From the First to the Last, From Cradle to Grave:

An Probabilistic Analysis on the Growth and Decline of Given/Surnames

CS 109 Autumn 2020 Project

YouTube Link

(**https://tinyurl.com/y66medde**)

Edward Park

Idea/Background:

Flora and fauna aren't the only things that can go extinct. While they get all the airtime, there are other things that are on a direct course with extinction, such as last/surnames.

But how do names become extinct? Names linked to common professions - Bakers, Smiths, Fishers, Masons, Potters - have a miniscule chance of dying out, names that are linked to uncommon professions or geographical locations are dying out.

Building upon the analogy of extinction, we can classify names as endangered, on the brink, and extinct. To provide a frame of reference, we will be using the classification of last names given by myHeritage.com[1] wherein, names with less than 200 bearers are endangered, names with less than 20 bearers are on the precipice, and extinct names have no, or close to no, bearers.

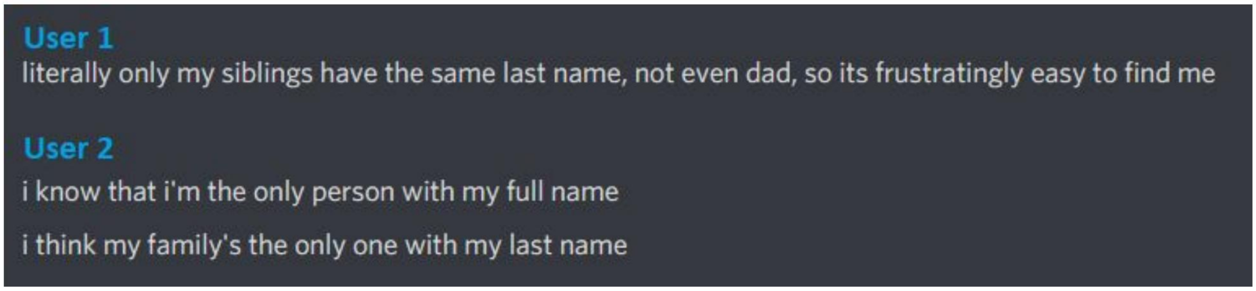
How long does it take for a name to go extinct. And furthering the extinction analogy, how long would it take for a name to be brought back from the brink.

Impetus:

The main driving force behind this project is personal curiosity stemming from late night, online conversations. As one can assume, anonymous, online conversations deep into the early hours of the morning can venture into various wild topics. During one of these conversations, the topic of having "generic" surnames was raised, to which some participants were quick to note that they had "rare" surnames - being one of a handful of people to possess said name.

[1] https://blog.myheritage.com/wp-content/uploads/2011/05/Old-British-Surnames-On-The-Brink-of-extinction_final.pdf

From there, my curiosity peaked. How long would such a name survive; how long would it take to reach the average, if at all; at what point would the name be wiped from memory. I simply had to know.



Going off the words of these users, their names are already classified as extinct, or at least, exceedingly close. Let's see if we can't use probabilistic analysis to see how long their names will last.

Design:

The general design for this analysis is deceptively simple. Boiled down to the base parts, it is population analysis. The difficulty lies within accurately modelling the population's growth and decline. This is made more difficult as we are not tracking the population in and of itself, but rather a trait of the population.

For the sake of simplicity, we are going to assume that males are the ones to pass down surnames, and will always do so if given the chance.

Now to get on with the modelling. For this model we will have to first model the growth of the population, then simulate how long it takes for said population to subsequently decline.

To model these simulations, I will be making use of the Galton-Watson process, a process that models how long it takes for the last male of a given population to pass on. This process models family names as being passed from father to son, while children are randomly either male or female, with names becoming extinct if the family name line truly dies out - either no more male descendents of the family to carry on the name.

While I know the rough population for the males in Users 1 and 2's families, it is not fun to simply have proportions for them, therefore, I made use of data compiled by Statista, namely their data on the proportions of the US population by age and gender[2], to create a list of male population proportions based on age ranges so that we can simulate any given surname, not just those of Users 1 and 2.

[2] https://www.statista.com/statistics/241488/population-of-the-us-by-sex-and-age/

From the provided data, we are now able to draw an approximation of the distribution of age among males in the US, which is as follows:

Under 5	~3.15% of the total population
5 to 9	~3.24% of the total population
10 to 14	~3.34% of the total population
15 to 19	~3.38% of the total population
20 to 24	~3.48% of the total population
25 to 29	~3.77% of the total population
30 to 34	~3.57% of the total population
35 to 39	~3.42% of the total population
40 to 44	~3.11% of the total population
45 to 49	~3.17% of the total population
50 to 54	~3.17% of the total population
55 to 59	~3.34% of the total population
60 to 64	~3.10% of the total population
65 to 69	~2.58% of the total population
70 to 74	~2.04% of the total population
75 to 79	~1.36% of the total population
80 to 84	~0.8% of the total population
85+	~0.7% of the total population

Beyond making a US male population proportion table for fun, going back into the general means of the Galton-Watson process, we need to make several assumptions. Specifically, we need to assume that family sizes are independent of each other and can take on values of *0* to *INF*. Additionally, the families are assumed to have independent identical distribution probabilities (IID) and that the probability that the number of children, *C*, born to a single parent is equal to some number *k* is written as $P(C = k)$ where $k = 0, 1, 2, [...]$. Therefore, we can use the number of children born in each generation to determine:

X_n = number of children born in generation n
(size of the n-th generation)

For creating the probability distribution, we will be making use of a Poisson Distribution. The Poisson Distribution is used as a descriptor of the number of children per parent, making use of the Poisson Distribution's property of lambda being the expectation, and in this case, the lambda would be the average number of kids per family. In this case, the average birth rate in the US is roughly 1.93 children per family[3], $Poi(\lambda = 1.93)$.

However, using a Poisson is not the end all be all, we must also make use of a Bernoulli Random Variable to see if a child is born to a given person at all. While data for *Births per 1000 Males* is not readily available - the closest being a line graph from 2002[4] - data for *Births per 1000 Females* is readily available. While not the most ideal, as we are tracking male populations, the birth rate data will, regardless, allow us to determine whether or not a given person has a child using: $Ber(p = \text{Birth Rate Data})$.

The *Births per 100 Females* data[5] was compiled by the CDC in 2018, and from said data we can glean the following data[*], wherein the birth rate per age range is divided by 1000:

10 to 14	~0.0002
15 to 19	~0.0174
20 to 24	~0.0679
25 to 29	~0.0952
30 to 34	~0.0996
35 to 39	~0.0526
40 to 44	~0.0118
45 to 49[**]	~0.00045
50 to 54[**]	~0.00045

[*]For age ranges not represented, it is assumed that the birth rate is 0
[**]For age ranges [45, 49] and [50, 54], the data presented gives them as a combined total. They were equally redistributed for the purposes of this analysis.

[3]https://www.statista.com/statistics/718084/average-number-of-own-children-per-family/#:~:text=The%20typical%20American%20picture%20of,18%20per%20family%20in%201960.&text=If%20there's%20one%20thing%20the,is%20known%20for%2C%20it's%20diversity.
[4] https://medium.com/@Episona/ages-impact-on-male-fertility-4f0ddc521e55
[5] https://www.cdc.gov/nchs/data/vsrr/vsrr-007-508.pdf

Now that we have the birth rates per 1000 women per age range, we can create a Bernoulli Distribution of whether a given person has offspring.

With these variables now accounted for, we can now simulate how long a given surname will last. While the length of the simulation can vary, however, what will not vary is the processes that will occur in the simulation.

Simulation:

The simulation of this problem is fairly straightforward, we simply have a starting population, divide that starting population into the age groups listed above, and then for each age group, check the fertility rate, then check if a given person has a child - *Ber(p = Birth Rate Data)* - and if the person has a child check how many - $Poi(\lambda = 1.93)$ - then add all the children of a generation together - X_n - then remove the oldest generation, and add half of the new generation - we add half as we are interested in the population of males and the birth ratio of males to females is 105 males:100 females, which for the sake of simplicity, we simplify to 1:1. After this we then should check if the population is extinct: check if the sum total of the population is zero. If it is zero we should note at which point the population went extinct and resimulate, otherwise continue on to the next generation until extinction.

After the simulations are complete, we should find the sample mean of generations until extinction, adding all the generations at which the population went extinct and dividing by the number of simulations, giving us the average number of generations until the population bites the dust. Alternatively, we can multiply the sample mean by five, as we have defined generations as age ranges of five years, to find the number of years until the population goes into the long night.

Results:

I was truly surprised by the results of this Galton-Watson simulation/analysis. User 1's surname had a starting sample size of two, however, it managed to survive roughly 100 years, while User 2's surname lasted a little while longer. I would have assumed that their names would have disappeared much earlier than what the simulations returned, as their names, for better or worse, were already considered "extinct."

Future:

While this project proved to be both enjoyable and engaging, there are still a couple of things that could have been better. Firstly, as noted above, it would have been greatly beneficial to know the *Births per 1000 Men* and used that as our p for the Bernoulli Random Variable, as we are interested solely in the male population for the purposes of this process. Secondly, an issue arises in the simulation, as the simulation will have males possibly reproduce every five years, which is not in the least bit realistic. A later refinement of this process should take into account how many males already had children in the previous run and give them an adapted probability of having more offspring after five years. The simulation also does not take into account deaths outside of the removal via the final age range, so future implementations should also look into removing some proportion of the population before the age range removal. And in a similar vein, future implementations should also take into account those who willingly choose not to have children or are otherwise unable to have children, essentially removing them from the population of interest.