CASE STUDY: EXTINCTION OF FAMILY NAMES

The idea that families 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.

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 extinction probability of Chinese female lines of descent.

Learning Objectives:

• Learn about Galton-Watson-Bienaymé braching process;
• Develop concrete examples for applications of probability generating function;
• Utilize elementary matrix theory to solve recursive relations;
• Use recursive relations to compute various probabilities of extinction.

Suppose that we know from statistical analysis the probabilities in a particular family that a man has 0, 1, 2, ⋯ sons and that each son has the same probabilities of sons of his own, and so on. There are two questions that of essence to this problem.
(1) What is the probability for any given number of descendants in the male line in any given generation?
(2) What is the probability of eventual extinction of the family name?

1. INTRODUCTION

Suppose that one man starts a new family name. Let
\[ Z_n = \text{the number of sons in the } n\text{-th generation}, \quad n = 0, 1, 2, \cdots. \]
Then \( Z_0 = 1 \). Let \( X \) be a generic random variable standing for the number of sons from a male in the family and \( p_k \) be the probability that a male has \( k \) sons, i.e. \( P(X = k) = p_k \) 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
\[
P(s) = P_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} p_k s^k.
\]

We shall write
\[
P_n(s) = \text{the pgf of } Z_n, \quad n = 0, 1, 2, \cdots.
\]
For notational brevity, we shall also write
\[
m = \mathbb{E}(X) = P'(1) = \text{the mean of the number of sons from a given male,}
\]
and
\[
s^2 = \mathbb{V}(X) = P''(1) + P'(1) - (P'(1))^2 = \text{the variance of the number of sons from a given male.}
\]

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.
CASE STUDY: EXTINCTION

We carefully examine how the $n$-th generation carry forward the family name to the $n+1$-st generation. Let us label the males in the $n$-th generation by $1, 2, 3, \ldots, Z_n$ and let $X_i$ be the number of sons from the male with the label $i$. Then the total number of males 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

$$P_{n+1}(s) = P_n(P(s)), \quad n = 0, 1, 2, \ldots.$$  

Exercise 1.2. Show that

$$P_{n+1}(s) = P(P_n(s)), \quad n = 0, 1, 2, \ldots.$$  

Let us consider a concrete example where the fundamental equation (2) can be used to derive the pgf $P_n(s)$ for any $n$, which can be then “inverted” to find the terms $P(Z_n = k)$.

Exercise 1.3. Suppose that the probabilities $p_k$'s are given by

$$p_k = pq^k, \quad k = 0, 1, 2, \ldots.$$  

Then

$$P(s) = \sum_{k=0}^{\infty} pq^k s^k = \frac{p}{1 - qs}.$$  

Show that

$$P_2(s) = \frac{p(1 - qs)}{1 - qs - pq}.$$  

Furthermore, assume that the probabilities of having a son and having a daughter are equal, i.e. $p = q = 1/2$.

1. Calculate the probability of having no male in the second generation without using a pgf.
2. Write down the Taylor expansion of $P_2(s)$ and calculate the same probability. Check if it agrees with your answers from part (1).

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

Exercise 1.4. Show that:

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

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

(Hint: differentiate the fundamental equation (2) to find recursive formulas for moments.)
2. Probability distribution of the number of descendants

In this section, we apply the methodology developed in the previous section to actual data and calculate the probability distribution of descendants in a female line. We will show by example of how to compute the iterative function from the previous section in matrix form.

Define \(c_i\) to be the proportion of women who have \(i = 0, 1, 2, \ldots\) children (e.g., \(c_0\) is the proportion of women who have no children, \(c_1\) is the proportion of women who have one child and so on). Let \(g\) define the proportion of births which are girls. Then, of the \(c_1\) mothers with one child, the proportion of daughters is \(gc_1\). Of the \(c_2\) mothers with two children, we have \(g^2c_2\) with two daughters, \(2g(1 - g)c_2\) with one daughter and \((1 - g)^2c_2\) with no daughters. This division according to the number of daughters is a binomial distribution.

Let \(E_j\) denote the event of having \(j = 0, 1, 2, \ldots\) daughters over the course of one’s life. Denote the probabilities of obtaining these events to be \(\pi_j\),

\[
\pi_j = P(E_j)
\]

where \(\pi_0 + \pi_1 + \ldots + \pi_n = 1\).

An illustration of these definitions is shown in Table 1. From the table, the summation on rows correspond to \(\pi_j\).

<table>
<thead>
<tr>
<th>Number of Daughters</th>
<th>Total Children</th>
<th>No. of Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>(c_0)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>(c_1)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>(c_2)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>(c_3)</td>
</tr>
<tr>
<td></td>
<td>(n)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td></td>
<td>Total Women</td>
<td>(c)</td>
</tr>
</tbody>
</table>

Exercise 2.1. Table 2 summarizes the total female population in China and Mexico by number of children. According to the 2000 Chinese Census sex ratio, the probability of having a girl is 0.455. Based on this information, write a Python program to compute the distribution of Chinese women according to total number of children and number of daughters. Compute that the probabilities for \(i = 0, 1, 2, 3, 4, 5\) so you should have a \(6 \times 6\) matrix in the form of Table 1.
### Table 2. China Female Population by Number of Children

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>No. of Women</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td></td>
<td>9,080,779</td>
<td>11,519,885</td>
<td>8,548,763</td>
<td>3,187,520</td>
<td>895,589</td>
<td>315,130</td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td>9,421,296</td>
<td>4,070,609</td>
<td>4,852,040</td>
<td>4,164,527</td>
<td>2,643,692</td>
<td>1,743,119</td>
</tr>
</tbody>
</table>

Source: United Nations Demographic Yearbook 2000

(Hint: Apply binomial expansion to get rest of the values in the table. You can check that your calculations are correct if your π_j’s sum up to 1.)

Let Π be a vector of the probability values, \( \Pi = [\pi_0 \pi_1 \pi_2 \pi_3 \ldots] \). To carry one generation into the next, we multiply Π by the infinite transition matrix \( P \)

\[
\Pi \cdot P = [\pi_0 \pi_1 \pi_2 \pi_3 \ldots] \cdot \begin{bmatrix}
1 & 0 & 0 & \ldots \\
\pi_0 & \pi_1 & \pi_2 & \ldots \\
\pi_0^2 & 2\pi_1\pi_2 & 2\pi_2\pi_0 + \pi_1^2 & \ldots \\
\pi_0^3 & 3\pi_1\pi_2^2 & 3\pi_1\pi_0^2 & \ldots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\]

\( P \) represent the transition from one woman existing with certainty, to the probabilities of 0, 1, 2, ... daughters. Thus, \( \Pi P \) is the probabilities for numbers of daughters; for 0, 1, 2, ... granddaughters through daughters are \( \Pi P^2 \) and, in general for the nth generation is \( \Pi P^n \). The values of \( P \) comes from the generating function which can be obtained by noting that the jth cell of the ith row is the coefficient of \( s^{i-1} \) from Equation 1. In general, the values of the transition matrix is given by

\[
p_{ij} = \pi_0p_{i-1,j} + \pi_1p_{i-1,j-1} + \pi_2p_{i-1,j-2} + \ldots
\]

The cells of each row are obtained by multiplying the cells above and to the left by \( \pi_0, \pi_1, \ldots \) and so on. For example, for cell (3,3), we have \( p_{33} = \pi_0p_{23} + \pi_1p_{22} + \pi_2p_{21} = \pi_0(\pi_2) + \pi_1(\pi_1) + \pi_2(\pi_0) = \pi_1^2 + 2\pi_2\pi_1 \).

**Exercise 2.2.** Using Python, construct the transition matrix \( P \) for the Chinese line. The size of your matrix should be 6 × 6.

**Exercise 2.3.** Using your solution from Exercise 2.1 and 2.2, compute the probability of \( N = 0, 1, \ldots, 5 \) descendants up to 5th generation of the Chinese female line.

The computation for the Mexico data is given in the tables below. You may verify your program code with the Mexico data as a test case.
TABLE 3. Mexico Distribution of Women According to Total
Children and Number of Daughters

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0.0000</th>
<th>0.0000</th>
<th>0.0000</th>
<th>0.0000</th>
<th>0.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0825</td>
<td>0.0689</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>0.0536</td>
<td>0.0895</td>
<td>0.0373</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>0.0251</td>
<td>0.0628</td>
<td>0.0524</td>
<td>0.0146</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>0.0087</td>
<td>0.0290</td>
<td>0.0363</td>
<td>0.0202</td>
<td>0.0042</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>0.0031</td>
<td>0.0130</td>
<td>0.0217</td>
<td>0.0181</td>
<td>0.0076</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

Table 4. Transition Matrix for Mexico

<table>
<thead>
<tr>
<th></th>
<th>1.0000</th>
<th>0.0000</th>
<th>0.0000</th>
<th>0.0000</th>
<th>0.0000</th>
<th>0.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5232</td>
<td>0.2631</td>
<td>0.1477</td>
<td>0.0529</td>
<td>0.0118</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>0.2738</td>
<td>0.2753</td>
<td>0.2238</td>
<td>0.1331</td>
<td>0.0620</td>
<td>0.0232</td>
</tr>
<tr>
<td></td>
<td>0.1432</td>
<td>0.2161</td>
<td>0.2300</td>
<td>0.1837</td>
<td>0.1183</td>
<td>0.0635</td>
</tr>
<tr>
<td></td>
<td>0.0749</td>
<td>0.1507</td>
<td>0.1983</td>
<td>0.1961</td>
<td>0.1573</td>
<td>0.1064</td>
</tr>
<tr>
<td></td>
<td>0.0392</td>
<td>0.0986</td>
<td>0.1545</td>
<td>0.1810</td>
<td>0.1721</td>
<td>0.1384</td>
</tr>
</tbody>
</table>

Table 5. Probability of N Descendants in X Generations
(Mexico)

<table>
<thead>
<tr>
<th>N</th>
<th>X=1</th>
<th>X=2</th>
<th>X=3</th>
<th>X=4</th>
<th>X=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7098</td>
<td>0.8064</td>
<td>0.8640</td>
<td>0.9004</td>
<td>0.9240</td>
</tr>
<tr>
<td>1</td>
<td>0.1232</td>
<td>0.0701</td>
<td>0.0435</td>
<td>0.0280</td>
<td>0.0183</td>
</tr>
<tr>
<td>2</td>
<td>0.0866</td>
<td>0.0535</td>
<td>0.0344</td>
<td>0.0225</td>
<td>0.0148</td>
</tr>
<tr>
<td>3</td>
<td>0.0458</td>
<td>0.0321</td>
<td>0.0216</td>
<td>0.0144</td>
<td>0.0095</td>
</tr>
<tr>
<td>4</td>
<td>0.0206</td>
<td>0.0170</td>
<td>0.0121</td>
<td>0.0082</td>
<td>0.0055</td>
</tr>
<tr>
<td>5</td>
<td>0.0086</td>
<td>0.0085</td>
<td>0.0063</td>
<td>0.0044</td>
<td>0.0029</td>
</tr>
</tbody>
</table>

3. Probability of eventual extinction

In the earlier calculation, we have already observed that in each generation there is a significant probability of extinction. Now we turn to a possibly more interesting question - 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 1.4). The only uncertain case is when $p_0 > 0$ in which case we may have $Z_n = 0$ for some $n$. 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
\[
\xi = P(\text{extinction}) = P\left(\lim_{n \to \infty} \bigcup_{k=1}^{n} \{Z_k = 0\}\right) = P\left(\lim_{n \to \infty} \{Z_n = 0\}\right) = \lim_{n \to \infty} P(\{Z_n = 0\}).
\]

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

Let us write \( x_n = P_n(0) = P(Z_n = 0) \) 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 \( \xi \) 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 demonstrate that \( \xi \) has to be the smallest nonnegative root. Suppose \( \rho \) 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 \( \xi \) is indeed the smallest such root.

**Exercise 3.1.** 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 \( P(1) = 1 \).) \( \square \)

**Exercise 3.2.** Let us return the probability of extinction of family names. In 1939, Lotka used the methods of this section to calculate the extinction probability for American male lines of descent. From the 1920 U.S. census, he found that the probabilities \( \{p_k, k \geq 0\} \) were well represented by the distribution
\[
p_k = be^{k-1}, \quad k \geq 1, \quad p_0 = 1 - \sum_{k=1}^{\infty} p_k,
\]
where \( b = 0.2126 \) and \( c = 0.5893 \). Reproduce Lotka’s result on the probability of eventual extinction. \( \square \)
REFERENCES


