

Do many of us really have rare surnames or do we just think we do?

By Donald Hatch

We all know that for many day-to-day purposes, and especially for family history research, having an uncommon or rare surname can be beneficial. Those with surnames that occur uniquely or very rarely can be more easily traced than the 100,000 John Smiths or Mary Browns. Many of us may, however, mistakenly think we have a (very) rare name when this is not the case.

My interest in this subject was awakened by an article by Trevor Ogden entitled "How rare are surnames?" in the *Journal of One-Name Studies* in April 1998, which suggested that there might be at least several hundred thousand different surnames in the UK and also that a large proportion of these occurred only once, i.e. there was only one person with this name in the UK.

In this article I will call such a unique occurrence of a surname a "hapax", after the Greek word meaning once-only and used in an expression to describe words that are found only once in all extant classical literature. More recent work by Ogden¹ suggests that there are perhaps 200,000 persons with names that only occur once in the UK and 140,000 whose names only occur twice.

Can there really be so many unique names and how can this be? Apart from immigrants, most people must surely have at least some relatives with the same name; and as soon as a male marries his name will automatically acquire a second holder. My conclusions in this article from research into 19th century names is that although a considerable proportion of all names are "hapax", the number and proportion of people with such a name is smaller than at first sight might appear.

Estimates based on small samples, or even large samples that are only a small proportion of the total population, give rise to erroneous estimates of the

frequency of rare or unique names. In addition, many apparently hapax names are transcription errors, spelling variants or homonyms for more common names. Even after correction for these factors, many remaining hapax names appear to be dubious, and very few individuals can be found to be hapax in both the 1881 and 1901 census returns.

Extinction of names – a theoretical approach

Given assumptions about longevity, and about the chances of a holder of a hapax name marrying and producing sons, it is possible to estimate the chances of a hapax name becoming extinct, remaining hapax or multiplying. Assuming, then, that our Mr or Miss Hapax is a 20-year-old orphan without relatives of the same name, marries at 25, produces two children by the time he or she is 30 and dies at 70, and this also applies to all their children, Table 1, below, indicates the likely outcome.

The table indicates that under these assumptions the chance that a hapax name will disappear altogether is high and rises over time, and that if it does not disappear it will cease to be a hapax name! For example, if our hapax is a Miss Hapax, after 30 years she will have married and lost her maiden name, so there is a 50% chance of extinction in the first generation. If it is a Mr Hapax he will also have married, but will now have a wife and two children, so that there are four persons with his "hapax" surname.

If both children are girls (25% chance) he and his wife will have died after another 30 years and both girls will have married and lost their maiden name, leading to an additional 12.5% chance of extinction. If they are a boy and a girl (50% chance) there will be still one family of four after 60 years (50% of 50% = 25%). After 120 years there is a nearly 75% chance that the name will have disappeared, and a small chance that from the original one Mr Hapax

more than 10 descendants have appeared. Eventually, the chance that the name becomes extinct levels out at 81%. Given the simple assumptions above, a real hapax cannot occur, since all the occurrences come in family units of four persons, but, of course, reality is much more complex so that

Table 1 – Rates of extinction of "hapax" surnames

Number of persons with hapax name after	30 years	60 years	90 years	120 years	150 years
0 i.e name is extinct	50.0%	62.5%	69.5%	74.2%	77.5%
4	50.0%	25.0%	15.6%	10.8%	7.6%
8		12.5%	10.9%	8.9%	14.9%*
12			3.1%	3.9%	
16			0.8%	1.6%	
20				0.4%	
24				0.2%	
28				<0.1%	
32				negl.	

* (8 or more)

(The formula used to make these computations is given in an Appendix at the end of the article)

hapax names can appear and disappear regularly. Indeed, given the rate of extinction there must be a similar rate of hapax creation.

This finding that hapax names will tend to become extinct or multiply supports the belief that the total number at any time may not be as great as some estimates based on samples suggest. To test this hypothesis, I examined some large samples of names. To avoid the problems caused by the influx in recent years of large numbers of immigrants with unusual names, I started by analysing a large sample of Victorian marriage registrations. Because even this large sample seriously underestimated the number of hapax names, it was later augmented by a 100% sample from the 1881 census.

Samples and sampling error

If you take a random sample of persons, the proportion of those with apparently hapax names is at first 100%. All the names are different but after say, 50, persons are sampled some names begin to be repeated. In this way the number of hapax names continues to rise, but they decline as a proportion of all the names found. Most of the hapax

Table 2 – Characteristics of Ha names from a sample of Free BMD marriages and the 1881 census

Name*	Percentage with main spelling*	Frequency in 1867-86 sample	1867-86 marriage sample	All names in 1853**	1867-86 marriages	1881 census ***
			Rankings		%	
Hall	100%	17387	1	19	9.0	8.5
Harris	98%	14728	2	21	7.6	7.5
Harrison	99%	13360	3	34	6.9	6.5
Harvey	99%	6171	4	87	3.2	3.2
Hart	99%	4691	5	134	2.4	2.5
Hawkins	100%	4532	6	135	2.3	2.1
Hayes	94%	4357	8	136	2.3	1.9
Harding	99%	4292	7	150	2.2	2.1
Hartley	99%	4010	9	243	2.1	1.8
Hardy	96%	3776	10	188	2.0	2.1
Haynes	63%	3731	11	>300	1.9	1.8
Hammond	99%	3529	12	194	1.8	1.7
Harper	100%	3210	13	185	1.7	1.9
Hayward	100%	3014	14	252	1.6	1.2
Hancock	100%	2968	15	257	1.5	1.3
Haigh	66%	2919	16	>300	1.5	1.6
Hargreaves	100%	2612	17	>300	1.4	1.2
Hanson	94%	2255	18	>300	1.2	0.9
Hale	95%	2079	19	>300	1.1	1.1
Hamer	69%	1829	20	>300	0.9	0.8
All others		87992			45.5	48.3
Total sample	93,6%	193442			100.0	100.0

* Including homonyms such as Haynes/Haines, Haigh/Hague, Hamer/Harmer etc.

** Excluding homonyms for England and Wales, source Registrar General

*** Including Scotland; the Scottish name Hamilton came in the 6th place with 2.2%

NB: In 1853 the most common surname, Smith, was held by 1.36% of all persons, or slightly more than 4 times as frequent as Hall

names turn out not to be so, as a second, third, etc., holder is found. It will be shown that even if a 20% sample of the population is taken, (a very large number of persons), many of the apparent hapax names are not, in fact, so.

A similar effect occurs with the words in this article: if we just take the first sentence nearly all the words are hapax, and in the whole article very many will be, but this does not mean that if all English literature is examined such a large proportion will remain so. As a test I analysed the 25,000 words in a report I had written, which suggested that of the 2,615 different words used, 1,059 (41%) occurred only once, and a further 402 (15%) only twice. But together these uncommon words represented only 7% of the total number of words. The most common 10 ("the", "of", "and", "to" etc) were good for 70% of all the words used. A similar distribution is found with surnames.

To achieve as large a proportional sample as possible I restricted my analysis to names beginning with the letters HA. All the marriages between 1867 and 1886 of brides and grooms whose surname began with these letters were extracted from the FreeBMD database as at November 2001. At that time, the coverage of marriages was about 73% – 5.7 million persons marrying, compared to the ONS

total of 7.7 for England and Wales in this period. Surnames beginning with "Ha" were, and remain, about 3.4% or 1 in 30 of all UK surnames. "Ha" does not include any of the most common surnames such as Smith or Brown but has a good scatter of names in the top 200. If we assume that the distribution of "Ha" names is similar to that of all other names, we can make estimates for the whole population of names by using this factor of ± 30 . This assumption has not been verified, and is commented on later in this paper.

Table 2 (left) summarises the

distribution of the most common names from this sample of marriages (the data from the 1881 census is discussed later in this article and added here for comparison).

The total sample amounted to 193,972 persons, but still only represented about 20% of all persons with "Ha" names at the time. A small number (340) of obvious transcription errors, spelling mistakes, and double-barrelled names were removed. Dealing with "foreign" names, which were relatively few at this period – slightly less than 1% of the births of people in the 1901 census took place abroad – was more difficult.

Foreign names

Obvious foreign names with small numbers of holders (e.g. Haberstumpf), were removed, since they were likely to be recent immigrants or temporary visitors. Hagestadt, with six occurrences, was the most common of the 190 persons with foreign names. Foreign names that occurred more often were left in the sample, since it could be assumed that the name had become indigenous, e.g. Habershon, with 17 persons.

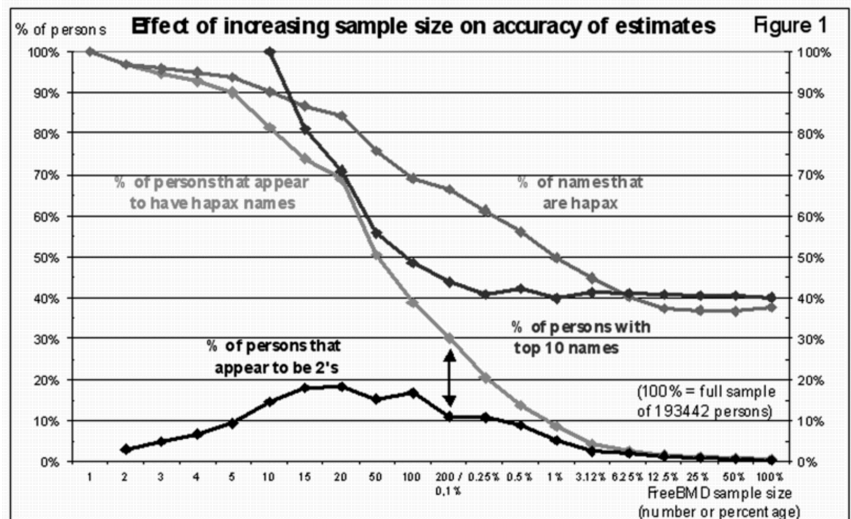
After removal of all these persons (0.3%) the sample for further analysis became the 193,442 in Table 2. Amongst these persons there were 3,775 "different" names, of which about 1,900, or half, were hapax names. This would suggest that there might have been 285,000 hapax names in the total population (1900 x 5 x 30, as the sample is only 20% of the total "Ha" names, and "Ha" names are 1/30 of the total population). This total is similar to Trevor Ogden's original estimate for modern-day Britain. However, inspection of the sample indicated that there were many spelling variations of what could realistically be considered the same name, either because of transcription errors or simply from the existence of homonyms – different spellings of the same sound.

Homonyms

The process of reducing the complete sample to account for this factor is somewhat subjective. In principle, I combined all the various spellings of surnames that were pronounced similarly – homonyms – but even here there is doubt about the pronunciation used. For example, did Haman rhyme with Harman, Hamman or Hayman? The most frequently occurring spelling variants combined into one name were Haynes (2,338 persons) and Haines (1,263 persons). Only about six per cent of all persons had a name that was regarded as a homonym

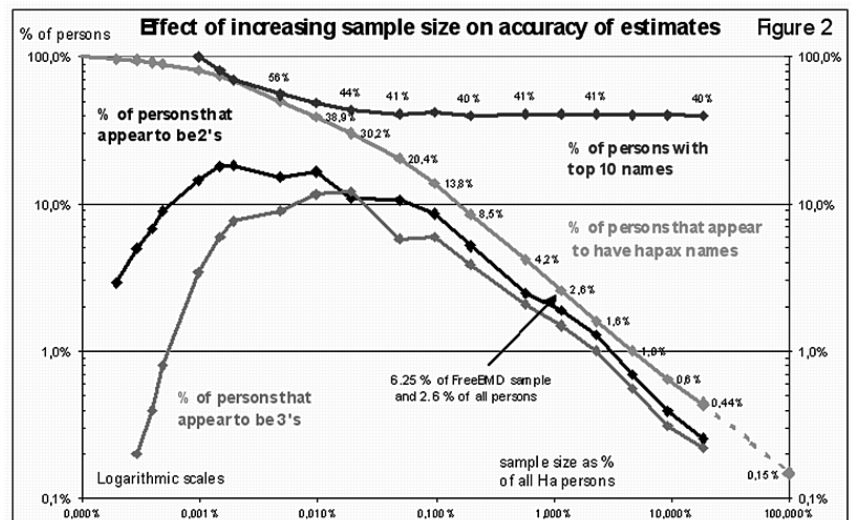
and combined with another name. The result was that the total number of names was reduced from 3,775 to 2,274, and the number of hapax names fell more sharply, from 1,900 to just 854, (since relatively many of these names were homonyms for, or spelling variants of, more common names). This meant that 38% of all names were hapax names, but only 0.44% of persons had such a name.

To estimate how many of these 854 names were, in fact, hapax names, and not just "apparent hapax" names because of the relatively small sample, two further tests were carried out. Firstly, a large number of random subsamples were taken from the



FreeBMD sample, at levels of 50%, 25%, 12.5% etc., down to samples of just one person. From these, summary statistics were calculated and graphed.

Figure 1 illustrates how a quite small sample gives a good estimate of the percentage of persons with



the top 10 names ($\pm 40\%$), but that the percentage of persons who appear to have a hapax name (or, for example, to have just one other holder of their name, a "2") declines continuously. A sample of 200, or 0.1%, of the full sample of 193,442 persons suggests that about 30% have hapax names and a further 10% 2s (marked by arrows).

Figure 2 uses logarithmic scales, which generate a straight line trend, to clarify the decline in percentage

hapax as the sample size increases (the 100% sample size in this figure refers to all "Ha" names, about five times as many as the FreeBMD sample of 193,442 persons). Extrapolation suggests that if the FreeBMD sample were increased to 100% coverage of the "Ha" names, no more than about 0.15% (or about 1,500) of all persons would have one. Applied to the total population, there would be only about 45,000 (1,500 x 30) hapax names, instead of the 200,000 suggested by earlier research.

To sum up, the sample of 193,442 persons with 854 hapax names suggests that 0.44% of all persons had a hapax name, or about 132,000 in the whole population of about 30 millions at the time (854 x 5 x 30), but extrapolation to a 100% sample reduces this total considerably to less than 50,000.

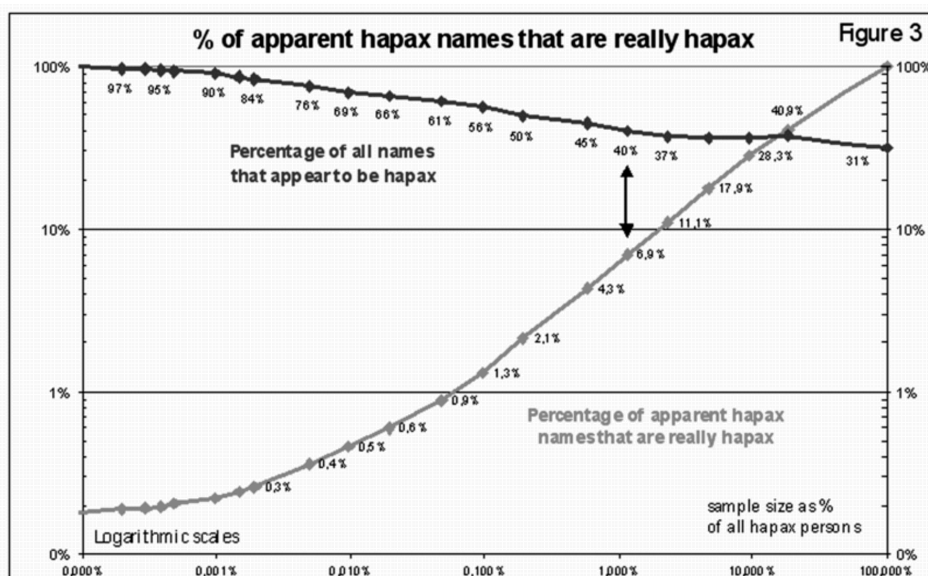
1881 census check

To confirm this estimate, a full sample of the 1881 UK census for names beginning with "Ha" was analysed. My thanks are due to Steve Archer for extracting this sample from the 1881 census. A similar process of analysis was used as for the FreeBMD sample: of the 1,041,336 persons with "Ha" names, only 2,075 were removed as foreign, and 302 as double-barrelled or clear transcription errors. The process of correcting for homonyms reduced the number of different (spellings of) names from 10,439 to 5,961. The number of hapax names fell from 3,886 to 1,876.

This is somewhat higher than the estimate of 1,500 given above; it may be related to the way names are transcribed. In the FreeBMD sample the original names in the marriage registers were possibly correctly given and spelled at the time of registration, with the errors only creeping in during

the original 19th century, and later FreeBMD, transcription processes; with the census the original enumerator may well have misspelled a considerable number of names, and there is well documented evidence that in recent times the census transcription process has also added many errors. But all transcriptions are subject to considerable errors from various sources. Table 3 gives some summary statistics from both samples.

Figure 3 illustrates how the percentage of appar-



ent hapax names that are really hapax rises sharply as the sample size increases; even a 20% sample (e.g. the 193,442 persons in the FreeBMD sample) only means that about 41% of apparent hapax names really are so.

Samples

On the assumption that there are actually 1876 real hapax names amongst "Ha" names in the 1881 census, the FreeBMD sample of 18.6% would suggest that only 18.6%, or 349, of the 854 apparent hapax names were, in fact, really so. The next step was to examine the individual hapax names in the two samples to see if this was the case.

It transpired that there were only 54, not 349, names that were hapax in both samples. Two-hundred-and-sixty-seven of the other 800 hapax names in the FreeBMD sample occurred more than once in the census, and the remaining 533 did not occur at all.

Conversely, of the 1876 hapax names in the 1881 census, 1,822 did not appear at all in the

Table 3 – Summary statistics from Ha samples

Sample:-	"Small sample"*	FreeBMD sample	1881 census
Number of persons with Ha names	12074	193442	1038959
Percentage of all Ha names	1%	18.6%	100%
Number of different names found	779	2274	5961
Persons per name, average	15.5	85.1	174.3
% with top 10 names	41%	40%	39%
Number of (apparent) hapax names	315	854	1876
% of names that are (apparent) hapax	40.4%	37.6%	31.5%
% of persons with hapax names / persons per hapax name	2.6% 38	0.44% 227	0.18% 555
% of apparent hapax names that are real hapax	7%	41%	100%

* 6.25% of the FreeBMD sample (= ± 1% of the total population of Ha names)

** see also figures 2 and 3

Table 4 – Examples of hapax names in 1881 census

Name	Note	Name	Note	Name	Note	Name	Note
Habisham		Hamblick		Haronesty		Haveeker	1
Hadwitt		Hammie		Harrups		Hawbrow	
Hainault	1	Hancatty	1	Harsomb		Hawlock	
Haitch		Hanigigton		Hartcham	1	Haxt	
Hallatson		Hanscourt		Hasdrick		Haybold	
Hallswatt		Harberer		Hasthead		Haygest	
Halsno	1	Hardslup		Hatterbey		Haynesmus	
Haltiwell		Harkard		Hauds		Hayslow	

(Note 1 indicates persons with a birthplace outside England; these were France, Norway, Ireland, N America and and Belgium respectively. In addition it can be clearly seen that some of the other names could easily be spelled differently, e.g. Habersham, Hadwit etc.)

FreeBMD sample and just 17 occurred more than once. However, the FreeBMD sample of marriages has the peculiar property that the marriage being registered immediately abolishes a hapax name, since a male hapax acquires a wife with his name and a female loses her maiden name! A further

Table 5 – Number of persons in 1901 with names that were hapax in 1881

Nr of persons in 1901 census with an 1881 hapax name	Number of occurrences	%
0 = extinct in 1901	146	73
1 = still hapax in 1901	17	9
2 occurrences	10	5
3	5	3
4	3	1
5	3	1
6	4	2
7	1	<1
8 or more*	11	5
Total sub-sample of 1881 hapax names	200	100

* One hapax name in 1881, Haxt, with no variants such as Haxed or Hackst, occurred no less than 19 times in the 1901 census

check was, therefore, carried out by examining what had happened to the 1881 census hapax names by the time of the 1901 census. An examination of the 1876 hapax names in the 1881 census suggests that many of these may well have been foreign and of recent origin, or a variant spelling of another name, in spite of the correction process carried out. Table 4 gives 32 randomly selected names as examples.

Next, a proportion of the 1881 census hapax

Table 6 – Names that were hapax in both the 1881 and 1901 census

Hablon	Halfold	Hardles
Hackard	Hamshall	Harpick
Hadred	Handor	Hatward
Hafliger	Handslow	Hawlock
Hagday	Hanscourt	Hawloy
Hainbrook	Happerton	

names were checked against the 1901 census. Two hundred of the 1876 names were examined, randomly chosen but excluding names that could have been foreign or a possible spelling variation.

Good English-sounding names such as Hallingford, Hallmore or Haperton remained.

Table 5 gives the number of holders in 1901 of these 1881 census hapax names. It appears that nearly three-quarters of the 1881 hapax names had become extinct by 1901, a similar but faster result than that predicted in Table 1.

Finally, which were these names that were hapax in both the 1881 and 1901 censuses, and were they, in fact, the same persons? The total is estimated to be about 170 (9% [Table 5] of 1,876 names) giving 5,100 (170 x 30) for the population at large. But none of the 17 persons identified in Table 5 were actually the same person in both censuses! Their names are given in Table 6.

Indeed, there was only one occurrence of any of the sample of 200 hapax persons in 1881 being found in 1901, even if the name was by then no longer hapax. This was a certain John Haytread, born Hitchin, a locksmith aged 23 living at Willenhall

in 1881, and a superintendent of an assurance company, aged 44, living at Swadlincote in 1901.

These findings suggest, as many users of the census will be aware, that these databases are by no means as accurate and comprehensive as we would like.

Conclusions

What conclusions can we draw from these analyses about the rarity or not of surnames? On the basis of the 1881 census sample, corrected for homonyms, etc., as described above, the following distribution is suggested, as shown in Table 7 on page 17.

(Please turn to page 17 for the conclusion of this article.)

(Continued from page 13)

Table 7 – Estimate of the frequency distribution of names, late 19th century

Description	Nr of "Ha" names	Nr of names in whole population*	Nr of persons in whole population* (millions)	% of population	Frequency (one occurrence in x persons)	Examples of names
Common	4	116	7.7	25.7%	more than 1 in 1000	Hall, Harris, Harrison, Harvey
Quite common	24	694	10.0	33.3%	1 in 1000 to 1 in 5000	Hartley, Hancock, Hay, Harwood
Quite rare	255	7370	9.8	32.8%	1 in 5000 to 1 in 100000	Hatch, Handford, Hamlet, Harbottle
Very rare	5678	164094	2.5	8.2%	less than 1 in 100000	Hawcroft, Hamper, Havage, Halber
Total	5961	172273	30.0	100.0%		

* Grossed up to England and Wales 1881 census population of approximately 30 millions by a factor of 28.9 (30 / 1.039 (Ha names)). From a modern sample of 55.9 million recent registrations in England and Wales the first 116 names only account for 23% of the population, with 152 names representing the most common 25.7%. The first 10 Ha names account for 1,3% of all persons in the 1881 census, and 1,2% in this modern sample. These differences may suggest that the assumption that the Ha names were representative of all names is not entirely correct, or that a modern sample has a slightly different composition to a 19th century one. See www.taliesin-arlein.net/names/ for the source of the modern data.

The estimate of about 172,000 different names applies to the situation in England and Wales in the late 19th century. Since then the population has nearly doubled and many new names have been imported by immigrants. This would suggest that more than 200,000 different names are currently to be found amongst the British population. Of the 2.5 million persons classified as having "very rare"

rare" name and less than 1% have an "extremely rare" or "unique" name, but one in three may claim to have a name classified as "quite rare". An interesting further study would be to investigate whether the distributions found here for England and Wales in the 19th century are also to be found at other times and in other countries, where the generation and history of names might be very different.

Table 8 – Estimate of the number of extremely rare and hapax names, late 19th century

	names	persons (000)	% of population
hapax names	54216	54	0.18%
occurrence = 2	19565	39	0.13%
= 3	11618	35	0.12%
= 4	10635	43	0.14%
= 5	9190	46	0.15%
= 6	7138	43	0.14%
Total 6 or less	112362	260	0.86%
Total more than 6	59911	29740	99.14%
Total	172273	30000	100.00%

names in table 7, about 10% or 260,000 had "extremely rare" (two to six holders) or hapax names, a breakdown of which is given in table 8.

Taking a slightly broader view of what qualifies as a unique name, i.e. a single family unit, we may conclude that there were about 112,000 names held by six or fewer persons at this time, comprising a total of 260,000 persons. Only 8% of us have a "very

I would like to acknowledge my thanks for constructive comment and encouragement from Philip Dance in the preparation of this article. Any errors or omissions are entirely my own responsibility. ○

DONALD HATCH
Bilthoven
Netherlands

Reference

1) To be found on Philip Dance's "Modern British surnames website at: homepages.newnet.co.uk/dance/webpjd/

Appendix – Formula for rates of extinction of surnames:

The formula for generating the rate of extinction becomes quite extensive after a few generations. After 60 years (two generations) it is $1/2 + 1/2^3$ (= 62.5%); after 90 years $1/2 + 1/2^3 + 1/2^4 + 1/2^7$ (=69.5%); after 120 years $1/2 + 1/2^3 + 1/2^4 + 1/2^5 + 1/2^6 + 1/2^8 + 1/2^9 + 1/2^{10} + 1/2^{11} + 1/2^{15}$ (=74,2%). This only applies of course under the assumption made here.