Week 4: Markov chains
Branching processes and mitochondrial DNA

Mitochondrial DNA is passed from mother to children without genetic contribution from the father. All the variability in mitochondrial DNA is due to random mutations accumulated over time. Using estimates of the mutation rate and differences in mitochondrial DNA between humans, it would be possible to estimate the periods of time at which groups became distinct populations. However, we are not going to compute these times here. We will instead focus on a observing a few facts about a Markov chain (MC) that models the propagation of mitochondrial DNA. Notice that since male mitochondrial DNA is not passed on to children, it suffices to focus on female descendants of females, i.e., from mothers to daughters to daughters of daughters and so on.

Let us start by letting X_n denote the number of women in the n-th generation and X_{nr} be the number of women whose mitochondrial DNA is of type r. Assign indexes $i = 1, 2, \ldots, X_n$ to all of the n-th generation individuals. Independently of time n or her type r, the i-th woman has D_i daughters with probability distribution

$$\mathbb{P}\left[D_i = j\right] = p_j. \tag{1}$$

Because mutations are rare, the type of each daughter coincides with that of their mother in most cases. Once in a while, however, a mother of type r bears a daughter of a different type, say s. This occurs with probability q. When this happens, we assume that type s is new, i.e., that it is different from any other type that has ever existed. This is a reasonable assumption because mutations are rare and can happen in a large number of genes. The probability that the same mutation occurs twice can therefore be ignored. To simplify the calculations, we assume that mutations are "latent", i.e., if a mother of type r has a daughter for which a mutation occurs, then the daughter will still be of type r, but all of her daughters will be type s. In other words, when a mutation occurs, it occurs for all daughters. Then, assuming the mother is of type r, we can separate the probability of bearing D_{ir} daughters of type r or D_{is} daughters of type s as

$$\mathbb{P}\left[D_{ir} = j\right] = (1-q)p_j \qquad \text{and} \qquad \mathbb{P}\left[D_{is} = j\right] = qp_j. \tag{2}$$

The first probability accounts for the case when no mutations occur. The second one accounts for the daughters of a woman in which the mutation first arises.

For future reference, define ν to be the expected value of the total number of daughters and ν_r to be the expected number of daughters that share their mother's type, i.e.,

$$\nu = \mathbb{E}\left[D_i\right] = \sum_{j=1}^{\infty} jp_j, \quad \nu_r = \mathbb{E}\left[D_{ir}\right] = (1-q)\sum_{j=1}^{\infty} jp_j. \tag{3}$$

A Is the total number of women a Markov chain? Consider the total number of women X_n. Is the process $\{X_n\}_{n\in\mathbb{N}}$ an MC? If so, describe the transition probabilities $P_{0j} = \mathbb{P}\left[X_{n+1} = j \mid X_n = 0\right]$ and $P_{1j} = \mathbb{P}\left[X_{n+1} = j \mid X_n = 1\right]$ for all j. What are the transition probabilities into state 0, i.e., $P_{i0} = \mathbb{P}\left[X_{n+1} = 0 \mid X_n = i\right]$ for all i? Is the probability $P_{ii} = \mathbb{P}\left[X_{n+1} = i \mid X_n = i\right]$ of a state transitioning into itself strictly positive? Is this MC recurrent?

B Is the number of women of a certain DNA type a Markov chain? Consider the number of women X_{nr} of mitochondrial DNA of type r. As defined, the process $\{X_{nr}\}_{n\in\mathbb{N}}$ is *not* a MC. Why? Suppose now that we are given the information that at some time N, $X_{Nr} > 0$. Consider the stochastic process $\hat{X}_{ir} = X_{(N+i)r}$, for $i = 0, 1, \dots$, with the information that $\hat{X}_{0r} = X_{Nr} > 0$. This process is a MC. Why? Describe the transition probabilities $P_{0j} = \mathbb{P}\left[\hat{X}_{(n+1)r} = j \mid \hat{X}_{nr} = 0\right]$ and $P_{1j} = \mathbb{P}\left[\hat{X}_{(n+1)r} = j \mid \hat{X}_{nr} = 1\right]$ for all j. What are the transition probabilities into state 0, i.e., $P_{i0} = \mathbb{P}\left[\hat{X}_{(n+1)r} = 0 \mid \hat{X}_{nr} = i\right]$ for all i? Is the probability $P_{ii} = \mathbb{P}\left[\hat{X}_{(n+1)r} = i \mid \hat{X}_{nr} = i\right]$ of a state transitioning into itself strictly positive? Is this MC recurrent?

C System simulation. Write a simulation of this stochastic system. You can model the number of children as Poisson with mean $\lambda = 1.05$, which is half the fertility rate of the United States. If you are up for a challenge, you can approximate the probabilities from the following distribution taken from the number of children ever born to women in the age group 40-44[1]

Number of children	Percentage	Number of children	Percentage
0	0.179	1	0.174
2	0.354	3	0.189
4	0.068	5,6	0.028
> 7	0.008		

If you decide to use the data in this table, notice that the above distribution is for *all* children, both male and female, and that you are interested in girls only. Hand in your code.

D Simulation tests one. Run a simulation with rate of mutation $q = 10^{-2}$, using $X_0 = 100$ women, all with different mitochondrial DNA types. Run for $n = 50$ generations—approximately 1000 years at 20 years per generation. Show plots for the number of women in each type, number of existing types, and accumulated number of extinct types as a function of the generation. Show a bar plot of the number of individuals per type.

E Simulation tests two. Repeat part D with rate of mutation $q = 0$ and $X_0 = 400$ individuals of different types. This experiment shows the chances of your direct female line surviving for the next 10 centuries. Notice that most of the types go extinct, a few have a moderate number of individuals, and one or two have a large number of individuals. This means that far into the future, most of your direct female lines will be extinct except for one of you that will have a very large number of descendants. Who among you will be the one surviving, however, is determined by chance.

F Expected value of the number of direct line female descendants. The number of individuals in the $(n+1)$-th generation can be written in terms of the number X_n of individuals in the n-th generation and the numbers of daughters D_i of each individual. Explicitly,

$$X_{n+1} = \sum_{i=1}^{X_n} D_i. \tag{4}$$

[1] US census bureau, "Distribution of Women by Average Number of Children Ever Born, by Race, Marital Status, and Age," June 2002

Use (4) to prove that if the number of women in the first generation ($n = 0$) is X_0, then the expected number of female individuals in the n-th generation is

$$\mu_n = \mathbb{E}\left[X_n\right] = X_0 \nu^n. \tag{5}$$

Compare (5) with the number of individuals as a function of time you obtained in parts D and E. If they are similar, explain the similarity. If they are not, explain the differences.

Likewise, prove that the expected number of descendants sharing the mitochondrial DNA type of one type r woman from the zeroth generation is

$$\mu_{nr} = \mathbb{E}\left[X_{nr}\right] = \nu_r^n = (1-q)^n \nu^n. \tag{6}$$

G Extinction in probability and almost sure extinction. Show that if $\nu_r < 1$ then type r goes extinct in probability independently of the number of individuals in the original generation, i.e,

$$\lim_{n\to\infty} \mathbb{P}\left[X_{nr} = 0\right] = 1. \tag{7}$$

Using the fact that after X_{nr} becomes 0 for the first time it stays at 0 show that when $\nu_r < 1$ types become extinct almost surely, i.e.,

$$\mathbb{P}\left[\lim_{n\to\infty} X_{nr} = 0\right] = 1. \tag{8}$$

Note that these are very different conclusions: (7) is the limit of a probability whereas (8) is the probability of a limit.

H Probability of extinction in m generations. Denote as $P_{em}(x)$ the probability that a certain mitochondrial type becomes extinct before the m-th generation when the number of individuals with this type in zeroth generation x. For instance, consider there is a single individual of type r at time 0. Then, $P_{e1}(1)$ is the probability that type r goes extinct at generation 1, i.e., that she does not have any descendants in the first generation. In other words, it is the probability that she bears no daughters. The probability $P_{e2}(1)$ refers to the event that she has no descendants in the second generation, either because she did not have daughters or because her daughters did not have daughters. In general, $P_{em}(1)$ is the probability that her direct female line died out some time before the m-th generation. Compute the probability $P_{e1}(1)$ of extinction in one generation. For extinction in more than one generation prove that the following recursion is true

$$P_{em}(1) = \sum_{j=1}^{\infty} p_j \left[P_{e(m-1)}(1)\right]^j. \tag{9}$$

For $x \neq 1$ argue that the probability of extinction is simply $P_{em}(x) = \left[P_{em}(1)\right]^x$. Compare this value as a function of time with the numerical estimate you obtain from the simulations in parts D and E.

I Probability of eventual extinction. Denote as $P_e(x)$ the probability of eventual extinction—i.e., extinction at sometime between now and infinity—of a mitochondrial DNA type when the number of original individuals is x. Show that $P_e(1)$ is the solution of the following equation

$$P_e(1) = \sum_{j=1}^{\infty} p_j \left[P_e(1)\right]^j . \tag{10}$$

Argue that in general $P_e(x) = [P_e(1)]^x$.