatrix}\sum x_i^2 & \sum x_i^3\\ \sum x_i^2 y_i & \sum x_i^2 y_i\end{vmatrix}}{\begin{vmatrix}\sum x_i^2 & \sum x_i^3\\ \sum x_i^3 & \sum x_i^4\end{vmatrix}}=\frac{\sum x_i^2\sum x_i^2 y_i-\sum x_i^3\sum x_i y_i}{\sum x_i^2\sum x_i^4-\left(\sum x_i^3\right)^2}$$

As before, these values can easily be obtained by computer using an Excel sheet or BASIC program.

For example, the following one:

PROGRAM REGNOMS2.BAS

```
10 REM ***** Programa per al càlcul de les correlacions de segon grau en “Variacions de noms en una comunitat”
20 REM ***** Josep M. Albaigès i Olivart, albaiges@ciccp.es
30 REM ***** © 09.09.11
40 REM ***** L’autor agrairà qualsevol comentari, suggerència o millora en el programa
50 INPUT “N”;N
60 OPEN “\x.txt” FOR INPUT AS 1
70 OPEN “\y.txt” FOR INPUT AS 2
80 SX=0:SX3=0:SX4=0:SX2Y=0
90 INPUT#1,X
100 INPUT#2,Y
110 SX=SX2+X^2:SX3=SX3+X^3:SX4=SX4+X^4:SX2Y=SX2Y+X^2*Y
120 IF NOT EOF(1) THEN GOTO 80
130 CLOSE
140 DELTA=SX2*SX4-SX3^2
150 PRINT “delta”
160 PRINT DELTA
170 ALFA=(XSY*SX4-SX2Y*SX3)/DELTA
180 BETA=(SX2*SX2Y-SX3*SXY)/DELTA
190 PRINT “alfa”,”beta”
200 PRINT ALFA, BETA
```

A-2.3. Determination of the best potential curve.

The shape of the distribution curve of the intermediate zone suggests a potential curve, say, with an equation:

$$y = kx^n$$

For its determination it will be sufficient to take logarithms of both sides:

$$\log y = n \log x + \log k$$

That is, there will be a linear relationship between the logarithms of *x* and *y*. We are thus in the case seen in A-2.1. It is enough to take logarithms of the values to be correlated and proceed under the same program.

Annex 3

Population BCN-10-m (first 100 numbers)

N	Log(N)	MALE NAMES	F	F/N	F/N accum	log(F/N)
1	0.000	MARC	194	2.627%	2.627%	-1.580
2	0.301	POL	148	2.004%	4.632%	-1.698
3	0.477	ÀLEX	142	1.923%	6.555%	-1.716
4	0.602	PAU	137	1.855%	8.410%	-1.732
5	0.699	MARTí	124	1.679%	10.089%	-1.775
6	0.778	ERIC	117	1.585%	11.674%	-1.800
7	0.845	ARNAU	110	1.490%	13.164%	-1.827
8	0.903	DANIEL	103	1.395%	14.559%	-1.855
9	0.954	GERARD	100	1.354%	15.913%	-1.868
10	1.000	JAN	98	1.327%	17.240%	-1.877
11	1.041	DAVID	90	1.219%	18.459%	-1.914
12	1.079	BIEL	88	1.192%	19.651%	-1.924
13	1.114	ALEJANDRO	80	1.083%	20.734%	-1.965
14	1.146	HUGO	79	1.070%	21.804%	-1.971
15	1.176	PABLO	78	1.056%	22.860%	-1.976
16	1.204	VÍCTOR	77	1.043%	23.903%	-1.982
17	1.230	ADRIÀ	69	0.934%	24.837%	-2.029
18	1.255	NIL	69	0.934%	25.772%	-2.029
19	1.279	BRUNO	68	0.921%	26.693%	-2.036
20	1.301	GUILLEM	66	0.894%	27.587%	-2.049
21	1.322	LUCAS	65	0.880%	28.467%	-2.055
22	1.342	NICOLÁS	64	0.867%	29.334%	-2.062
23	1.362	MAX	60	0.813%	30.146%	-2.090
24	1.380	ORIOL	59	0.799%	30.945%	-2.097
25	1.398	IKER	57	0.772%	31.717%	-2.112
26	1.415	ALEIX	54	0.731%	32.449%	-2.136
27	1.431	JOAN	50	0.677%	33.126%	-2.169
28	1.447	JORDI	49	0.664%	33.789%	-2.178
29	1.462	JOEL	47	0.637%	34.426%	-2.196
30	1.477	ROGER	46	0.623%	35.049%	-2.206
31	1.491	ÁLVARO	45	0.609%	35.658%	-2.215
32	1.505	ADRIÁN	43	0.582%	36.241%	-2.235
33	1.519	GABRIEL	43	0.582%	36.823%	-2.235
34	1.531	MARIO	41	0.555%	37.378%	-2.256
35	1.544	IVAN	40	0.542%	37.920%	-2.266
36	1.556	SERGI	40	0.542%	38.462%	-2.266
37	1.568	UNAI	39	0.528%	38.990%	-2.277
38	1.58	ADAM	38	0.515%	39.504%	-2.289
39	1.591	HÉCTOR	35	0.474%	39.978%	-2.324
40	1.602	IZAN	35	0.474%	40.452%	-2.324
41	1.613	MATEO	34	0.460%	40.913%	-2.337
42	1.623	OSCAR	33	0.447%	41.360%	-2.350
43	1.633	XAVIER	33	0.447%	41.807%	-2.350
44	1.643	AITOR	32	0.433%	42.240%	-2.363
45	1.653	BERNAT	32	0.433%	42.673%	-2.363
46	1.663	DIEGO	32	0.433%	43.107%	-2.363
47	1.672	LEO	32	0.433%	43.540%	-2.363
48	1.681	ALBERT	31	0.420%	43.960%	-2.377
49	1.690	LUCA	31	0.420%	44.380%	-2.377
50	1.699	TEO	31	0.420%	44.800%	-2.377

51	1.708	IAN	29	0.393%	45.192%	-2.406
52	1.716	CARLOS	27	0.366%	45.558%	-2.437
53	1.724	MIQUEL	27	0.366%	45.924%	-2.437
54	1.732	TOMÀS	27	0.366%	46.289%	-2.437
55	1.740	MARCOS	25	0.339%	46.628%	-2.470
56	1.748	SANTIAGO	25	0.339%	46.966%	-2.470
57	1.756	FERRAN	24	0.325%	47.291%	-2.488
58	1.763	GUILLERMO	24	0.325%	47.616%	-2.488
59	1.771	GAEL	23	0.311%	47.928%	-2.507
60	1.778	MARTÍN	23	0.311%	48.239%	-2.507
61	1.785	ERIK	22	0.298%	48.537%	-2.526
62	1.792	IGNACIO	21	0.284%	48.822%	-2.546
63	1.799	MARCEL	21	0.284%	49.106%	-2.546
64	1.806	MIGUEL	21	0.284%	49.391%	-2.546
65	1.813	SAMUEL	21	0.284%	49.675%	-2.546
66	1.820	JAVIER	20	0.271%	49.946%	-2.567
67	1.826	JOSEP	20	0.271%	50.217%	-2.567
68	1.833	LLUC	20	0.271%	50.488%	-2.567
69	1.839	SERGIO	20	0.271%	50.758%	-2.567
70	1.845	AARÓN	19	0.257%	51.016%	-2.590
71	1.851	ELOI	19	0.257%	51.273%	-2.590
72	1.857	LUIS	19	0.257%	51.530%	-2.590
73	1.863	MARCO	19	0.257%	51.788%	-2.590
74	1.869	ARAN	18	0.244%	52.031%	-2.613
75	1.875	ISAAC	18	0.244%	52.275%	-2.613
76	1.881	MATÍAS	18	0.244%	52.519%	-2.613
77	1.886	RUBÉN	18	0.244%	52.763%	-2.613
78	1.892	ÀNGEL	17	0.230%	52.993%	-2.638
79	1.898	RAYAN	17	0.230%	53.223%	-2.638
80	1.903	DÍDAC	16	0.217%	53.440%	-2.664
81	1.908	ENRIQUE	16	0.217%	53.657%	-2.664
82	1.914	JUAN	16	0.217%	53.873%	-2.664
83	1.919	AXEL	15	0.203%	54.076%	-2.692
84	1.924	MAURO	15	0.203%	54.280%	-2.692
85	1.929	OMAR	15	0.203%	54.483%	-2.692
86	1.934	PEDRO	15	0.203%	54.686%	-2.692
87	1.940	ROC	15	0.203%	54.889%	-2.692
88	1.944	ASIER	14	0.190%	55.079%	-2.722
89	1.949	EDUARD	14	0.190%	55.268%	-2.722
90	1.954	IGNASI	14	0.190%	55.458%	-2.722
91	1.959	ISMAEL	14	0.190%	55.647%	-2.722
92	1.964	MANUEL	14	0.190%	55.837%	-2.722
93	1.968	RAÜL	14	0.190%	56.027%	-2.722
94	1.973	EDGAR	13	0.176%	56.203%	-2.754
95	1.978	ALAN	12	0.163%	56.365%	-2.789
96	1.982	ALEXANDER	12	0.163%	56.528%	-2.789
97	1.987	JOSÉ	12	0.163%	56.690%	-2.789
98	1.991	KEVIN	12	0.163%	56.853%	-2.789
99	1.996	MOHAMED	12	0.163%	57.015%	-2.789
100	2.000	QUIM	12	0.163%	57.178%	-2.789

Please note that the columns of names and percentages were added to the logarithms of the order number and the same percentages which are used in the calculation of the potential regression curve.

Annex 4

The Monte-Carlo method is based on taking (usually by computer) random elements of a set, each one provided with a certain probability.

To apply it, proceed as follows:

1. They have the probability of occurrence (i.e. relative frequencies) of the "Population BCN-10-m" and the corresponding accumulative probabilities.
2. They will be assigned to a table of probabilities; it will be called *p*(*x*).
3. Then we will obtain a random number between 0 and 1.
4. Multiplying this number by the total population, you get a table number that will be considered as an extracted element.
5. Proceeding in this way 89 times, you have a set of 89 items, each corresponding to a name. This set of values will be regarded as the "Reduced Columbian Population".

You can use a program like this:

PROGRAM FRAC.BAS

It is necessary to put the relative population frequencies in the sequential file "\FRAC.TXT"

```
10 REM ***** Programa per a la determinació de població derivada pel mètode de
Monte-Carlo
20 REM ***** Josep M. Albaigès i Olivart, albaiges@ciccp.es
30 REM ***** © 09.09.11
40 REM ***** L'autor agrairà qualsevol comentari, suggerència o millora en el programa
50 OPEN "FRAC.TXT" FOR INPUT AS 1:OPEN "\FRAC2.TXT" FOR OUTPUT AS
2:OPEN "\FRAC3.TXT" FOR OUTPUT AS 3
60 FOR I=1 TO 2328
70 INPUT#1,X
80 F(I)=X
90 PRINT I,F(I)
100 NEXT
110 F(I)=194
120 S=0
130 FOR I=1 TO 2328
140 S=S+F(I)
150 NEXT
160 PRINT S
170 FOR I01 TO 2328
180 FRAC(I)0FRAC(I-1)+F(I)/S
190 NEXT
200 FOR I=1 TO 2328
210 PRINT I,FRAC(I)
220 NEXT
230 FOR J=1 TO 89
240 R=RND
250 I=1
260 IF R>FRAC(I) THEN I=I+1:GOTO 260
270 FRACUM(I-1)=FRACUM(I)+1
```

```
280 NEXT
290 FOR K=1 TO 2328
300 IF FRACUM(K)>0 THEN PRINT K,FRACUM(K)
310 NEXT
320 FOR K=1 TO 2328
330 IF FRACUM(K)>0 THEN PRINT#2,K:PRINT#3,FRACUM(K)
340 NEXT
350 CLOSE
```

The values x_i will appear in the sequential file \FRACUM2.TXT, ant the ones for $f(x)$ in FRACUM3.TXT.

In our case we have obtained the following table:

Population reduced to Columbian one

x	x’	x’/71	F	FR	FRAC	FRAC %
1	1	1.41%	4	4.49%	4.49%	1.41%
2	2	2.82%	2	2.25%	6.74%	2.82%
3	3	4.23%	2	2.25%	8.99%	4.23%
5	4	5.63%	2	2.25%	11.24%	5.63%
6	5	7.04%	3	3.37%	14.61%	7.04%
8	6	8.45%	2	2.25%	16.85%	8.45%
9	7	9.86%	2	2.25%	19.10%	9.86%
10	8	11.27%	2	2.25%	21.35%	11.27%
11	9	12.68%	1	1.12%	22.47%	12.68%
12	10	14.08%	2	2.25%	24.72%	14.08%
13	11	15.49%	2	2.25%	26.97%	15.49%
14	12	16.90%	1	1.12%	28.09%	16.90%
15	13	18.31%	2	2.25%	30.34%	18.31%
16	14	19.72%	1	1.12%	31.46%	19.72%
17	15	21.13%	1	1.12%	32.58%	21.13%
18	16	22.54%	2	2.25%	34.83%	22.54%
20	17	23.94%	1	1.12%	35.96%	23.94%
22	18	25.35%	1	1.12%	37.08%	25.35%
24	19	26.76%	1	1.12%	38.20%	26.76%
25	20	28.17%	2	2.25%	40.45%	28.17%
28	21	29.58%	1	1.12%	41.57%	29.58%
31	22	30.99%	1	1.12%	42.70%	30.99%
33	23	32.39%	2	2.25%	44.94%	32.39%
34	24	33.80%	1	1.12%	46.07%	33.80%
35	25	35.21%	1	1.12%	47.19%	35.21%
38	26	36.62%	1	1.12%	48.31%	36.62%
44	27	38.03%	1	1.12%	49.44%	38.03%
45	28	39.44%	1	1.12%	50.56%	39.44%
46	29	40.85%	1	1.12%	51.69%	40.85%
47	30	42.25%	2	2.25%	53.93%	42.25%
54	31	43.66%	1	1.12%	55.06%	43.66%
65	32	45.07%	1	1.12%	56.18%	45.07%
73	33	46.48%	1	1.12%	57.30%	46.48%
77	34	47.89%	1	1.12%	58.43%	47.89%
90	35	49.30%	1	1.12%	59.55%	49.30%
94	36	50.70%	1	1.12%	60.67%	50.70%
111	37	52.11%	1	1.12%	61.80%	52.11%
132	38	53.52%	1	1.12%	62.92%	53.52%

149	39	54.93%	1	1.12%	64.04%	54.93%
174	40	56.34%	1	1.12%	65.17%	56.34%
199	41	57.75%	1	1.12%	66.29%	57.75%
245	42	59.15%	1	1.12%	67.42%	59.15%
277	43	60.56%	1	1.12%	68.54%	60.56%
328	44	61.97%	1	1.12%	69.66%	61.97%
365	45	63.38%	1	1.12%	70.79%	63.38%
366	46	64.79%	1	1.12%	71.91%	64.79%
390	47	66.20%	1	1.12%	73.03%	66.20%
401	48	67.61%	1	1.12%	74.16%	67.61%
450	49	69.01%	1	1.12%	75.28%	69.01%
543	50	70.42%	1	1.12%	76.40%	70.42%
639	51	71.83%	1	1.12%	77.53%	71.83%
711	52	73.24%	1	1.12%	78.65%	73.24%
966	53	74.65%	1	1.12%	79.78%	74.65%
931	54	76.06%	1	1.12%	80.90%	76.06%
976	55	77.46%	1	1.12%	82.02%	77.46%
1022	56	78.87%	1	1.12%	83.15%	78.87%
1094	57	80.28%	1	1.12%	84.27%	80.28%
1200	58	81.69%	1	1.12%	85.39%	81.69%
1248	59	83.10%	1	1.12%	86.52%	83.10%
1318	60	84.51%	1	1.12%	87.64%	84.51%
1453	61	85.92%	1	1.12%	88.76%	85.92%
1590	62	87.32%	1	1.12%	89.89%	87.32%
1645	63	88.73%	1	1.12%	91.01%	88.73%
1832	64	90.14%	1	1.12%	92.13%	90.14%
1943	65	91.55%	1	1.12%	93.26%	91.55%
1948	66	92.96%	1	1.12%	94.38%	92.96%
2001	67	94.37%	1	1.12%	95.51%	94.37%
2111	68	95.77%	1	1.12%	96.63%	95.77%
2143	69	97.18%	1	1.12%	97.75%	97.18%
2235	70	98.59%	1	1.12%	98.88%	98.59%
2321	71	100.00%	1	1.12%	100.00%	100.00%
			89	**Sum:**	41.42	
					0.5	
					40.92	
				S=	0.58451	

In this table, *x* represents the value of the generic name for the "Population BCN-10-m". *x*' is the value obtained randomly, but this time referring to its own table, which contains (in our case) 71 different names, which will be considered as abscissas of the "Columbian reduced Population," and they will allow us to do the same calculations for other populations.

Indeed, the same table is used to obtain the area of the curve by the method of approximate integration seen in Annex 1. The sum of the ordinates is 41.42, from which .5 must be deducted from the sum and then divide the result by the number of intervals (or multiply by the amplitude of each). Thus the value of the subtended area between the curve and the abscissa is obtained. It comes to 0.58451.