CASE STUDY: EXTINCTION OF THE LINE OF DESCENDANTS (IN PARTICULAR, FAMILY NAMES)

The idea that family names die out originated in antiquity, particularly since the establishment of patrilineality (a common kinship system in which an individual's family membership derives from his/her father's lineage.) As offsprings of each generation can be either male or female and only the males carry on family names, the family name becomes extinct if male descendants die out. This is also related to the Y chromosome transmission in genetics. Hence the problem of the extinction of family names has been of great interests to the studies of demographics, anthropology and genetics.

Chinese names are well-known examples of family name extinction. There are currently only about 3,100 surnames in use, in contrast with close to 12,000 recorded in the ancient literature. Within the existing surnames, 22% of the population share three most common family names and the top 200 names cover 96% of the whole population. According to Du et al. (1992), "the Chinese population uses fewer surnames and includes much larger isonymous groups than Caucasoids or Japanese because surnames appeared in China at least 3000 years earlier than in Europe or Japan. Since that time, when the population was much smaller than now, many surnames have become extinct."

While there are many other factors that affect the survival of family names in reality, we shall only consider the simplest model which was first independently studied by Bienaymé (1845) and Galton and Watson (1874). The model, known as the Galton-Watson-Bienaymé branching process, provides an elegant description of how family name can die out by nature.

Similarly to a family name extinction (which is the probability that the line of male descendants gets extinct), one can look, more generally, at the line of all descendants getting extinct.

The purpose of this case study is to give a brief introduction to the Galton-Watson-Bienaymé branching process and develop a numerical example for calculating the probability of a line of descent extinction.

Suppose that we know from statistical analysis the probabilities that an individual has $0, 1, 2, \cdots$ children and that each them has the same probabilities of children of his/her own, and so on. There are two questions that of essence to this problem.

(1) What is the probability of a given number of descendants in any given generation?
(2) What is the probability of eventual extinction of the descendants?

1. Background: Branching Process

Consider descendants of one particular individual. Let

$Z_n =$ the number of descendants in the n-th generation, $\qquad n = 0, 1, 2, \cdots$.

Then $Z_0 = 1$. Let X be a generic random variable standing for the number of children of an individual and p_k be the probability that an individual has k childrens, i.e. $\mathbb{P}(X = k) = p_k$ for every $k = 0, 1, 2, \cdots$. A very important mathematical device we need for this problem is the probability generating function (pgf), $P(s)$, defined by

$$P(s) = P_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} p_k s^k. \tag{1}$$

We shall write

$$P_n(s) = \text{ the pgf of } Z_n, \qquad n = 0, 1, 2, \cdots .$$

For notational brevity, we shall also write

$m = \mathbb{E}(X) = P'(1) =$ the mean of the number of children of a given individual,

and

$$\begin{aligned} \sigma^2 = \mathbb{V}(X) &= P''(1) + P'(1) - (P'(1))^2 \\ &= \text{ the variance of the number of children.} \end{aligned}$$

The stochastic process $\{Z_n, n \geq 0\}$ is called the Galton-Watson-Bienaymé branching process.

Let us first observe some nice useful properties of the probability generating function. (All random variables considered here have values in the non-negative integers.)

Exercise 1.1. (1) If X and Y are independent, then show that

$$P_{X+Y}(s) = P_X(s)P_Y(s).$$

More generally, if $X_1, X_2, \cdots, X_n$ are $n \geq 1$ independent random variables, and if $S_n = X_1 + X_2 + \cdots + X_n$, then

$$P_{S_n}(s) = P_{X_1}(s)P_{X_2}(s) \cdots P_{X_n}(s).$$

(2) Let $X_1, X_2, \cdots$ be i.i.d. random variables, and N be a random variable independent of the X_i's. Let the random variable $S_N = \sum_{k=1}^{N} X_k$. Then show that

$$P_{S_N}(s) = P_N(P_X(s)),$$

where $P_X(s)$ stands for the pgf of each of the X_i's.

□

We carefully examine how the n-th generation carry forward to the $n+1$-st generation. Let us label the descendants in the n-th generation by $1, 2, 3, \cdots, Z_n$ and let X_i be the number of children of the descendant with the label i. Then the total number of descendants in the $n+1$-st generation would be

$$Z_{n+1} = X_1 + X_2 + \cdots + X_{Z_n}.$$

By assumption, the X_i's are independent of each other and of Z_n. Hence by Exercise 1.1, we have the fundamental equation

$$P_{n+1}(s) = P_n(P(s)), \qquad n = 0, 1, 2, \cdots. \tag{2}$$

Exercise 1.2. Show that

$$P_{n+1}(s) = P(P_n(s)), \qquad n = 0, 1, 2, \cdots. \tag{3}$$

□

We can use the fundamental equation (2) to calculate moments of Z_n for any n.

Exercise 1.3. Show that:

(1) If $m < \infty$, then $\mathbb{E}(Z_n) = m^n$ for all $n \geq 0$.
(2) If $\sigma^2 < \infty$, then

$$\mathbb{V}(Z_n) = \begin{cases} \frac{\sigma^2 m^n (m^n - 1)}{m^2 - m}, & \text{if } m \neq 1, \\ n\sigma^2, & \text{if } m = 1. \end{cases}$$

(Hint: differentiate the fundamental equation (2) to find recursive formulas for moments.) □

2. Background: Probability of eventual extinction

Now we turn to the probability of eventual extinction.

Consider first the extreme cases. If $p_0 = 0$, then every generation has at least one descendant for the next generation and the branching process will never die out. Hence the probability of eventual extinction is zero. If furthermore, $p_1 = 1$, there will be exactly one descendent every generation, while if $p_1 < 1$, we will have $Z_n \to \infty$ as $n \to \infty$ (why? consider Exercise 1.3). The only uncertain case is when $p_0 > 0$ in which case we may have $Z_n = 0$ for some n. Assume $p_0 > 0$. Note that once the number of descendents reaches zero, there would be no chance of reproduction and hence the process would have died out. Therefore, the event $\{Z_n = 0\}$ is a subset of the event $\{Z_{n+1} = 0\}$ for all n. We shall write the event of extinction as

$$\{\text{extinction}\} = \{Z_n = 0, \text{ for some } n \geq 1\} = \bigcup_{n=1}^{\infty} \{Z_n = 0\}.$$

By what is known as the "continuity" property of probabilities, we then have for $\xi = \mathbb{P}(\text{extinction})$:

$$\xi = \mathbb{P}(\text{extinction}) = \mathbb{P}\left(\lim_{n\to\infty} \bigcup_{k=1}^{n} \{Z_k = 0\} \right) = \mathbb{P}\left(\lim_{n\to\infty} \{Z_n = 0\} \right) = \lim_{n\to\infty} \mathbb{P}\left(\{Z_n = 0\} \right).$$

Note that the probabilities $\mathbb{P}(Z_n = 0)$ are getting larger as n increases.

Let us write $x_n = \mathbb{P}_n(0) = \mathbb{P}(Z_n = 0)$ for $n \geq 0$. By the fundamental equation (3), we have for all $n \geq 0$,

$$x_{n+1} = P_{n+1}(0) = P(P_n(0)) = P(x_n).$$

Using this fact and the continuity of the pgf, we get

$$\xi = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} P(x_n) = P(\lim_{n\to\infty} x_n) = P(\xi).$$

This shows that the probability of eventual extinction ξ has to be a root of the equation $P(s) = s$. However, be careful that the root to the equation may not be unique! We can prove the following:

The probability of eventual extinction ξ is the smallest root of the equation $P(s) = s$ in the interval $[0, 1]$.

Indeed, suppose ρ is another nonnegative root of $P(s) = s$. Recall that $x_0 = 0$ so $x_0 \leq \rho$. Since $P(s)$ is an increasing function of s for $s \geq 0$, we find

$$x_1 = P(x_0) \leq P(\rho) = \rho.$$

Repeating the argument, we get

$$x_2 = P(x_1) \leq P(\rho) = \rho,$$

and so on, finding that $x_n \leq \rho$ for every n. But then $\xi = \lim_{n\to\infty} x_n \leq \rho$ and so ξ is indeed the smallest such root.

Exercise 2.1. Recall that $m = \mathbb{E}(X) = P'(1)$ is the mean number of first-generation descendants (children) of one individual. Show that

(1) if $m \leq 1$ then $\xi = 1$;
(2) if $m > 1$, then $\xi < 1$.

(Hint: draw a picture and recall that $p_0 > 0$ and $P(1) = 1$.) □

Exercise 2.2. Let us $P(s)$ be the generation function of the number of children of one individual. Let ϕ be the probability that a child is a female. Assume that the genders of children are independent. Using Exercise 1.1(2) show that the generating function of the number of daughters of one individual is

$$\hat{P}(s) = P((1-\phi) + \phi s).$$

□

3. Your assignment

Table 1 provides the statistics of the number of children per woman in China and Mexico: N_k is the number of women polled who have k children. (Assume, it is at most 5 children.) So, the total number of women polled is

$$N = \sum_{k=1}^{5} N_k.$$

The statistical estimate of the distribution of the number of children per woman is

$$p_k = N_k/N, \quad k = 0, 1, \dots, 5.$$

Table 1. Sample of Female Population: Number of Children per Woman

	Number of Children k					
	0	1	2	3	4	5
N_k (China)	9,080,779	11,519,885	8,548,763	3,187,520	895,589	315,130
N_k (Mexico)	9,421,296	4,070,609	4,852,040	4,164,527	2,643,692	1,743,119

Source: United Nations Demographic Yearbook 2000

Do the following exercise.

Exercise 3.1. Assume that $\{p_k\}$ is the distribution of the number of children not only for a woman, but for any individual – female or male. Assume that the probability that a child is female is $1/2$. For both China and Mexico, calculate: (a) the probability that the line of descendants of a given individual gets eventually extinct; (b) the probability that the line of female descendants of a given woman gets eventually extinct. You have to:

- Produce a report that clearly describes what you do and how, and your conclusions. (This is one pdf file.)
- Produce a python code(s), which has numbers $\{N_k\}$ as an input, and the extinction probability as an output. (This is another, plain text file.) *Suggestion:* When you solve $P(s) = s$, you can use, for example, binary search, up to some accuracy. But specify what you do in the report.
- So, when you submit this project, you upload two files, report (pdf) and code (plain text: .txt or .py)
- In this project, both the report and the code are important. The code must be compilable. Submission without report or without code is not valid.

References

[1] Anderson, W. Elementary stochastic processes. Class notes. McGill University.

[2] Du, R., Yuan, Y., Hwang, J., Mountain, J. and Cavalli-Sforza, L. (1992) Chinese surnames and the genetic differences between North and South China. *Journal of Chinese Linguistics Monograph Series*. **5**: 1–66, 68–93.

[3] Bienayme, I. J. (1845) De la loi de multiplication et de la durée des familles. *Soc. Philomath. Paris, Extraits, Sér.*. **5**: 37–39.

[4] Galton, F. and Watson, H.W. (1874) On the probability of extinction of families. *Journ. Antrop. Inst. of Great Britain and Irland*, **4**, 138–144.

[5] Keyfitz, N. and Tyree, A. (1967) Computerization of the branching process. *Behavioral Science*, **12**, 329–336.