Distribution of Forenames, Surnames, and Forename-Surname Pairs in the United States

D. K. Tucker

Esson and Tucker Management

Unavailability of data and computational resources has generally limited the study of personal names to that of individual forenames and surnames, small populations, and dictionaries of name types. Considerable attention has been paid to the comparative popularity of forenames but little to the frequency distribution of forenames, surnames, and forename-surname pairs. Frequency distributions for names in the United States are presented and are seen to approximate power law curves. The paradox of the commonality of the rare forename or surname is investigated and the puzzle of the plot of the occupied frequencies is presented.

Introduction

For many years people have struggled to get some idea of the number of forenames and surnames in a major country such as the United States and determine the relative popularity of particular forenames and surnames. Here I set out to describe the actual distribution of personal names in the United States. For my purposes, *name* will be used, unless otherwise noted, to mean personal name, either a forename or surname.

Assuming that everyone in the United States has a personal name which is comprised, as a minimum, of a forename and a surname, we can say that if the population is x, then there are x surnames, x forenames used as first names, and x forename-surname pairs. My name, David Kenneth Tucker, would have David as the forename, Tucker as the surname, and David Tucker as the forename-surname pair. Kenneth does not feature further in this discussion, as it is almost impossible to obtain such information on a grand scale, whereas the other information is readily available from CD-based telephone

Names 49.2 (June 2001):69-96
ISSN:0027-7738

directories, albeit with their well-known limitations. (For a discussion of these limitations, see Hanks and Tucker 2000.)

Many people share their forename and surname with others; some have unique forenames or surnames, or both. Each name that is different from all other names is a name *type* and every example of that name is a name *token*. There are 1,321,612 tokens of the type David, 56,636 tokens of the type Tucker, and 807 tokens of the type David Tucker in the directory. David, Tucker, and David Tucker represent three name classes: forename, surname, and forename-surname pair. We know that the total number of tokens of all the types within a type-class equals the population, but what is not obvious is the relationship between the types and tokens. Both *David* and *Tucker* are popular forename and surname types respectively, so how many types are there in the population? This article answers that question and a few others, but in turn raises questions for others to answer.

The source data was the 1997 edition of INFOUSA ProCD Select Phone, a pack of six CDs listing almost 100 million telephone subscribers. Using the standard export function supplied with the product and the greater than 50,000 records export facility authorized by an unlock code from ProCD, the subscriber name and state for all residential listings, as opposed to business listings, were extracted.

The extract was subjected to extensive analysis to remove the remaining non-residential listings such as municipalities, universities, services, hospitals, religious houses, utilities, military, and others. The compound names were repaired where necessary[1] and the individual forenames extracted and extraneous qualifiers, such as *Realtor, The Man,* and *Psychologist,* removed.

Extraction and analysis revealed the following statistics, as shown in tables 1 and 2.

Table 1. Number of Types and Tokens (in Millions).

Class Type and Class Tokens	Count
Surname Tokens	88.7
Unknown Forename Tokens	15.7
Forename Tokens	73.0
Forename-Surname Pairs Tokens	73.0
Surname Types	1.75
Forename Types	1.25
Forename-Surname Pairs Types	27.3

For the sake of completeness, the mean and standard deviation of tokens per type for each class is given in table 2, but as we shall see, because of the skewness of the distribution, these measures are of little value.

Table 2. Means and Standard Deviations.

Type	Mean	Standard Deviation
Surname	51	1544
Forename	58	4703
Pairs	3	70

Unknown means that a forename was shown to exist but it was unknown. An entry such as *Mr. and Mrs. Frank Churchill* shows one forename but the other is unknown; the two forename-surname pairs from this entry are thus *Frank Churchill* and *unknown Churchill.* In this case we know that *unknown Churchill* is a female. In the case of *Mr. and Mrs. F. Churchill* we get two *unknown Churchill* forename-surname pairs: one female and the other male. We may deduce that the majority of unknown forenames are thus forenames of females. In the case of an entry such as *Mark and Karen Mulligan* we see two known forenames, with the two forename-surname pairs: *Mark Mulligan* and *Karen Mulligan.*

We see from table 1 that there are 73 million known forenames, and that there is evidence of another 15.7 million forenames but what they are is unknown; we will call these the unknown forenames. The vast majority of these unknown forenames, if we knew them, would probably be subsumed within the 1.25 million forename types.

There are 27.3 million forename-surname pairs, not counting the unknown forename-surname pair types. This number is surprisingly low as the number of different surnames and the number of different forenames would allow over 2 million million $(2.10\hat{\ }12)^2$ unique forename-surname pairs; more than enough to allow every American a unique name. We thus suspect that there is order in the naming of people.

Graphic Representation of the Data

Cumulative Curves: Tokens Plotted Against Types

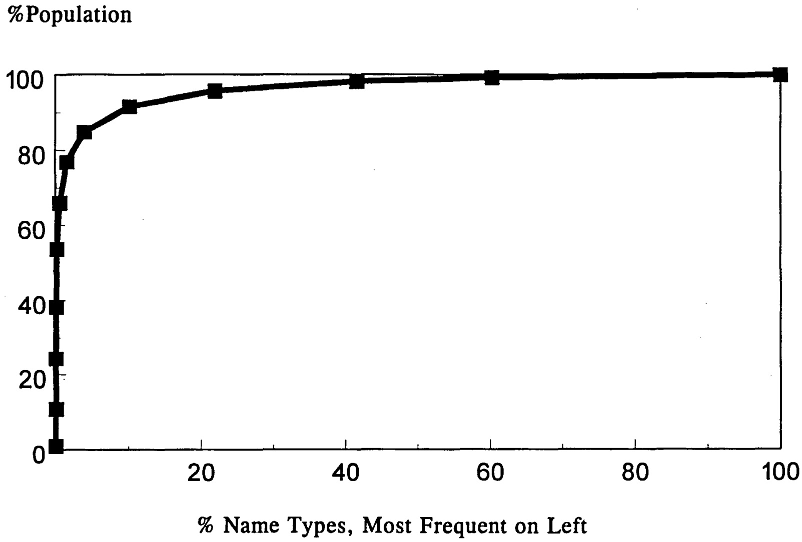
Even a cursory look at almost any telephone directory would show that there are many people with popular names and also many people with rare names, but we need to be more descriptive than this. For example, a dictionary publisher may ask: "What is the smallest number of name types required to include 75% of the population?" Since both the names and their frequencies are available, this can easily be calculated. For example, we can arrange the names in order of their descending frequency beginning with the surname *Smith*, which is the most common surname in America with a frequency, or count, of 832 thousand (832k).

We know that the sum of the counts for all name types must equal the total name tokens, which is the population, so we know for each name type what proportion of the population it covers. Our sample population is 88.7 million, so *Smith* represents almost 1% of that population; in fact, 0.937749%, to be more exact. However, it is only one name type in 1.75 million types, or 0.000057% of the name types. Thus the origin of our graph is at the point where 0.937749 and 0.000057 intersect. We can now add the next most frequent name, *Johnson* (with a count of 610k), to the list. The cumulative effect of adding *Johnson* to *Smith* is to generate a point at 1.625577; 0.000114. We can continue to do this until we have added all the name types in descending frequency order until we arrive at the final point 100; 100, which says that all the name types (100%) represent all the name tokens, or population (100%).

If we plotted the results on a normal graph with linear scales for population and names, we would get a graph that looks like figure 1. The graph starts near the 0; 0 point, rapidly rises to about 90; 10 and then slowly rises to 100; 100. It is difficult to understand what is going on in this presentation, as all the activity seems to take place for low values of percentage of name types.

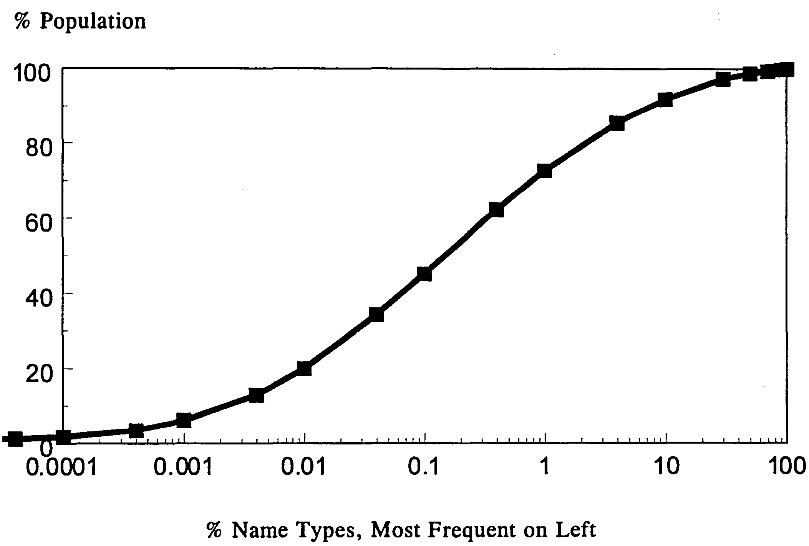
The graph shown as figure 1 has linear scales for both its axes; thus it is lin-lin. A linear scale is one where the increment is constant. Starting at, say, 0, the scale goes to 10, 20, and so on up to, say, 100. A logarithmic scale, in contrast, is one where the increment is the power of a base number. Consider a base number of 10. We might start at 10^-4, which is 0.0001, and increase by 10 times each increment: to 10^-3, which is 0.001, and so on up to 10^2, which is 100.

Figure 1. U.S. Surnames Distribution–Linear Plot.



If we use a logarithmic scale for the x-axis (% of name types) we get the plot shown in figure 2, where a log-lin, or semi-log, plot allows us to see much more detail.

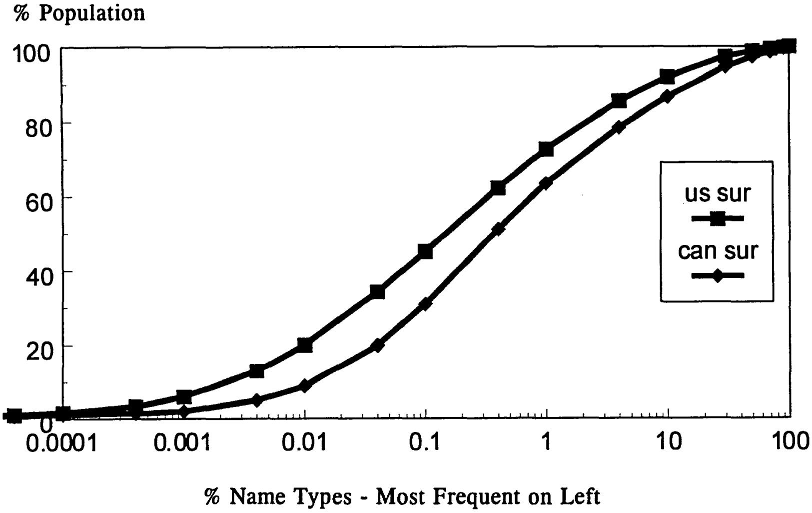
Figure 2. U.S. Surnames Distribution–Semi-Log Plot.



We can see for example that the most popular 1% of name types accommodate over 70% of the American population, and that 90% of the name types, from 10% to 100%, the rare name types, accommodate a mere 9% of the population. The distribution of surnames in the U.S. is thus highly skewed.

In order to see if the U.S. situation is unique, we can compare the distribution of surnames in Canada. When we do, we find that Canadian surnames show a similar skewness (figure 3).

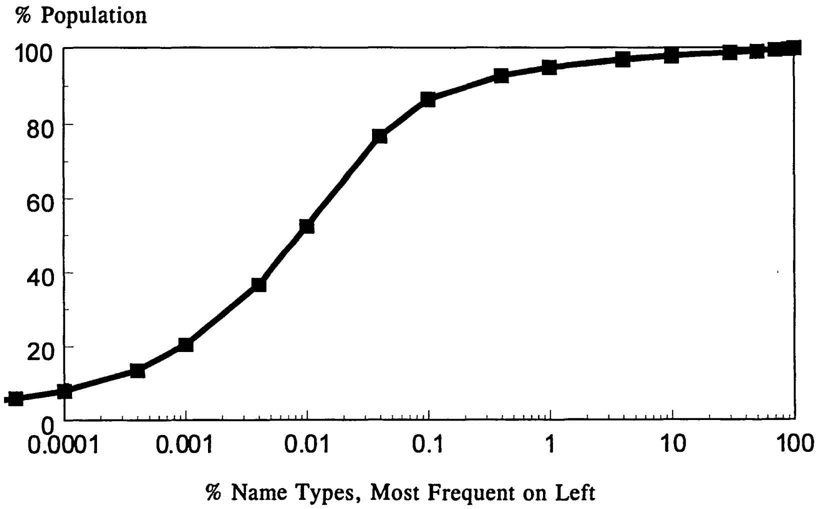
Figure 3. U.S. and Canada Surname Distributions.



It should be pointed out that although Canada has only one tenth the population of the U.S., it is possible to plot both on the same curve as the results have been normalized by using percentages. It should also be noted that both Canadian and U.S. personal names have a multilingual nature but those in the UK are, for the most part, unilingual. However, from other data which I have, if we plotted the UK curve, it would lie between the curve for the U.S. and that for Canada. This suggests that the shape of the curve is not peculiar to the U.S., or to Canada, but is intrinsic to at least some surname distributions.

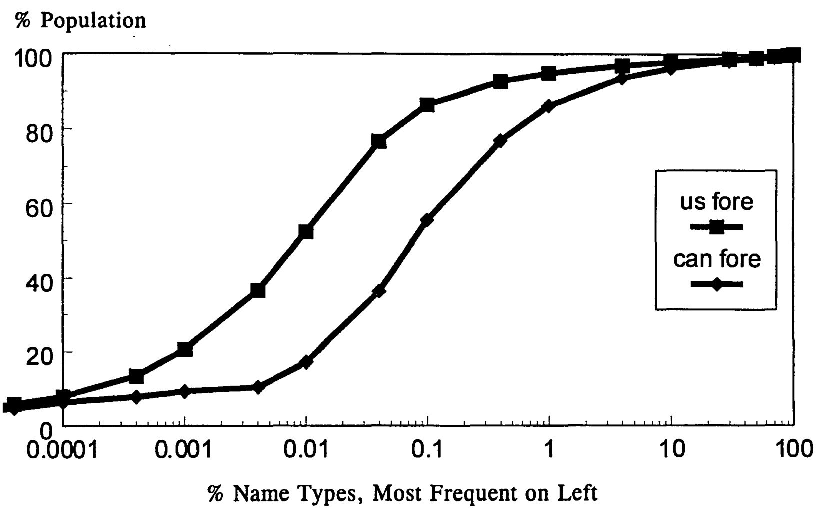
Forenames can be plotted in the same way as surnames, as shown in figure 4.

Figure 4. U.S. Forenames Distribution.



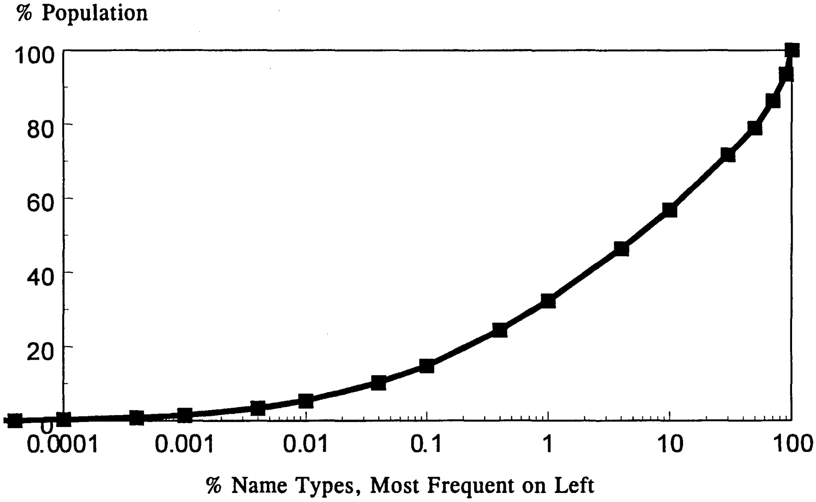
This curve rises faster than the surname curve and shows that 1% of forename types, the most popular, accommodate about 95% of the population. This means that 99% of forename types are shared by only 5% of the population. The forename distribution is more skewed than the surname distribution. For comparison, the Canadian forename curve is shown in figure 5, where we see again the same skewness as we found in the U.S. curve.

Figure 5. U.S. and Canada Forenames Distribution.



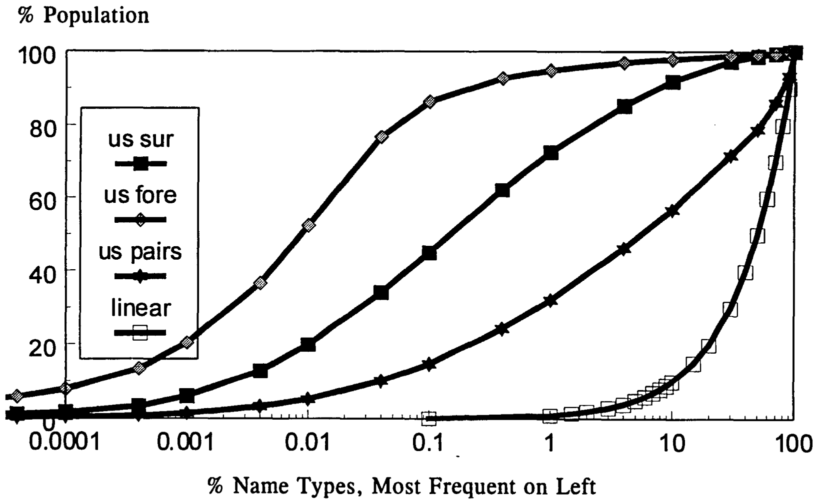
The next plot is forename-surname pair types, shown in figure 6.

Figure 6. U.S. Distribution of Forename-Surname Pairs.



This distribution is less skewed than the surname distribution, but it is still skewed; 1% of the forename-surname pair types accommodates nearly a third of the population. The final plot, figure 7, shows the U.S. forename, surname, forename-surname pairs distribution and, for comparison, a curve for a hypothetical distribution where x% of the name types would represent x% of the population; in other words, a totally unskewed distribution.

Figure 7. Forenames, Surnames, and Forename-Surname Pairs.



Non-Cumulative Representations

The cumulative curves are very useful in describing the distribution in everyday terms, but other researchers, such as Ogden (1998), have attempted to identify the frequency at which a name type appears with a given count; in other words, the number of types with a given number of tokens. Going back to our surname data we find that there are about 707k surname types that are unique; they have only one token each. Since there are only 1.75 million surname types to begin with we come to the stunning conclusion that about 40% of all surname types are unique. The paradoxical observation is that it is not uncommon to have the rarest name in the country since there are 707k rarest name *equals* in the population. To have a rare name is less common than having the name *Smith* but more common than having the second most popular name, *Johnson.*

We have described frequency as count: the number of tokens for a particular name type. The frequency range of our surname data is 1 to 832k. As we have seen with unique types, it is not uncommon for a number of types at low frequencies to have the same frequency. There are 707k surname types with a frequency of one, 222k with a frequency of two, and 115k with a frequency of 3. Ranked by increasing frequency the type count is generally descending but there are exceptions. The first occurs at a frequency of 36 which is shared by 3302 types, but more, 3320 types, share a frequency of 37.

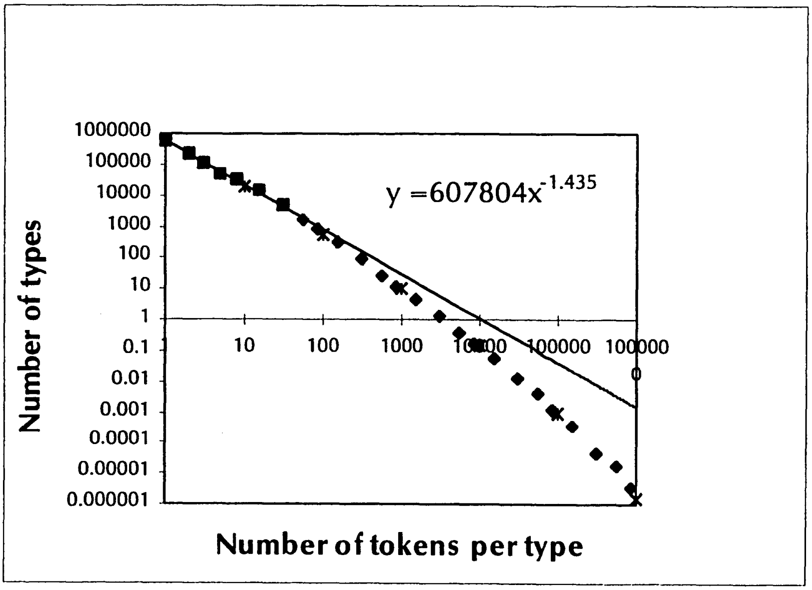
However, not all frequencies have name types. The general decline of number of types sharing a frequency continues with increase in frequency until the number of types sharing a frequency reaches zero. Of the stated range, only 5,845 have frequencies less than 1%. The first empty frequency is at 1,373; this means that there are no name types with 1,373 tokens. The gaps get bigger with increase in frequency; there are 221,678 empty frequencies between 610,104 (*Johnson*) and 831,783 (*Smith*). There are thus two related events as we increase frequency; there is a reduction in the number of types at that frequency, and an increase in the empty frequencies, hence more and bigger gaps.

One can see how the number of names with one token, two tokens, etc., can be plotted but what can we do about the gaps? I had trouble with this until Dr. Trevor Ogden suggested that I consider particles in the air where there are many small particles and fewer large particles. In attempting to determine the frequency of particles of a certain size, a progressive filter is built and the number of particles trapped at each stage allows the frequency to be calculated. Consider a three-stage filter with the first stage capturing particles between 20 and 10 microns, the second stage between 10 and 4 microns, and the final stage between 4 and 0 microns. Say that the filter captures 3, 13, and 29 particles, respectively.

The first plot would be at the mean position of the range, i. e., (20+10)/2=15 microns. The value would be the count divided by the range, i. e., 3/(20-10)=0.3 particles per micron. The second plot would be 7 microns with 2.17 particles per micron and the third plot at 2 microns with 7.25 particles per micron.

Surnames, of course, are different than particles and can only be integers; a person cannot have 1.3 surnames. So this averaging needs only to be introduced prior to any gaps occurring in the sequence. We soon find that the data are best plotted on a log-log scale with the resulting curve shown in figure 8.

Figure 8. U.S. Surname Frequency.



Ogden (1988) generated such curves for UK data (using smaller samples) and found that the data approximates to a power law curve.[3] Indeed a respectable fit for x < 100 would be:

y=607804*x^(-1.435).

However, the curve drops away substantially for higher values of x. I am indebted to Ogden for fitting a curve to the data and deriving population and number of types from the fitted curve. The fitted curve is described by the expression:

y=875000(x^-1.435)*1.25^(-(x^0.263)).

This predicts 700k names occurring once, a total of 1.73 million surname types and a population of 88.2 million. The actual data are 707k, 1.75 million, and 88.7 million respectively. This is a remarkably good fit.

However, with surnames, forenames, and forename-surname pairs, the unique frequencies are overstated because that is where the typos and other detritus settle. Nothing other than eyeballing these data for non-names and having knowledge of all legitimate forms is required to resolve this. Unfortunately, this knowledge is not available and the difficulty of determining whether or not a particular sequence of characters is a name, is anything but a trivial task, and in some cases may be insoluble. For instance, are *Spring Sage, Sodny, Skky, Syxx* and *Shh'kyia* personal names or not?[4]

The points predicted by this expression are shown in figure 8 as asterisks. I have no idea why this curve fits; I only know that it does. I have attempted to describe the *what* of name distributions; I am hopeful that there are experts in the growth of surnames who will find the data useful and who will be able to tell us the *why*. Population can be derived from the Non-Cumulative Curves; it is the product y*x. where y is the frequency and x is the number of population at that frequency. Taking the surnames as an example, the population p=y*x, which is:

875000(x^-1.435)*1.25^(-(x^0.263))*x

which simplifies to

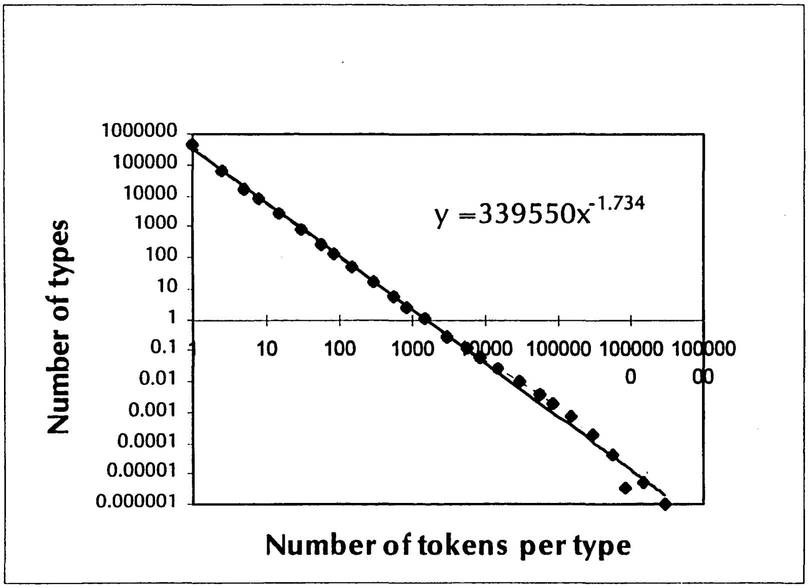
875000(x^-0.435)*1.25^(-(x^0.263)).

This gives the population for a particular x value. To get the population over a range of x it is necessary to integrate y.dx over the desired range.

The number of types is the sum of y/x for each x value. As an example, the value of y for x=1 is, from the formula, 700k. This divided by x equals 700k types. For x=10 the formula gives a value of 21,452 which divided by 10 gives 2,145 types. For x=100 the value is 558 which gives 6 types, and so on.

The curve for forenames plotted in the same way as for surnames, is also a power law curve, as shown in figure 9.

Figure 9. U.S. Forenames Frequency.



The best fit for the complete series is again a simple power law relationship:

y=339,550*x^(-1.734).

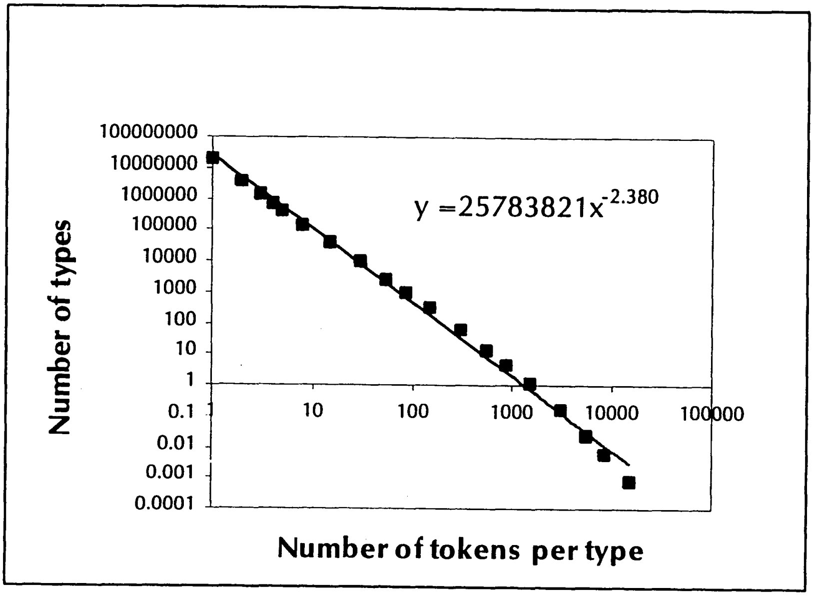
However, this underestimates the number of unique forenames as 340k, whereas the actual sample number is 879k. However, the sample number itself is overstated as this is where the flotsam and jetsam gravitates to: mainly typographical errors. More work is required to further rationalize, downward, the number of unique forenames.

The curve for forename-surname pairs plotted in figure 10 in the same way as for forenames, is a power law curve:

y=25,783.821*x^(-2.380).

This gives an overestimate of the number of unique forename-surname pairs of 25.8 million whereas the actual number is 20.3 million, but the estimate is in the same general area.

Figure 10. U.S. Forename-Surname Pairs Frequency.



Zipf's Law and Mandelbrot's Generalizations

Zipf's Law (1949) can be stated as: the frequency of each type in a large corpus multiplied by the rank of the type is a constant, where the constant is peculiar to the text under consideration: frequency * rank= constant, or frequency=constant/rank.

This law is widely referred to in the study of natural-language text corpora and it has been suggested that it might hold for personal names. For surnames, the mean (frequency*rank); that is, the constant for the 88.7 million surnames was calculated at 2,774,861 with a standard deviation of 1,851,396. Comparison of the actual frequency (count) against the predicted Zipf frequency with this constant shows no obvious correlation between the two overall.

However, for the first 100 surnames, frequency, tokens per type, plotted against rank shows a power law relationship described by:

Frequency=1000000/(Rank^0.59).

Mandelbrot (1959) generalized Zipf's Law by introducing an adjustable constant and modification of the power, in this case from the calculated mean of 2,774,861 to 1 million, and 1.00 to 0.59, respectively. However, the relationship breaks down at about rank 200, thereafter projecting higher than real frequencies.

For forenames, the mean (frequency*rank); that is, the constant, for the 73 million forenames, was calculated at 790,871 with a standard deviation of 397,565. Comparison of the actual frequency (count) against the predicted Zipf frequency using 790,871 as the constant, shows, again, no obvious correlation overall.

However, frequency plotted against rank for the first 100 surnames shows a similar power law relationship described by:

Frequency=3212507/(Rank^0.68).

Here again, both the constant and power have been modified from 790,871 to 3,212,507 and from 1.00 to 0.58 respectively. However, as in the case of surnames, the relationship breaks down at about rank 200, and thereafter projecting higher than real frequencies. No doubt the relationships can be further modified to better reflect the actual data but this is beyond the scope of this investigation.

The Simon Equation

In mentioning Mandelbrot, it is necessary to also mention Herbert Simon. Ogden (1998) refers to a work by Simon which he (Ogden) adapted to the study of surnames with some success. However, Simon's paper of 1955 was challenged by Mandelbrot in 1959 and the lengthy correspondence ended in 1961 with no agreement. The Simon equation, which uses a type against tokens/type plot, does not provide as good a fit as the expressions given: "the Simon equation gives close to y=x^-1.85 for small and moderate values of x, so it will have a steeper gradient than the U.S. data"(Ogden 2000).

Population Curves for Occupied Frequencies

In the cumulative curves we plotted population against name types. In the non-cumulative curves we plotted number of name types against frequency, or tokens per type. For example there were 707k unique surnames; i.e, each had one token. In plotting those curves we made use of averaging over the higher frequency values because of the gaps in the frequencies. In examining Zipf's Law we were interested only in rank and the gaps were not an issue. In this section we will look at another way to calculate population by looking at the occupied frequencies only. In this case we ignore the gaps; of the 832k frequencies we will use only the 5,844 occupied frequencies and plot the population for each frequency. The resulting plot is shown in figure 11.

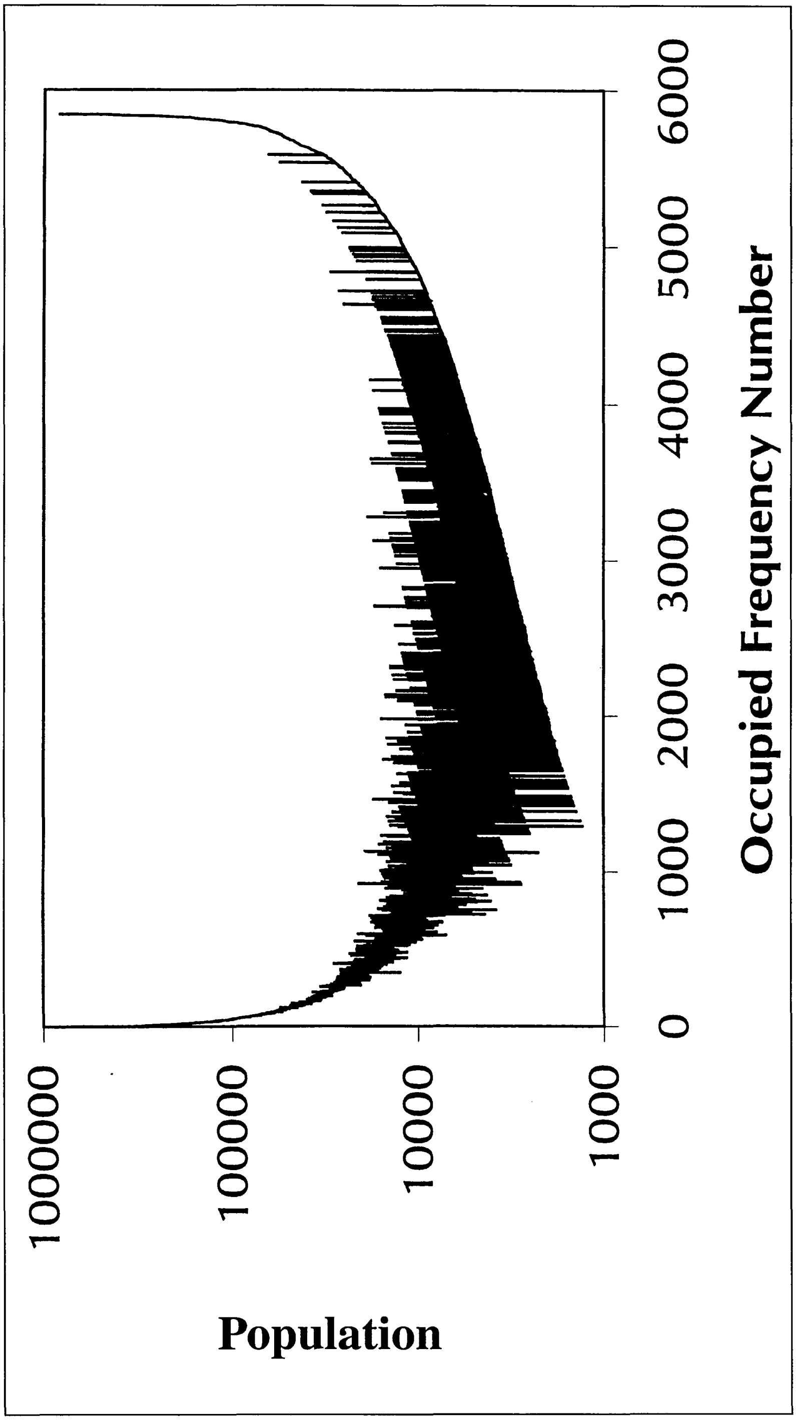


Figure 11. Occupied Frequencies.

This is not the latest in Viking longboat keel design, but it is an unusual curve. The verticals are lines joining adjacent plots. The origin and finish are not easy to see but are the population value of 707k for the first occupied frequency, and 832k for the last at 5844. The curve descends from the origin with an overall reduction in the number of types. However the number of types vibrates about this downward trend which gives the first part of the curve its fuzziness. This is to be contrasted with the smoothness of the finish of the curve where there is only one type. The curve reaches a minimum at a frequency of 1287, the first that has just one type. This is the frequency that represents minimum population. (The surname at frequency 1287, incidentally, is *Hord*.) The line of *one types* continues from there along the bottom line of the curve until it reaches the end. Descending from the end we can see the last of the *two types* at frequency 5806 on the x axis. We can thereon see the density of the *two types* grow and also see the *three types* and *four types* emerge until further increases in types are lost in the detail.

The beginning and end shapes of the curve seem to mirror each other. We have already established the general shape of the far end of this curve since the number of types at these high frequencies is 1; then the population per frequency is the same as the surname at this frequency.

We established that for the first 100 high frequency surnames:

Frequency = 1000000/(Rank^0.59).

So for this same group:

Population (hi-freq) = 1000000/(Rank^0.59).

Note that in terms of the expression we are at the high frequency end and looking at the next 99 lower frequencies.

At the low frequency end the population is the frequency multiplied by the number of types at that frequency. For a frequency of 1 the population is 707k, for 2 it is 444k, for 3 it is 344k, and so on. The first 100 populations are described by the power law:

Population (*lo-freq*) = 877606/(Rank^0.58).

The *hi-freq* and *lo-freq* population expressions are thus virtually mirror images. It is an odd characteristic of name distribution that the population maximums are at the beginning and end of the distribution curve. Table 3 shows how the populations are interlaced.

Table 3 is ordered in descending population with the maximum occurring at line 1 at the highest frequency: 831,783 (832k). The second highest population occurs at line 2 at the lowest frequency: 1, where there are 706,762 (707k) surnames at that frequency. The frequency extremes are not only shown by the absolute frequencies 1 and 831,783 but also by the population # which is the number for occupied frequencies being 1 and 5,844, respectively.

The table shows clearly the interlacing of the increasing low frequency population with the decreasing high frequency population; the curve shapes of each are mirror images of each other. There are 24 ascending frequencies and 26 descending frequencies. Again, this seems to be more than mere chance.

The population curve for forenames has the same overall shape as the surname curve but with a significant difference. It is U shaped with a minimum population at frequency 416 with the forename *Robert Scott.* The next minimum is *Carissa* at 573. However, the rise for the second upright of the U is anything but a mirror image of the rise on the first. Interlacing is weak in that the high frequency names occupy 97 of the top 100 population frequencies; the other three being frequencies 1, 2, and 3 at positions 8, 29, and 72.

In the surname case the population for the most popular name was of the same order as the population for all the rare names of frequency one: 832k cf. 707k. In the forename case the frequency of the most popular name, *John,* is over twice the size of the population for all forenames of frequency one: 2230k cf. 879k. Even within the 879k there is perhaps more flotsam and jetsam than in the surname case.

Popular Types: Surnames

Table 4 lists in descending frequency order the 50 most popular surnames of the 1.75 million types in the U.S. The count given is out of the sample population of 88.7 million tokens. The Zipf-Mandelbrot numbers in the last column are those predicted from the formula discussed above.

The list compares well with that provided by the U.S. Census Bureau (http://www.census.gov.genealogy/names), which is based on a sample of 6 million tokens with deliberate over-sampling for African-Americans and Hispanics.

Table 3. Maximum Population per Frequency.

LINE#	POP#	FREQUENCY	# AT FREQ	POPULATION
1	5844	831783	1	831783
2	1	1	706762	706762
3	5843	610104	1	610104
4	5842	452360	1	452360
5	5841	447208	1	447208
6	2	2	221848	443696
7	5840	432177	1	432177
8	5839	421078	1	421078
9	5838	354880	1	354880
10	3	3	114683	344049
11	4	4	81917	327668
12	5	5	61991	309955
13	6	6	49519	297114
14	5837	285232	1	285232
15	7	7	40570	283990
16	8	8	34034	272272
17	5836	269682	1	269682
18	9	9	28835	259515
19	10	10	24765	247650
20	5835	241254	1	241254
21	11	11	21770	239470
22	5834	239230	1	239230
23	5833	238747	1	238747
24	12	12	19198	230376
25	5832	226220	1	226220
26	13	13	16975	220675
27	5831	219271	1	219271
28	5830	217049	1	217049
29	14	14	15295	214130
30	15	15	13659	204885
31	16	16	12324	197184
32	5829	195819	1	195819
33	17	17	11242	191114
34	18	18	10340	186120
35	5828	185777	1	185777
36	5827	184136	1	184136
37	19	19	9507	180633
38	5826	180338	1	180338
39	20	20	8696	173920
40	21	21	7945	166845
41	5825	166842	1	166842
42	5824	166370	1	166370
43	22	22	7524	165528
44	5823	163390	1	163390
45	23	23	7041	161943
46	5822	160864	1	160864
47	5821	160009	1	160009
48	5820	158845	1	158845
49	24	24	6593	158232
50	5819	153615	1	153615

Table 4. Most Popular Surnames in the United States.

RANK	SURNAME	COUNT	ZIPF-MAN
1	Smith	831783	1000000
2	Johnson	610104	664343
3	Williams	452360	522996
4	Brown	447208	441351
5	Jones	432177	386908
6	Miller	421078	347449
7	Davis	354880	317243
8	Anderson	285232	293209
9	Wilson	269682	273525
10	Taylor	241254	257040
11	Moore	239230	242984
12	Martin	238747	230825
13	Thompson	226220	220178
14	Thomas	219271	210758
15	White	217049	202351
16	Clark	195819	194791
17	Harris	185777	187947
18	Jackson	184136	181714
19	Lee	180338	176009
20	Lewis	166842	170762
21	Hall	166370	165917
22	Walker	163390	161425
23	Young	160864	157246
24	Nelson	160009	153347
25	Allen	158845	149698
26	King	153615	146274
27	Robinson	153159	143052
28	Baker	148669	140016
29	Wright	148099	137147
30	Adams	144377	134431
31	Hill	141823	131855
32	Scott	137971	129408
33	Roberts	132659	127080
34	Campbell	132126	124861
35	Green	131873	122744
36	Phillips	126669	120721
37	Mitchell	122922	118785
38	Evans	117387	116930
39	Carter	116042	115152
40	Murphy	115601	113445
41	Parker	112936	111804
42	Turner	112377	110226
43	Peterson	110846	108706
44	Morris	110158	107242
45	Cook	109743	105829
46	Stewart	109121	104465
47	Collins	107617	103148
48	Rogers	106345	101875

49	Garcia	105882	100643
50	Edwards	105393	99451
51	Wood	98424	98295
52	Morgan	97713	97176
53	Kelly	94726	96090
54	Cox	94703	95036
55	Martinez	94105	94013
56	Rodriguez	94100	93018
57	Bailey	93393	92052
58	Cooper	92926	91112
59	Reed	92556	90198
60	Ward	92242	89308
61	Bell	89728	88441
62	Sullivan	86937	87597
63	Bennett	86539	86774
64	Myers	84848	85971
65	Gray	84423	85189
66	Hughes	84186	84425
67	Howard	84046	83679
68	Long	83277	82951
69	Watson	82750	82239
70	Ross	81892	81544
71	Richardson	81637	80864
72	Price	80852	80200
73	Russell	79186	79550
74	Fisher	78653	78914
75	Brooks	78647	78291
76	Foster	76761	77682
77	Powell	74080	77085
78	Hernandez	73728	76500
79	Perry	72800	75928
80	Olson	72486	75366
81	Reynolds	72366	74816
82	Lopez	72076	74276
83	Butler	70457	73747
84	Sanders	70393	73228
85	James	70272	72718
86	Barnes	70136	72218
87	Graham	69312	71727
88	Henderson	69047	71245
89	Hamilton	68294	70772
90	Patterson	67787	70307
91	West	67177	69850
92	Cole	66813	69401
93	Jenkins	66617	68960
94	Murray	66484	68526
95	Wallace	66195	68099
96	Gonzalez	65991	67680
97	Stevens	65676	67267
98	Meyer	65510	66862
99	Hayes	64858	66462
100	Kennedy	64834	66069

Popular Types: Forenames

Table 5 lists in descending frequency order the 100 most popular forenames of the 1.25 million types in the United States. The count given is out of the sample population of 73 million tokens. The Zipf-Mandelbrot numbers in the last column are those predicted from the formula discussed earlier.

The forenames listed, it should be noted, are self-declared. The forenames may appear to be contractions, diminutives, nicknames, and so forth, but this is the way the people list themselves. A fine example would be *Willie Williamson.* No attempt has been made to correct the form with the exception of standard abbreviations such as *Edw, Robt,* and *Wm* for *Edward, Robert,* and *William,* respectively, that have been expanded. However, even this rule is broken for *Chas,* which may be an abbreviation for *Charles* but is treated as a name in its own right. The over-riding rule is that if what is written can be said it is a name.

Where the forename is made of multiple segments, all segments are included in the name, even when there is no hyphen, such as in the forenames *Johnnie Gay* (1), *John Robert* (1421), *Jose Luis* (6791), *Willie Mae* (5259), *Ann Marie* (3982), *Le Roy* (3937), and *Yuk Shing,* (9). It can be argued that in some cases the person is merely listing their forenames such as in *John Robert* above; perhaps so, perhaps not.

The list compares well with the U.S. Census Bureau list of male forenames. There is no gender information in the sample used for this study; thus I am aware of the dangers of discussing "male" and "female" name lists especially as there is considerable evidence, e.g., Schwegel (1997) that girls are being given names that were previously considered to be exclusively male names, such as *John, Robert, William, James,* and *David.* However, on the assumption that the vast majority of usage of these names is still for males, I will in the following discussion include these names as male. By *unisex* I mean names that are currently recognized in society as being used by either gender, names such as *Leslie.*

Few female forenames appear on the list. The low count for forenames of females is a function of the source. Women listed with men are often in the form of *Mr. and Mrs. John Smith* and sometimes simply not listed in the household entry. Furthermore, some women tend not to use their forenames in phone listings for security reasons, especially solo women. From table 3 we know that the missed *Mrs.* is part of 15.7 million unknown forenames, so the lower counts are to be expected.

Table 5. Most Popular Forenames in the United States.

RANK	FORENAME	COUNT	ZIPF-MAN
1	John	2229952	3212507
2	Robert	2057921	2005135
3	James	1508651	1521954
4	William	1487740	1251536
5	David	1321612	1075337
6	Michael	1147838	949951
7	Richard	1147833	855416
8	Charles	739153	781165
9	George	684525	721040
10	Paul	674480	671188
11	Thomas	660147	629067
12	Donald	628017	592926
13	Joseph	577578	561517
14	Mark	549143	533921
15	Edward	527459	509451
16	Frank	489097	487576
17	Kenneth	463649	467885
18	Mary	451437	450048
19	Gary	446755	433802
20	Larry	403023	418932
21	Ronald	401590	405261
22	Daniel	361605	392642
23	Jack	333796	380951
24	Scott	312515	370084
25	Steve	304057	359952
26	Jerry	302528	350479
27	Jas	299949	341599
28	Harold	298175	333255
29	Steven	297890	325397
30	Raymond	292281	317981
31	Dennis	289236	310970
32	Stephen	283850	304328
33	Mike	276080	298026
34	Walter	275506	292037
35	Joe	272222	286337
36	Brian	270140	280904
37	Peter	261327	275719
38	Kevin	260339	270764
39	Fred	258937	266024
40	Jim	256447	261483
41	Linda	250828	257129
42	Carl	245456	252950
43	Bill	244173	248935
44	Anthony	243216	245073
45	Jeff	234752	241357
46	Roger	230729	237776
47	Henry	228600	234324
48	Don	227390	230994
49	Ralph	225511	227777
50	Gerald	224118	224670

51	Arthur	223092	221665
52	Tom	222777	218757
53	Wayne	220757	215942
54	Susan	218611	213214
55	Barbara	216646	210570
56	Terry	215029	208006
57	Chris	214246	205518
58	Bruce	211648	203101
59	Harry	210612	200754
60	Douglas	203674	198473
61	Jos	196894	196255
62	Albert	196884	194097
63	Chas	191131	191996
64	Roy	190346	189951
65	Howard	186461	187959
66	Karen	186229	186018
67	Jeffrey	184507	184125
68	Lisa	184192	182280
69	Timothy	178818	180479
70	Louis	178172	178722
71	Dale	177256	177006
72	Ray	176352	175331
73	Patrick	175915	173694
74	Nancy	174748	172094
75	Keith	172336	170531
76	Tim	171465	169002
77	Andrew	171166	167506
78	Eugene	171136	166043
79	Thos	170219	164611
80	Patricia	166399	163209
81	Dan	166332	161836
82	Randy	161222	160491
83	Carol	158758	159174
84	Eric	153862	157883
85	Russell	150534	156617
86	Lawrence	149154	155376
87	Earl	148509	154160
88	Alan	148098	152966
89	Donna	146214	151795
90	Greg	144745	150647
91	Bob	143953	149519
92	Betty	143475	148412
93	Dorothy	142869	147325
94	Lee	142288	146257
95	Norman	138089	145208
96	Jennifer	137301	144178
97	Stanley	136676	143166
98	Leonard	135123	142171
99	Helen	134813	141193
100	Ron	131421	140231

Table 6 shows the top 50 female and unisex forenames. I have attempted to include all unisex names but I regret that I do not know them all. *Mary* is number 1, *Maria,* at about a quarter of the count for *Mary,* is number 28, and *Marie* is number 42.

Table 6. Most Popular Female and Unisex Forenames.

#	FORENAME	COUNT	#	FORENAME	COUNT
1	Mary	451437	26	Ann	115059
2	Jerry	302528	27	Sandra	114435
3	Linda	250828	28	Maria	113728
4	Susan	218611	29	Diane	108878
5	Barbara	216646	30	Michelle	108739
6	Chris	214246	31	Julie	103337
7	Karen	186229	32	Shirley	103230
8	Lisa	184192	33	Laura	103091
9	Dale	177256	34	Sam	99581
10	Nancy	174748	35	Judy	98330
11	Patricia	166399	36	Brenda	98162
12	Carol	158758	37	Amy	95183
13	Donna	146214	38	Lynn	93408
14	Betty	143475	39	Kelly	91495
15	Dorothy	142869	40	Janet	91296
16	Lee	142288	41	Deborah	91092
17	Jennifer	137301	42	Marie	89140
18	Helen	134813	43	Joan	86706
19	Elizabeth	129998	44	Debbie	85446
20	Sharon	122790	45	Joyce	85337
21	Kathy	120894	46	Leslie	82806
22	Kim	119903	47	Cindy	82540
23	Margaret	119604	48	Carolyn	81154
24	Jean	116755	49	Debra	80496
25	Pat	115254	50	Lori	77653

Popular Types: Forename-Surname Pairs

Table 7 shows the 50 most common forename-surname pairs. There are no forenames used by women in this list for reasons previously mentioned. It should be noted that entries are almost exclusively Anglo-Saxon-Celtic names.

Table 8 has been extracted in sequence to include only forenames for women in the forename-surname pairs. Table 8 presents a very different picture than table 7 in that there are 5 Hispanic names: *Rodriguez, Garcia, Hernandez, Martinez,* and *Gonzalez,* all with the forename *Maria. Maria* is clearly a favorite forename for Hispanic women.

Table 7. Most Frequent Forename-Surname Pairs.

#	FORENAME	SURNAME	COUNT
1	Robert	Smith	17822
2	James	Smith	14477
3	William	Smith	13144
4	Robert	Johnson	13070
5	David	Smith	11919
6	John	Smith	11668
7	Robert	Miller	10971
8	Robert	Brown	10326
9	Robert	Jones	9922
10	Richard	Smith	9744
11	James	Johnson	9273
12	Michael	Smith	9153
13	John	Miller	8945
14	Robert	Williams	8924
15	John	Williams	8561
16	David	Johnson	8306
17	William	Johnson	8287
18	James	Brown	8114
19	James	Williams	7952
20	Charles	Smith	7694
21	William	Brown	7509
22	John	Johnson	7453
23	William	Miller	7335
24	Robert	Davis	7228
25	Robert	Anderson	7219
26	John	Davis	7006
27	James	Davis	6923
28	James	Miller	6915
29	William	Jones	6852
30	Richard	Johnson	6809
31	David	Miller	6788
32	Donald	Smith	6731
33	David	Brown	6612
34	James	Jones	6540
35	Robert	Wilson	6450
36	Robert	Taylor	6217
37	David	Jones	6119
38	John	Jones	6056
39	David	Williams	5934
40	John	Anderson	5922
41	Richard	Miller	5921
42	John	Brown	5883
43	William	Davis	5834
44	George	Smith	5716
45	John	Martin	5701
46	John	Wilson	5645
47	James	Wilson	5626
48	Michael	Johnson	5537
49	Robert	Moore	5521
50	Robert	Martin	5433

Table 8. Most Common Forename-Surname Pairs (Female and Unisex Forenames).

#	FORENAME	SURNAME	COUNT
1	Mary	Smith	4359
2	Mary	Johnson	3579
3	Jerry	Smith	3262
4	Mary	Williams	3025
5	Jerry	Johnson	2408
6	Barbara	Smith	2273
7	Mary	Miller	2105
8	Mary	Davis	2015
9	Linda	Johnson	1974
10	Susan	Smith	1909
11	Maria	Rodriguez	1884
12	Maria	Garcia	1837
13	Karen	Smith	1798
14	Lisa	Smith	1772
15	Jerry	Brown	1755
16	Jerry	Williams	1736
17	Patricia	Smith	1708
18	Barbara	Johnson	1688
19	Jerry	Miller	1641
20	Terry	Johnson	1619
21	Maria	Hernandez	1614
22	Donna	Smith	1597
23	Nancy	Smith	1548
24	Jerry	Davis	1535
25	Mary	Wilson	1532
26	Maria	Martinez	1530
27	Mary	Anderson	1509
28	Chris	Johnson	1482
29	Linda	Williams	1468
30	Chris	Smith	1405
31	Mary	Moore	1392
32	Margaret	Smith	1391
33	Maria	Gonzalez	1389
34	Mary	Thomas	1388
35	Jennifer	Smith	1386
36	Dorothy	Johnson	1386
37	Linda	Brown	1381
38	Mary	Taylor	1375
39	Susan	Johnson	1375
40	Mary	Thompson	1344
41	Karen	Johnson	1335
42	Linda	Jones	1328
43	Linda	Miller	1319
44	Lisa	Johnson	1319
45	Mary	Martin	1316
46	Sharon	Smith	1313
47	Bobby	Smith	1303
48	Barbara	Brown	1302
49	Betty	Johnson	1297
50	Barbara	Williams	1292

Conclusion

Two graphic methods of representing the forename, surname, and forename-surname pairs data culled from the U.S. telephone directory have been demonstrated. The cumulative curve method allows immediate observation of the severe skew of the three distributions, particularly forenames. One can read from the forename curve that the most frequent 0.1% of forenames represent 86% of the population. The curves each have their own shape and the U.S. shapes are similar to the Canadian shapes for the same classes.

The non-cumulative or frequency method allows the derivation for algebraic expressions, basically power law expressions, for the various name classes. What needs to be done is to determine why these curves are the shape they are and what the parameters in the algebra mean, if anything, in the world of names.

The Zipf-Mandelbrot-Simon discussion seems to have limited application to these distributions, but may spur others to provide a reasoned argument for the distributions demonstrated.

From the algebraic expressions we can calculate the sample population. The calculated results match the actual sample population reasonably well. Population can also be drawn directly from the occupied frequencies. Mirrored population counts at the high and low ends of the surname distribution remain a puzzle yet to be solved. Why is it as common to have a unique surname as it is to be called *Smith* or *Johnson*?

The lists of popular surname and forename types have few surprises except for the under-representation of women in the source data. While telephone directories have the appeal of immediacy, further study of the personal names of the U.S. must have access to data which is currently outside the public domain. Personal name research is in the public interest–from genealogy to genetics and beyond. Extracts from the public records could be made available to serious researchers with no degradation in the privacy of the people. This is perhaps an issue for the American Name Society, and other interested parties, to champion. Meanwhile, the U. S. Census Bureau is to be congratulated for its leadership in this arena.

Notes

1. The source data treats a string after a space as a new name and generally assumes that within a given sequence of names the first will be the surname and the remainder will be the forename(s). With a name string like *Kets De Vrie Manfred,* it assumes the surname is *Kets* and the forenames are *De Vrie Manfred.* This has to be repaired to surname *Kets De Vrie* and forename *Manfred.* Similarly, *Many Fingers John* is presented as surname *Many* and forenames *Fingers John.* This is repaired to surname *Many Fingers* and forename *John.* The practice of many married couples to use both their surnames also presents a problem. If Bill Smith and Mary Jones decide to use *Smith Jones* as their surname, the string will be *Smith Jones Bill and Mary,* which will be presented as surname *Smith* and forenames *Jones Bill and Mary.* This must be repaired to one entry, *Smith Jones, Bill,* and a second entry, *Smith Jones, Mary.*

2. It is common notation to use "^" to mean "raised to the power of." Hence, "y=x squared" would be written "y=x^2."

3. A power law curve is one which represents the power (logarithmic rather than linear) expression of an equation.

4. Each of these names is listed in Schwegel (1997).

References:

Hanks, Patrick, and D. Kenneth Tucker. 2000. "A Diagnostic Database of American Personal Names." *Names* 48: 59-69.

Mandelbrot, Benoit. 1959. "A Note on a Class of Skew Distribution Functions, Analysis and Critique of a Paper by H. A. Simon." *Information and Control* 2: 90-99.

Ogden, Trevor. 1998. "How Rare Are Surnames?" *The Journal of One-Name Studies* 119-124 (April).

_____. 2000. Personal Communication. December.

Schwegel, Janet. 1997. *The Baby Name Countdown.* 4th Ed. New York: Marlowe.

Simon, Herbert A. 1955. "On a Class of Skew Distribution Functions." *Biometrika* 42: 425-440.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.