Universitat Autònoma de Barcelona

Applied Stochastic Processes

Applications: The Galton-Watson Process

Miquel Barcelona Poza (1359817)
Daniel López Badell (1561011)

January 2020

Contents

1 Introduction

Studies on population dynamics processes is one of the most studied cases throughout contemporary history. In this project we will work on a branching stochastic process aired in the 19th century and known as Galton-Watson process. At that time there were several aristocratic families that realized that their family names could become extinct. In front of such a problem, investigator Sir Francis Galton started to develop a process that could answer how many male children must each generation have on average in order to avoid the extinction of a family name. Nowadays this model is very useful in genetics to study Y chromosome propagation along generations and many other propgation issues. The purpose of this work is to study this model and its properties and finally to obtain the results for two different practical cases for our own surnames.

2 The Galton-Watson process

In this section we present the Galton-Watson process and its main definitions and properties. These are going to be very useful in the following sections in order to understand in a better way the obtained results and the used methods. The Galton-Watson process is a branching process that models population to describe the evolution of a family name among generations. As we have exposed before, the main objective of this model is to study extinction probabilities. Consider a number of individuals in a population that has not external influences and evolves according to the following rules: Considering generation $n = 0, 1, 2, ...$ each individual produces a random number i of descendants (it considers that only male's descendants keep family name). This process can be described graphically by the Figure 1. The offspring distribution is the PDF that describes the number of descendants per individual

$$p_i = P(\#\quad descendants = i), \quad for \quad i = 0, 1, 2, ...$$

To study the propagation of a family name it takes an initial amount of individuals Z_0 and it generates a set of random family trees that can be finite, if some generation has not any descendant, or infinite, if the family name never dies out. We can read this process as a Markov chain where Z_n are the number of individuals at nth generation.

Notice that if population decreases to 0 in certain n generation ($Z_n = 0$) then $Z_m = 0$ for all $m = n, n + 1, n + 2, ...$, so 0 is an absorbing state. We will define the following probabilities on extinction:

$$\delta_n = P(Z_n = 0)$$

$$\pi_0 = P(process \quad dies \quad out) = \lim_{n \to -\infty} P(Z_n = 0) = \lim_{n \to -\infty} \delta_n.$$

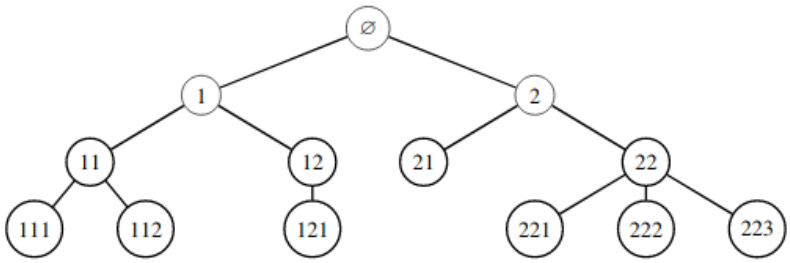


Figure 1: Scheme of the behaviour of a classical Galton-Watson process. Image extracted from [1].

Our main objective will be to compute π_0, that will depend on offspring distribution and the initial number of individuals Z_0. Now we will study the offspring distribution behaviour along generations and how impacts on extinction probabilities. For that it is necessary to set notations on some concepts, assuming that $Z_0 = 1$

$$\mu = E(Z_1), \sigma^2 = Var(Z_1)$$

$$m_n = E(Z_n), v_n == Var(Z_n).$$

By intrinsic properties of Markov chains studied in class, we know the following behaviour of this statistics

$$m_n = \mu^n,$$

$$v_n = \begin{cases} \frac{\sigma^2}{\mu(1-\mu)}\mu^n(1-\mu^n), & if \quad \mu \neq 1, \\ n\sigma^2, & \text{otherwise.} \end{cases}$$

In order to look for a relation between π_0 and offspring distribution moments, we study its generating function $\sigma(s)$ with $s \in [0, 1]$. Analogously, we define the generating function of X_n $\sigma_{X_n}(s)$, that can be expressed as the nth iterate of $\sigma(s)$. As we can see in Theorem 14.2 in [2], there is the following relation between offspring distribution and extinction probability.

If $\mu \leq 1, \pi_0 = 1$; otherwise, π_0 is the smallest solution on $[0, 1]$ to $s = \sigma(s)$.

This remarkable result will allow us to calculate extinction probability based on descendant's distribution and initial population Z_0. First we calculate $\widehat{\pi}_0$ for a unique individual with the theorem and distribution, then $\pi_0 = \widehat{\pi}_0^{Z_0}$.

3 Code explanation

In the following section we deal with the main part of the code in charge of generate the particular number of simulations for a given branching process. Due to our purposes, we define a function called simulation that receives as parameters the number of the last generation g that we want to compute, the initial number of population for the zero generation and the probability parameter p for a given distribution. For instance, for a Poisson process p will be the well-known λ parameter of the distribution function $f(k) = \lambda^k \exp -\lambda/k!$. In addition, we set the number of times n that we want to execute our simulation in order to obtain consistent results. Those are the most important values and our future results will depend basically on them.

The `simulation` function starts by declaring a matrix with $g+1$ rows and n columns. This matrix will contain all the generated information by our simulation process. In this case, on one hand each column of the matrix will define a single branch simulation process. And on the other hand, each row of the matrix will give us the values of a certain generation for the different simulations. All of these allows us to perform an easier analysis for the different results.

Once the matrix is well defined, a set of names are declared to make the matrix better identifiable. This is the only variable that we need to declare and this is the one which is going to be returned at the end of the function. Therefore, our computational time will depend directly on the size of this matrix. Finally, we need to initialize the first row of the matrix, the zero generation, with the input value for the initial population z_0.

At this moment, we are ready to run the simulation itself. For that we generate to different loops: the outside one is the loop that travels through the different generations and is identified by the variable i and the inside one generates every simulation value for a given generation, j. As it was said before, in every generation step i only those simulations which are not extinguished are able to have offspring, that is why we need to check if the previous value from the generation $i-1$ and same simulation j was greater than 0. In that case, the next generation value is computed using the given probability distribution. We will take later about how we compute this probability value.

At the end of the loops, the generation matrix is filled and it can be returned. The corresponding code to this function is presented below.

```
1 # Applied Stochastic Processes
2
3 # n simulation Galton-Watson Process
4 # p probability parameter and g generations
5 # z0 initial population
6 simulation = function(n,p,g,z0) { #n nombre de simulacions
```

```
  # defining a n x (g + 1) matrix
  generation = matrix(data=NA , nrow = g + 1 , ncol = n , byrow=TRUE)
  namesim = c(1:n)
  for (i in namesim) namesim[i]=paste("s",i)
  namegen = c(0:g)
  for (i in namegen) namegen[i + 1]=paste("Z",i)
  dimnames(generation) = list(namegen,namesim)
  # set initial condition
  generation[1,] = z0
  # until generation g
  for (i in (1:g)) {
    # for the n simulations
    for (j in (1:n)) {
      if (generation[i,j]>0) {
        generation[i+1,j]=rnbinom(1,generation[i,j],p)
      }
      else{
        generation[i+1,j]=0
      }
    }
  }
  generation
}
```

3.1 Probability distributions

Here below we present the procedures that are used to compute the number of individuals of a next generation depending on the previous one. In this assignment, only three probability distributions are used. As we have said before, the number of individuals of the generation n is given by the sum over all the children born from all the individuals of the previous generation, that means

$$Z_n = \sum_{n=1}^{Z_{n-1}} Z_{n,i}, \tag{1}$$

where $Z_{n,i}$ are the number of children from the generation n that are the offspring of the individual i of the generation $n-1$.

Using the equation 1 in our function would reduce the performance speed because we would need

to define another different loop to generate k random numbers if k was the number of the previous generation. Instead of doing this, we take advantage of some probabilistic properties related to the used distributions.

Poisson Process If the number of children for a given individual $Y_{i,j} \sim P(Z_{i,j} = k)$ follows a Poisson distribution with parameter λ, then the sum over all the k individuals will be,

$$X_i = \sum_{j=1}^{k} Y_{i,j} \sim Pois(k\lambda). \tag{2}$$

Then, the next generation can be calculated without performing any additional loop and only taking the right parameter [3]. Also we can compute the extinction probability π_0 as we have seen in section 2. If $\lambda \leq 1$ then $\pi_0 = 1$ but if $\lambda > 1$, then $\widehat{\pi}_0 = \{s \in [0,1] | \frac{ln(s)}{1-s} = -\lambda\}$ and $\pi_0 = \widehat{\pi}_0^{Z_0}$.

Binomial distribution In this case, the number $Y_{i,j}$ follows a binomial distribution with parameters $B(n,p)$. Then, the total number of children for the generation i will be

$$X_i = \sum_{j=1}^{k} Y_{i,j} \sim B(kn,p). \tag{3}$$

And again this value has to be calculated just once for every simulation [3]. Analogously, we calculate probability of extinction. If $E[X_i] = np \leq 1$ then $\pi_0 = 1$, but otherwise then $\widehat{\pi}_0 = \{s \in [0,1] | s = (1 + p(s-1))^n\}$ and $\pi_0 = \widehat{\pi}_0^{Z_0}$.

Geometric distribution Assuming that the offspring distribution follows a geometric distribution with parameter p, then the sum over all the children of the previous generation will be

$$X_i = \sum_{j=1}^{k} Y_{i,j} \sim NB(k,p) \tag{4}$$

a negative binomial with parameters k the number of individuals in the previous generation and p the parameter that points probability of success [3]. Expected value of this distribution is $\mu = \frac{p}{1-p}$ so if $p \leq \frac{1}{2}$ then $\pi_0 = 1$. Otherwise we calculate analogously the probability of extinction and we obtain that $\pi_0 = \left(\frac{1-p}{p}\right)^{Z_0}$.

Distribution	p_1	p_2	p_3	p_4
	0.35	0.45	0.65	0.75
Geometric simulation	0.531	0.816	0.999	1.000
Expected value	0.538	0.818	1.000	1.000
	0.6	0.5	0.4	0.3
Binomial simulation	0.099	0.233	0.553	0.966
Expected value	0.096	0.236	0.557	1.000
	0.8	0.9	1.0	1.1
Poisson simulation	0.995	0.976	0.911	0.804
Expected value	1.000	1.000	1.000	0.824

Table 1: Comparison between the computed and the theoretical probabilities of a branching process to die out depending on the probability parameter for different distributions.

4 Study of the process

In this section we will analyze some results regarding the three distributions exposed in previous section in order to validate if their behaviour follows the mentioned theoretical results. For this analysis we have done 10000 simulations with $Z_0 = 1$ constraint and 4 different p probabilities of success per each distribution. In Figure 2, the probability of extinction for the next 20 generations is exposed in each case.

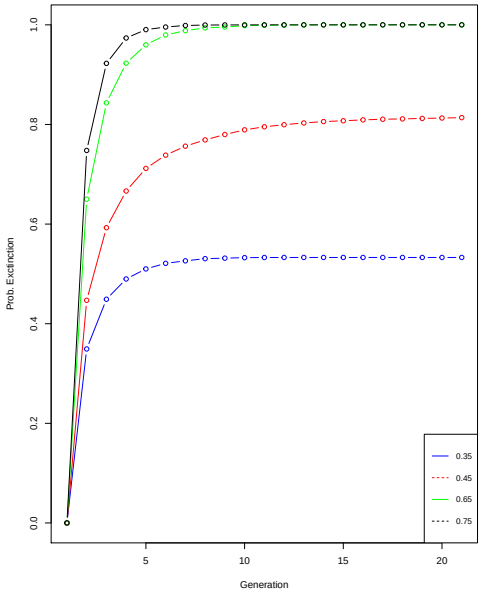
Finally, we compare our practical results obtained by the `simulations` function with the theoretical values. As we can observe in the Table 1, the results obtained follows the laws described previously. They are also displayed the values of p used to generate each simulation for the different distributions.

5 Practical case

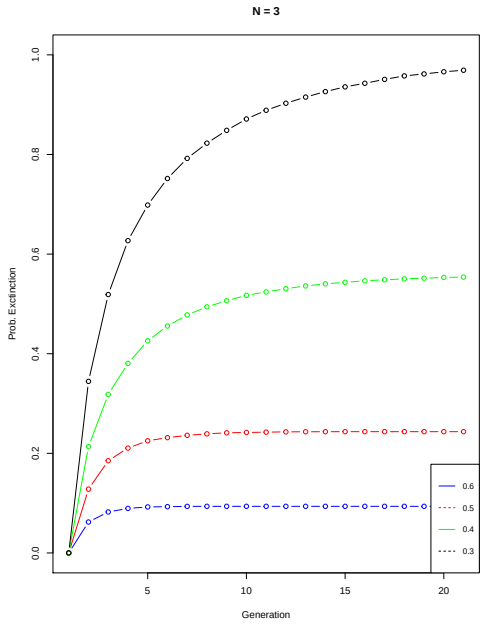
In the following section we extract some data from the INE concerning the number of individuals in Spain with the surnames Barcelona and Lopez [4]. The purpose of this study consists on analyzing the evolution and the survival of these surnames depending on different parameters.

As we are dealing with a branching process, we only consider the male individuals with their first surname satisfying our conditions. These are the ones that can contribute to the survival of the names. For that reason, it makes sense to assume that the half of the population with a given surname is male. Hence, the number of individuals for the zero generation Z_0 with the first surname Barcelona is 640 while those men with Lopez as first surname are 434730.

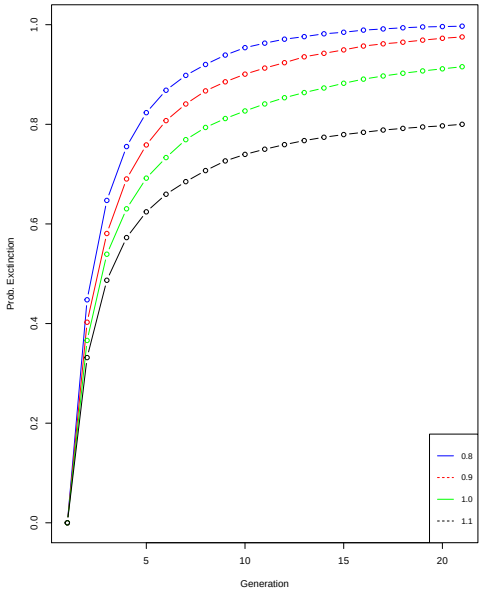
At this moment, we need to select which is the probability distribution that models the offspring of a single individual at a given generation. This is a tedious and very complicated task that would be another project by itself. However, in these kind of problems is very common to use a geometric



(a) Geometric distribution.



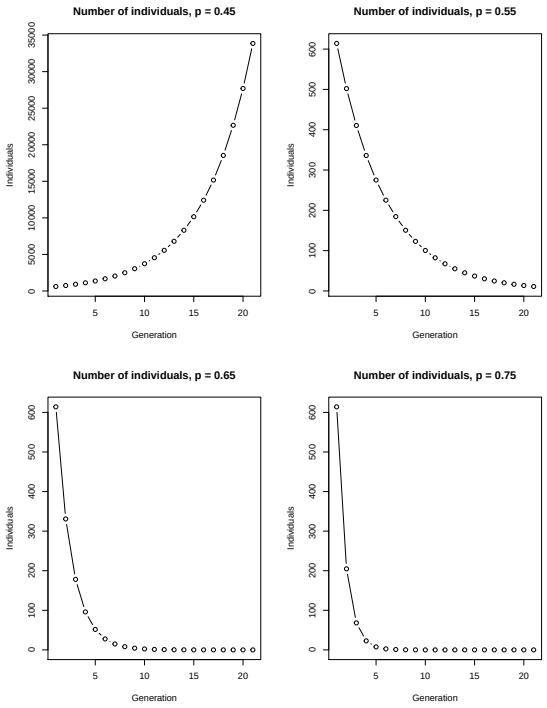
(b) Binomial distribution.



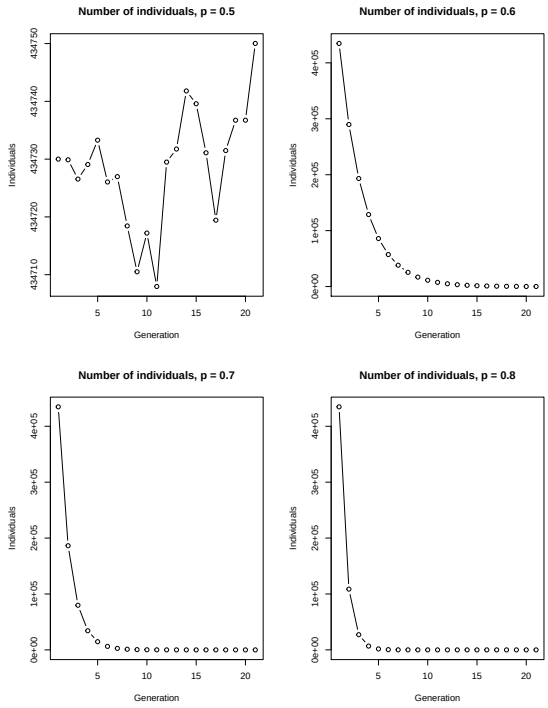
(c) Poisson distribution.

Figure 2: Probability of extinction for a Galton-Watson process following different probability distributions with different parameters and 1 initial individual.

distribution with a certain probability factor p. This distribution was firstly used by Lotka in 1920 [5]. By studying the census data and some other sociological parameters he estimated the parameters of the distribution. In that case, we will compare the different results for arbitrary



(a) Surname Barcelona.



(b) Surname Lopez.

Figure 3: Evolution of the number of individuals during the 20 next generations with the surnames changing the parameter p.

probabilistic parameters. Then, the offspring distribution will be

$$P(Z_{i,j} = k) = (1-p)^k p, \quad k = 0, 1, \ldots, \tag{5}$$

meaning that the probability of not having offspring for a given individual is p.

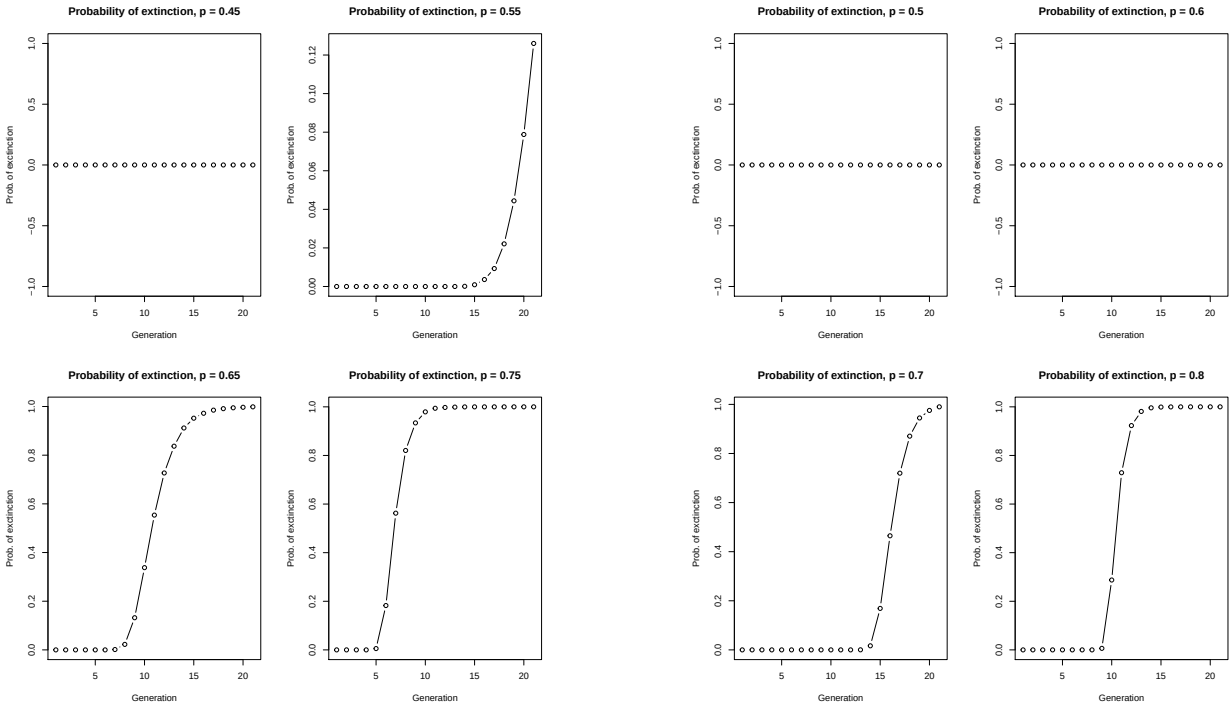
In this way, due tot the different initial number of individuals of both surnames, we will select different p values for each one of the cases in order to obtain significant results. For the Barcelona surname we will use

$$p = (0.45, 0.55, 0.65, 0.75),$$

while for the Lopez surname, p will be

$$p = (0.5, 0.6, 0.7, 0.8).$$

Moreover, to use the function `simulation`, we choose to produce 10000 simulations for the next 20 generations. With all of these values, we obtain the different evolution of the individual mean for each one of the two surnames. These results are show in Figure 3. As we can see, both surnames tends to the extinction when the parameter p is larger than $1/2$. The population will increase at



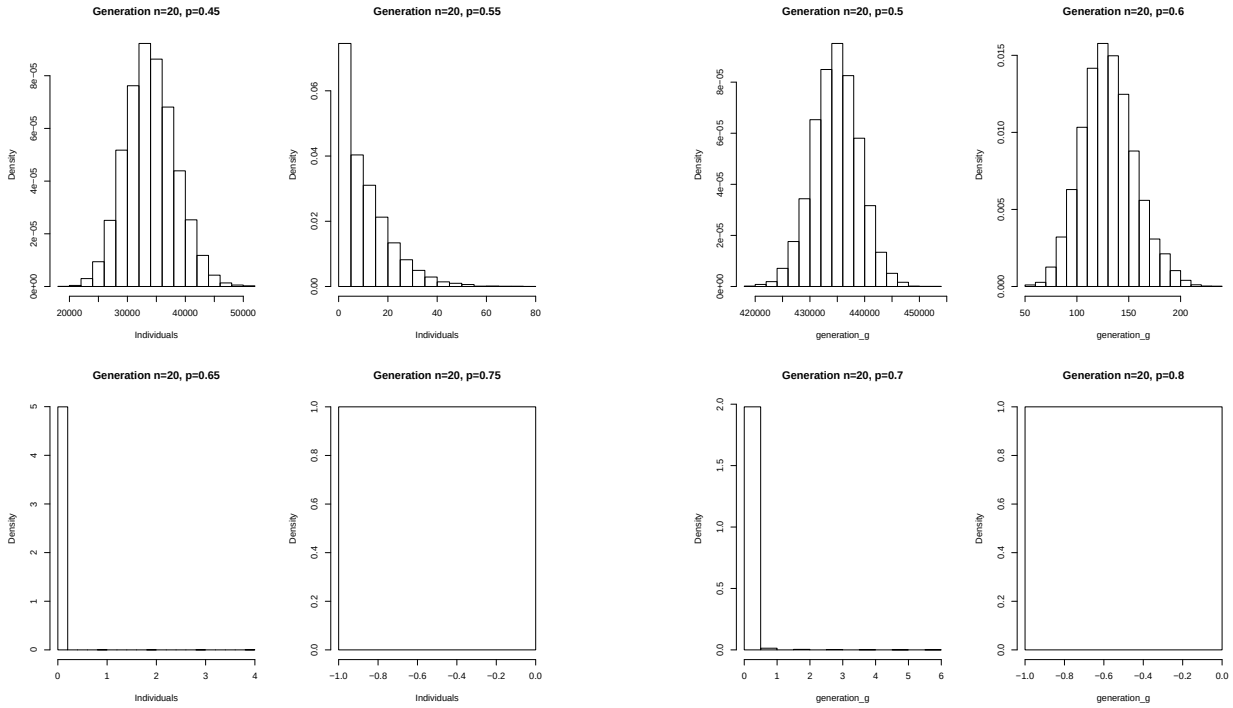
(a) Surname Barcelona. **(b)** Surname López.

Figure 4: Probability of extinction of the surnames during the next 20 generations for different parameters p.

different velocities depending on how large is this value. Nevertheless, when it is smaller or equal than this critical value, the population will survive.

Then, for the same parameters we compute the probabilities of extinction for both surnames and we present the results in Figures 4. In this case, due to the huge number of individuals in the first generation, the probabilities of extinction are 0 when the parameter p is smaller or equal than $1/2$ even though in the first generations. This result differs from what we saw in section §4 when we simulated a process with a geometric distribution for 1 initial individual. This behaviour is even more significant for the Lopez surname because of the great number of the initial sample. But, as we increase the value of p, the population increases its probability of extinction.

Finally, we draw some histograms showing the number of individuals for the different simulations at the generation number 20. These are displayed in Figure 5. Again, the sample will vary depending on the parameter p.



(a) Surname Barcelona.

(b) Surname López.

Figure 5: Histogram of the number of individuals with the surnames for the 20th generation with different parameters p.

6 Conclusions

In the results of the practical we have seen the huge impact of varying offspring distribution parameters on population's extinction, as theory points in section 2. To conclude, it is true that this model does not take in account important factors that could not affect on 19th century aristocracy but now they do, such as adoption or change of family names, but nowadays it still can be useful in other fields like genetics or computer virus propagation.

References

[1] "The simple Galton-Watson process: Classical approach". In: (2011), pp. 1–21.

[2] P. K. S., K. B. Athreya, and P. E. Ney. "Branching Processes." In: *Journal of the American Statistical Association* 69.345 (1974), p. 282. ISSN: 01621459.

[3] Marta Sanz-Solé. *Probabilitats*. Edicions Universitat de Barcelona, 1999.

[4] *Instituto Nacional de Estadística. (Spanish Statistical Office)*.

[5] F.W. Roush. "Branching processes". In: *Mathematical Social Sciences* 7.3 (1984), pp. 300–301. ISSN: 01654896.