23
Probability Redux: Stochastic Processes

23.1 Introduction

Most of this book has focused on IID sequences of random variables. Now we consider sequences of dependent random variables. For example, daily temperatures will form a sequence of time-ordered random variables and clearly the temperature on one day is not independent of the temperature on the previous day.

A **stochastic process** $\{X_t : t \in T\}$ is a collection of random variables. We shall sometimes write $X(t)$ instead of X_t. The variables X_t take values in some set $\mathcal{X}$ called the **state space**. The set T is called the **index set** and for our purposes can be thought of as time. The index set can be discrete $T = \{0, 1, 2, \ldots\}$ or continuous $T = [0, \infty)$ depending on the application.

23.1 Example (IID observations). A sequence of IID random variables can be written as $\{X_t : t \in T\}$ where $T = \{1, 2, 3, \ldots, \}$. Thus, a sequence of IID random variables is an example of a stochastic process. ■

23.2 Example (The Weather). Let $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. A typical sequence (depending on where you live) might be

$$\text{sunny}, \text{sunny}, \text{cloudy}, \text{sunny}, \text{cloudy}, \text{cloudy}, \cdots$$

This process has a discrete state space and a discrete index set. ■

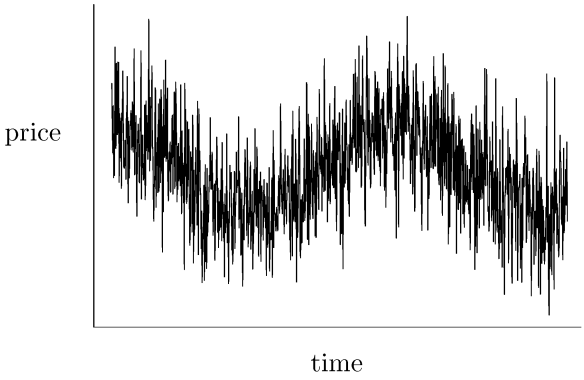


FIGURE 23.1. Stock price over ten week period.

23.3 Example (Stock Prices)**.** Figure 23.1 shows the price of a fictitious stock over time. The price is monitored continuously so the index set T is continuous. Price is discrete but for all practical purposes we can treat it as a continuous variable. ■

23.4 Example (Empirical Distribution Function)**.** Let $X_1, \ldots, X_n \sim F$ where F is some CDF on [0,1]. Let

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t)$$

be the empirical CDF. For any fixed value t, $\widehat{F}_n(t)$ is a random variable. But the whole empirical CDF

$$\left\{ \widehat{F}_n(t) : \ t \in [0,1] \right\}$$

is a stochastic process with a continuous state space and a continuous index set. ■

We end this section by recalling a basic fact. If $X_1, \ldots, X_n$ are random variables, then we can write the joint density as

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1) f(x_2|x_1) \cdots f(x_n|x_1, \ldots, x_{n-1}) \\ &= \prod_{i=1}^{n} f(x_i|\text{past}_i) \end{aligned} \tag{23.1}$$

where $\text{past}_i = (X_1, \ldots, X_{i-1})$.

23.2 Markov Chains

A Markov chain is a stochastic process for which the distribution of X_t depends only on X_{t-1}. In this section we assume that the state space is discrete, either $\mathcal{X} = \{1, \ldots, N\}$ or $\mathcal{X} = \{1, 2, \ldots, \}$ and that the index set is $T = \{0, 1, 2, \ldots\}$. Typically, most authors write X_n instead of X_t when discussing Markov chains and I will do so as well.

23.5 Definition. *The process* $\{X_n : n \in T\}$ *is a* **Markov chain** *if*

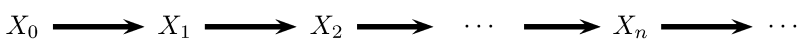
$$\mathbb{P}(X_n = x \mid X_0, \ldots, X_{n-1}) = \mathbb{P}(X_n = x \mid X_{n-1}) \tag{23.2}$$

for all n *and for all* $x \in \mathcal{X}$.

For a Markov chain, equation (23.1) simplifies to

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2) \cdots f(x_n|x_{n-1}).$$

A Markov chain can be represented by the following DAG:



Each variable has a single parent, namely, the previous observation.

The theory of Markov chains is a very rich and complex. We have to get through many definitions before we can do anything interesting. Our goal is to answer the following questions:

1. When does a Markov chain "settle down" into some sort of equilibrium?
2. How do we estimate the parameters of a Markov chain?
3. How can we construct Markov chains that converge to a given equilibrium distribution and why would we want to do that?

We will answer questions 1 and 2 in this chapter. We will answer question 3 in the next chapter. To understand question 1, look at the two chains in Figure 23.2. The first chain oscillates all over the place and will continue to do so forever. The second chain eventually settles into an equilibrium. If we constructed a histogram of the first process, it would keep changing as we got

more and more observations. But a histogram from the second chain would eventually converge to some fixed distribution.

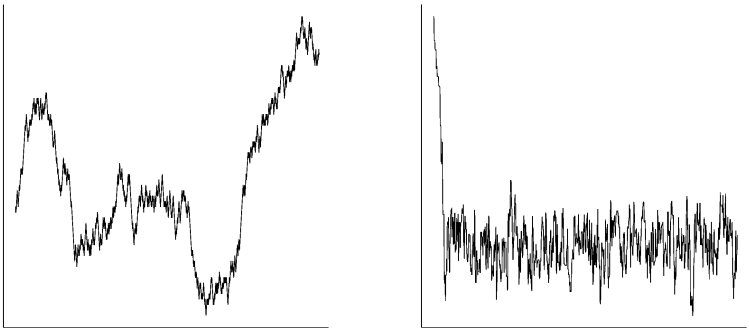


FIGURE 23.2. Two Markov chains. The first chain does not settle down into an equilibrium. The second does.

Transition Probabilities. The key quantities of a Markov chain are the probabilities of jumping from one state into another state. A Markov chain is **homogeneous** if $\mathbb{P}(X_{n+1} = j | X_n = i)$ does not change with time. Thus, for a homogeneous Markov chain, $\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_1 = j | X_0 = i)$. We shall only deal with homogeneous Markov chains.

23.6 Definition. *We call*

$$p_{ij} \equiv \mathbb{P}(X_{n+1} = j | X_n = i) \tag{23.3}$$

the **transition probabilities.** *The matrix* $\mathbf{P}$ *whose* (i, j) *element is* p_{ij} *is called the* **transition matrix.**

We will only consider homogeneous chains. Notice that $\mathbf{P}$ has two properties: (i) $p_{ij} \geq 0$ and (ii) $\sum_i p_{ij} = 1$. Each row can be regarded as a probability mass function.

23.7 Example (Random Walk With Absorbing Barriers)**.** Let $\mathcal{X} = \{1, \ldots, N\}$. Suppose you are standing at one of these points. Flip a coin with $\mathbb{P}(\text{Heads}) = p$ and $\mathbb{P}(\text{Tails}) = q = 1 - p$. If it is heads, take one step to the right. If it is tails, take one step to the left. If you hit one of the endpoints, stay there. The

transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad \blacksquare$$

23.8 Example. Suppose the state space is $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. Then X_1, X_2, ... represents the weather for a sequence of days. The weather today clearly depends on yesterday's weather. It might also depend on the weather two days ago but as a first approximation we might assume that the dependence is only one day back. In that case the weather is a Markov chain and a typical transition matrix might be

	Sunny	Cloudy
Sunny	0.4	0.6
Cloudy	0.8	0.2

For example, if it is sunny today, there is a 60 per cent chance it will be cloudy tomorrow. ■

Let

$$p_{ij}(n) = \mathbb{P}(X_{m+n} = j | X_m = i) \tag{23.4}$$

be the probability of of going from state i to state j in n steps. Let $\mathbf{P}_n$ be the matrix whose (i, j) element is $p_{ij}(n)$. These are called the **n-step transition probabilities.**

23.9 Theorem (The Chapman-Kolmogorov equations)**.** *The n-step probabilities satisfy*

$$p_{ij}(m+n) = \sum_k p_{ik}(m) p_{kj}(n). \tag{23.5}$$

PROOF. Recall that, in general,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y | X = x).$$

This fact is true in the more general form

$$\mathbb{P}(X = x, Y = y | Z = z) = \mathbb{P}(X = x | Z = z)\mathbb{P}(Y = y | X = x, Z = z).$$

Also, recall the law of total probability:

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y).$$

Using these facts and the Markov property we have

$$\begin{aligned} p_{ij}(m+n) &= \mathbb{P}(X_{m+n} = j | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j, X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k, X_0 = i) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j | X_m = k) \mathbb{P}(X_m = k | X_0 = i) \\ &= \sum_k p_{ik}(m) p_{kj}(n). \quad \blacksquare \end{aligned}$$

Look closely at equation (23.5). This is nothing more than the equation for matrix multiplication. Hence we have shown that

$$\mathbf{P}_{m+n} = \mathbf{P}_m \mathbf{P}_n. \tag{23.6}$$

By definition, $\mathbf{P}_1 = \mathbf{P}$. Using the above theorem, $\mathbf{P}_2 = \mathbf{P}_{1+1} = \mathbf{P}_1 \mathbf{P}_1 = \mathbf{P}\mathbf{P} = \mathbf{P}^2$. Continuing this way, we see that

$$\mathbf{P}_n = \mathbf{P}^n \equiv \underbrace{\mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P}}_{\text{multiply the matrix n times}}. \tag{23.7}$$

Let $\mu_n = (\mu_n(1), \ldots, \mu_n(N))$ be a row vector where

$$\mu_n(i) = \mathbb{P}(X_n = i) \tag{23.8}$$

is the marginal probability that the chain is in state i at time n. In particular, μ_0 is called the **initial distribution.** To simulate a Markov chain, all you need to know is μ_0 and $\mathbf{P}$. The simulation would look like this:

Step 1: Draw $X_0 \sim \mu_0$. Thus, $\mathbb{P}(X_0 = i) = \mu_0(i)$.
Step 2: Denote the outcome of step 1 by i. Draw $X_1 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_1 = j | X_0 = i) = p_{ij}$.
Step 3: Suppose the outcome of step 2 is j. Draw $X_2 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_2 = k | X_1 = j) = p_{jk}$.
And so on.

It might be difficult to understand the meaning of μ_n. Imagine simulating the chain many times. Collect all the outcomes at time n from all the chains. This histogram would look approximately like μ_n. A consequence of theorem 23.9 is the following:

23.10 Lemma. *The marginal probabilities are given by*

$$\mu_n = \mu_0 \mathbf{P}^n.$$

PROOF.

$$\begin{aligned} \mu_n(j) &= \mathbb{P}(X_n = j) \\ &= \sum_i \mathbb{P}(X_n = j | X_0 = i) P(X_0 = i) \\ &= \sum_i \mu_0(i) p_{ij}(n) = \mu_0 \mathbf{P}^n. \quad \blacksquare \end{aligned}$$

Summary of Terminology

1. Transition matrix: $\mathbf{P}(i,j) = \mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij}$.
2. n-step matrix: $\mathbf{P}_n(i,j) = \mathbb{P}(X_{n+m} = j | X_m = i)$.
3. $\mathbf{P}_n = \mathbf{P}^n$.
4. Marginal: $\mu_n(i) = \mathbb{P}(X_n = i)$.
5. $\mu_n = \mu_0 \mathbf{P}^n$.

STATES. The states of a Markov chain can be classified according to various properties.

23.11 Definition. *We say that i* **reaches** *j (or j is* **accessible** *from i) if $p_{ij}(n) > 0$ for some n, and we write $i \to j$. If $i \to j$ and $j \to i$ then we write $i \leftrightarrow j$ and we say that i and j* **communicate.**

23.12 Theorem. *The communication relation satisfies the following properties:*

1. *$i \leftrightarrow i$.*
2. *If $i \leftrightarrow j$ then $j \leftrightarrow i$.*
3. *If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.*
4. *The set of states $\mathcal{X}$ can be written as a disjoint union of* **classes** *$\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2 \bigcup \cdots$ where two states i and j communicate with each other if and only if they are in the same class.*

If all states communicate with each other, then the chain is called **irreducible**. A set of states is **closed** if, once you enter that set of states you never leave. A closed set consisting of a single state is called an **absorbing state.**

23.13 Example. Let $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The classes are $\{1, 2\}, \{3\}$ and $\{4\}$. State 4 is an absorbing state. ∎

Suppose we start a chain in state i. Will the chain ever return to state i? If so, that state is called persistent or recurrent.

23.14 Definition. *State i is* **recurrent** *or* **persistent** *if*

$$\mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i) = 1.$$

Otherwise, state i is **transient**.

23.15 Theorem. *A state i is recurrent if and only if*

$$\sum_n p_{ii}(n) = \infty. \tag{23.9}$$

A state i is transient if and only if

$$\sum_n p_{ii}(n) < \infty. \tag{23.10}$$

PROOF. Define

$$I_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i. \end{cases}$$

The number of times that the chain is in state i is $Y = \sum_{n=0}^{\infty} I_n$. The mean of Y, given that the chain starts in state i, is

$$\mathbb{E}(Y|X_0 = i) = \sum_{n=0}^{\infty} \mathbb{E}(I_n|X_0 = i) = \sum_{n=0}^{\infty} \mathbb{P}(X_n = i|X_0 = i) = \sum_{n=0}^{\infty} p_{ii}(n).$$

Define $a_i = \mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i)$. If i is recurrent, $a_i = 1$. Thus, the chain will eventually return to i. Once it does return to i, we argue again

that since $a_i = 1$, the chain will return to state i again. By repeating this argument, we conclude that $\mathbb{E}(Y|X_0 = i) = \infty$. If i is transient, then $a_i < 1$. When the chain is in state i, there is a probability $1 - a_i > 0$ that it will never return to state i. Thus, the probability that the chain is in state i exactly n times is $a_i^{n-1}(1 - a_i)$. This is a geometric distribution which has finite mean. ∎

23.16 Theorem. *Facts about recurrence.*

1. *If state i is recurrent and $i \leftrightarrow j$, then j is recurrent.*
2. *If state i is transient and $i \leftrightarrow j$, then j is transient.*
3. *A finite Markov chain must have at least one recurrent state.*
4. *The states of a finite, irreducible Markov chain are all recurrent.*

23.17 Theorem (Decomposition Theorem)**.** *The state space $\mathcal{X}$ can be written as the disjoint union*

$$\mathcal{X} = \mathcal{X}_T \bigcup \mathcal{X}_1 \bigcup \mathcal{X}_2 \cdots$$

where $\mathcal{X}_T$ are the transient states and each $\mathcal{X}_i$ is a closed, irreducible set of recurrent states.

23.18 Example (Random Walk)**.** Let $\mathcal{X} = \{\ldots, -2, -1, 0, 1, 2, \ldots, \}$ and suppose that $p_{i,i+1} = p$, $p_{i,i-1} = q = 1 - p$. All states communicate, hence either all the states are recurrent or all are transient. To see which, suppose we start at $X_0 = 0$. Note that

$$p_{00}(2n) = \binom{2n}{n} p^n q^n \tag{23.11}$$

since the only way to get back to 0 is to have n heads (steps to the right) and n tails (steps to the left). We can approximate this expression using Stirling's formula which says that

$$n! \sim n^n \sqrt{n} e^{-n} \sqrt{2\pi}.$$

Inserting this approximation into (23.11) shows that

$$p_{00}(2n) \sim \frac{(4pq)^n}{\sqrt{n\pi}}.$$

It is easy to check that $\sum_n p_{00}(n) < \infty$ if and only if $\sum_n p_{00}(2n) < \infty$. Moreover, $\sum_n p_{00}(2n) = \infty$ if and only if $p = q = 1/2$. By Theorem (23.15), the chain is recurrent if $p = 1/2$ otherwise it is transient. ∎

Convergence of Markov Chains. To discuss the convergence of chains, we need a few more definitions. Suppose that $X_0 = i$. Define the **recurrence time**

$$T_{ij} = \min\{n > 0 : \ X_n = j\} \tag{23.12}$$

assuming X_n ever returns to state i, otherwise define $T_{ij} = \infty$. The **mean recurrence time** of a recurrent state i is

$$m_i = \mathbb{E}(T_{ii}) = \sum_n n f_{ii}(n) \tag{23.13}$$

where

$$f_{ij}(n) = \mathbb{P}(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j | X_0 = i).$$

A recurrent state is **null** if $m_i = \infty$ otherwise it is called **non-null** or **positive.**

23.19 Lemma. *If a state is null and recurrent, then* $p_{ii}^n \to 0$.

23.20 Lemma. *In a finite state Markov chain, all recurrent states are positive.*

Consider a three-state chain with transition matrix

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Suppose we start the chain in state 1. Then we will be in state 3 at times 3, 6, 9, ... This is an example of a periodic chain. Formally, the **period** of state i is d if $p_{ii}(n) = 0$ whenever n is not divisible by d and d is the largest integer with this property. Thus, $d = \gcd\{n : \ p_{ii}(n) > 0\}$ where gcd means "greater common divisor." State i is **periodic** if $d(i) > 1$ and **aperiodic** if $d(i) = 1$. A state with period 1 is called **aperiodic**.

23.21 Lemma. *If state i has period d and $i \leftrightarrow j$ then j has period d.*

23.22 Definition. *A state is* **ergodic** *if it is recurrent, non-null and aperiodic. A chain is ergodic if all its states are ergodic.*

Let $\pi = (\pi_i : \ i \in \mathcal{X})$ be a vector of non-negative numbers that sum to one. Thus π can be thought of as a probability mass function.

23.23 Definition. *We say that π is a* **stationary** *(or* **invariant***) distribution if $\pi = \pi\mathbf{P}$.*

Here is the intuition. Draw X_0 from distribution π and suppose that π is a stationary distribution. Now draw X_1 according to the transition probability of the chain. The distribution of X_1 is then $\mu_1 = \mu_0\mathbf{P} = \pi\mathbf{P} = \pi$. The distribution of X_2 is $\pi\mathbf{P}^2 = (\pi\mathbf{P})\mathbf{P} = \pi\mathbf{P} = \pi$. Continuing this way, we see that the distribution of X_n is $\pi\mathbf{P}^n = \pi$. In other words:

> If at any time the chain has distribution π, then it will continue to have distribution π forever.

23.24 Definition. *We say that a chain has* **limiting distribution** *if*

$$\mathbf{P}^n \to \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}$$

for some π, that is, $\pi_j = \lim_{n\to\infty} \mathbf{P}^n_{ij}$ exists and is independent of i.

Here is the main theorem about convergence. The theorem says that an ergodic chain converges to its stationary distribution. Also, sample averages converge to their theoretical expectations under the stationary distribution.

23.25 Theorem. *An irreducible, ergodic Markov chain has a unique stationary distribution π. The limiting distribution exists and is equal to π. If g is any bounded function, then, with probability 1,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi(g) \equiv \sum_j g(j)\pi_j. \tag{23.14}$$

Finally, there is another definition that will be useful later. We say that π satisfies **detailed balance** if

$$\pi_i p_{ij} = p_{ji}\pi_j. \tag{23.15}$$

Detailed balance guarantees that π is a stationary distribution.

23.26 Theorem. *If π satisfies detailed balance, then π is a stationary distribution.*

PROOF. We need to show that $\pi\mathbf{P} = \pi$. The j^{th} element of $\pi\mathbf{P}$ is $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$. ∎

The importance of detailed balance will become clear when we discuss Markov chain Monte Carlo methods in Chapter 24.

Warning! Just because a chain has a stationary distribution does not mean it converges.

23.27 Example. Let

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Let $\pi = (1/3, 1/3, 1/3)$. Then $\pi P = \pi$ so π is a stationary distribution. If the chain is started with the distribution π it will stay in that distribution. Imagine simulating many chains and checking the marginal distribution at each time n. It will always be the uniform distribution π. But this chain does not have a limit. It continues to cycle around forever. ∎

Examples of Markov Chains.

23.28 Example. Let $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. Let

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Then $C_1 = \{1, 2\}$ and $C_2 = \{5, 6\}$ are irreducible closed sets. States 3 and 4 are transient because of the path $3 \to 4 \to 6$ and once you hit state 6 you cannot return to 3 or 4. Since $p_{ii}(1) > 0$, all the states are aperiodic. In summary, 3 and 4 are transient while 1, 2, 5, and 6 are ergodic. ∎

23.29 Example (Hardy-Weinberg)**.** Here is a famous example from genetics. Suppose a gene can be type A or type a. There are three types of people (called genotypes): AA, Aa, and aa. Let (p, q, r) denote the fraction of people of each genotype. We assume that everyone contributes one of their two copies of the gene at random to their children. We also assume that mates are selected at random. The latter is not realistic however, it is often reasonable to assume that you do not choose your mate based on whether they are AA, Aa, or aa. (This would be false if the gene was for eye color and if people chose mates based on eye color.) Imagine if we pooled everyone's genes together. The proportion of A genes is $P = p + (q/2)$ and the proportion of a genes is

$Q = r + (q/2)$. A child is AA with probability P^2, aA with probability $2PQ$, and aa with probability Q^2. Thus, the fraction of A genes in this generation is

$$P^2 + PQ = \left(p + \frac{q}{2}\right)^2 + \left(p + \frac{q}{2}\right)\left(r + \frac{q}{2}\right).$$

However, $r = 1 - p - q$. Substitute this in the above equation and you get $P^2 + PQ = P$. A similar calculation shows that the fraction of "a" genes is Q. We have shown that the proportion of type A and type a is P and Q and this remains stable after the first generation. The proportion of people of type AA, Aa, aa is thus $(P^2, 2PQ, Q^2)$ from the second generation and on. This is called the Hardy-Weinberg law.

Assume everyone has exactly one child. Now consider a fixed person and let X_n be the genotype of their n^{th} descendant. This is a Markov chain with state space $\mathcal{X} = \{AA, Aa, aa\}$. Some basic calculations will show you that the transition matrix is

$$\begin{bmatrix} P & Q & 0 \\ \frac{P}{2} & \frac{P+Q}{2} & \frac{Q}{2} \\ 0 & P & Q \end{bmatrix}.$$

The stationary distribution is $\pi = (P^2, 2PQ, Q^2)$. ■

23.30 Example (Markov chain Monte Carlo)**.** In Chapter 24 we will present a simulation method called Markov chain Monte Carlo (MCMC). Here is a brief description of the idea. Let $f(x)$ be a probability density on the real line and suppose that $f(x) = cg(x)$ where $g(x)$ is a known function and $c > 0$ is unknown. In principle, we can compute c since $\int f(x)dx = 1$ implies that $c = 1/\int g(x)dx$. However, it may not be feasible to perform this integral, nor is it necessary to know c in the following algorithm. Let X_0 be an arbitrary starting value. Given $X_0, \ldots, X_i$, draw X_{i+1} as follows. First, draw $W \sim N(X_i, b^2)$ where $b > 0$ is some fixed constant. Let

$$r = \min\left\{\frac{g(W)}{g(X_i)},\ 1\right\}.$$

Draw $U \sim \text{Uniform}(0, 1)$ and set

$$X_{i+1} = \begin{cases} W & \text{if } U < r \\ X_i & \text{if } U \geq r. \end{cases}$$

We will see in Chapter 24 that, under weak conditions, $X_0, X_1, \ldots,$ is an ergodic Markov chain with stationary distribution f. Hence, we can regard the draws as a sample from f. ■

INFERENCE FOR MARKOV CHAINS. Consider a chain with finite state space $\mathcal{X} = \{1, 2, \ldots, N\}$. Suppose we observe n observations $X_1, \ldots, X_n$ from this chain. The unknown parameters of a Markov chain are the initial probabilities $\mu_0 = (\mu_0(1), \mu_0(2), \ldots,)$ and the elements of the transition matrix $\mathbf{P}$. Each row of $\mathbf{P}$ is a multinomial distribution. So we are essentially estimating N distributions (plus the initial probabilities). Let n_{ij} be the observed number of transitions from state i to state j. The likelihood function is

$$\mathcal{L}(\mu_0, \mathbf{P}) = \mu_0(x_0) \prod_{r=1}^{n} p_{X_{r-1}, X_r} = \mu_0(x_0) \prod_{i=1}^{N} \prod_{j=1}^{N} p_{ij}^{n_{ij}}.$$

There is only one observation on μ_0 so we can't estimate that. Rather, we focus on estimating $\mathbf{P}$. The MLE is obtained by maximizing $\mathcal{L}(\mu_0, \mathbf{P})$ subject to the constraint that the elements are non-negative and the rows sum to 1. The solution is

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_i}$$

where $n_i = \sum_{j=1}^{N} n_{ij}$. Here we are assuming that $n_i > 0$. If not, then we set $\widehat{p}_{ij} = 0$ by convention.

23.31 Theorem (Consistency and Asymptotic Normality of the MLE). *Assume that the chain is ergodic. Let $\widehat{p}_{ij}(n)$ denote the MLE after n observations. Then $\widehat{p}_{ij}(n) \xrightarrow{\text{P}} p_{ij}$. Also,*

$$\left[\sqrt{N_i(n)}(\widehat{p}_{ij} - p_{ij})\right] \rightsquigarrow N(0, \Sigma)$$

where the left-hand side is a matrix, $N_i(n) = \sum_{r=1}^{n} I(X_r = i)$ and

$$\Sigma_{ij,k\ell} = \begin{cases} p_{ij}(1 - p_{ij}) & (i,j) = (k,\ell) \\ -p_{ij}p_{i\ell} & i = k, j \neq \ell \\ 0 & \text{otherwise.} \end{cases}$$

23.3 Poisson Processes

The Poisson process arises when we count occurrences of events over time, for example, traffic accidents, radioactive decay, arrival of email messages, etc. As the name suggests, the Poisson process is intimately related to the Poisson distribution. Let's first review the Poisson distribution.

Recall that X has a Poisson distribution with parameter λ — written $X \sim \text{Poisson}(\lambda)$ — if

$$\mathbb{P}(X = x) \equiv p(x;\ \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Also recall that $\mathbb{E}(X) = \lambda$ and $\mathbb{V}(X) = \lambda$. If $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\nu)$ and $X \amalg Y$, then $X+Y \sim \text{Poisson}(\lambda+\nu)$. Finally, if $N \sim \text{Poisson}(\lambda)$ and $Y|N = n \sim \text{Binomial}(n, p)$, then the marginal distribution of Y is $Y \sim \text{Poisson}(\lambda p)$.

Now we describe the Poisson process. Imagine that you are at your computer. Each time a new email message arrives you record the time. Let X_t be the number of messages you have received up to and including time t. Then, $\{X_t : t \in [0, \infty)\}$ is a stochastic process with state space $\mathcal{X} = \{0, 1, 2, \ldots\}$. A process of this form is called a **counting process**. A Poisson process is a counting process that satisfies certain conditions. In what follows, we will sometimes write $X(t)$ instead of X_t. Also, we need the following notation. Write $f(h) = o(h)$ if $f(h)/h \to 0$ as $h \to 0$. This means that $f(h)$ is smaller than h when h is close to 0. For example, $h^2 = o(h)$.

23.32 Definition. *A* **Poisson process** *is a stochastic process* $\{X_t : t \in [0, \infty)\}$ *with state space* $\mathcal{X} = \{0, 1, 2, \ldots\}$ *such that*

1. $X(0) = 0$.
2. *For any* $0 = t_0 < t_1 < t_2 < \cdots < t_n$, *the increments*

$$X(t_1) - X(t_0),\ X(t_2) - X(t_1),\ \cdots, X(t_n) - X(t_{n-1})$$

   are independent.

3. *There is a function* $\lambda(t)$ *such that*

$$\mathbb{P}(X(t+h) - X(t) = 1) = \lambda(t)h + o(h) \tag{23.16}$$
$$\mathbb{P}(X(t+h) - X(t) \geq 2) = o(h). \tag{23.17}$$

We call $\lambda(t)$ *the* **intensity function**.

The last condition means that the probability of an event in $[t, t+h]$ is approximately $h\lambda(t)$ while the probability of more than one event is small.

23.33 Theorem. *If* X_t *is a Poisson process with intensity function* $\lambda(t)$, *then*

$$X(s+t) - X(s) \sim \text{Poisson}(m(s+t) - m(s))$$

where

$$m(t) = \int_0^t \lambda(s)\, ds.$$

In particular, $X(t) \sim \text{Poisson}(m(t))$. *Hence,* $\mathbb{E}(X(t)) = m(t)$ *and* $\mathbb{V}(X(t)) = m(t)$.

23.34 Definition. *A Poisson process with intensity function $\lambda(t) \equiv \lambda$ for some $\lambda > 0$ is called a* **homogeneous Poisson process** *with rate λ. In this case,*

$$X(t) \sim \text{Poisson}(\lambda t).$$

Let $X(t)$ be a homogeneous Poisson process with rate λ. Let W_n be the time at which the n^{th} event occurs and set $W_0 = 0$. The random variables $W_0, W_1, \ldots,$ are called **waiting times**. Let $S_n = W_{n+1} - W_n$. Then $S_0, S_1, \ldots,$ are called **sojourn times** or **interarrival times**.

23.35 Theorem. *The sojourn times $S_0, S_1, \ldots$ are* IID *random variables. Their distribution is exponential with mean $1/\lambda$, that is, they have density*

$$f(s) = \lambda e^{-\lambda s}, \quad s \geq 0.$$

The waiting time $W_n \sim \text{Gamma}(n, 1/\lambda)$ i.e., it has density

$$f(w) = \frac{1}{\Gamma(n)} \lambda^n w^{n-1} e^{-\lambda t}.$$

Hence, $\mathbb{E}(W_n) = n/\lambda$ and $\mathbb{V}(W_n) = n/\lambda^2$.

PROOF. First, we have

$$\mathbb{P}(S_1 > t) = \mathbb{P}(X(t) = 0) = e^{-\lambda t}$$

with shows that the CDF for S_1 is $1 - e^{-\lambda t}$. This shows the result for S_1. Now,

$$\begin{aligned} \mathbb{P}(S_2 > t | S_1 = s) &= \mathbb{P}(\text{no events in } (s, s+t] | S_1 = s) \\ &= \mathbb{P}(\text{no events in } (s, s+t]) \quad \text{(increments are independent)} \\ &= e^{-\lambda t}. \end{aligned}$$

Hence, S_2 has an exponential distribution and is independent of S_1. The result follows by repeating the argument. The result for W_n follows since a sum of exponentials has a Gamma distribution. ∎

23.36 Example. Figure 23.3 shows requests to a WWW server in Calgary.[1] Assuming that this is a homogeneous Poisson process, $N \equiv X(T) \sim \text{Poisson}(\lambda T)$. The likelihood is

$$\mathcal{L}(\lambda) \propto e^{-\lambda T} (\lambda T)^N$$

[1] See http://ita.ee.lbl.gov/html/contrib/Calgary-HTTP.html for more information.

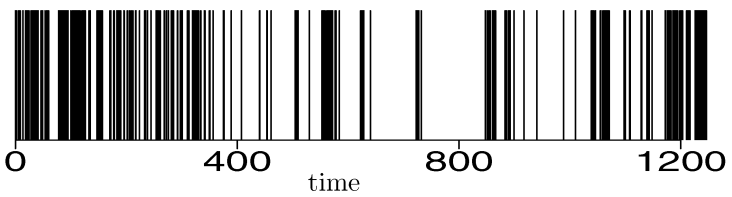


FIGURE 23.3. Hits on a web server. Each vertical line represents one event.

which is maximized at

$$\widehat{\lambda} = \frac{N}{T} = 48.0077$$

in units per minute. Let's now test the assumption that the data follow a homogeneous Poisson process using a goodness-of-fit test. We divide the interval $[0, T]$ into 4 equal length intervals I_1, I_2, I_3, I_4. If the process is a homogeneous Poisson process then, given the total number of events, the probability that an event falls into any of these intervals must be equal. Let p_i be the probability of a point being in I_i. The null hypothesis is that $p_1 = p_2 = p_3 = p_4 = 1/4$. We can test this hypothesis using either a likelihood ratio test or a χ^2 test. The latter is

$$\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the number of observations in I_i and $E_i = n/4$ is the expected number under the null. This yields $\chi^2 = 252$ with a p-value near 0. This is strong evidence against the null so we reject the hypothesis that the data are from a homogeneous Poisson process. This is hardly surprising since we would expect the intensity to vary as a function of time. ■

23.4 Bibliographic Remarks

This is standard material and there are many good references including Grimmett and Stirzaker (1982), Taylor and Karlin (1994), Guttorp (1995), and Ross (2002). The following exercises are from those texts.

23.5 Exercises

1. Let $X_0, X_1, \ldots$ be a Markov chain with states $\{0, 1, 2\}$ and transition matrix
$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.9 & 0.1 & 0.0 \\ 0.1 & 0.8 & 0.1 \end{bmatrix}$$
Assume that $\mu_0 = (0.3, 0.4, 0.3)$. Find $\mathbb{P}(X_0 = 0, X_1 = 1, X_2 = 2)$ and $\mathbb{P}(X_0 = 0, X_1 = 1, X_2 = 1)$.

2. Let $Y_1, Y_2, \ldots$ be a sequence of iid observations such that $\mathbb{P}(Y = 0) = 0.1$, $\mathbb{P}(Y = 1) = 0.3$, $\mathbb{P}(Y = 2) = 0.2$, $\mathbb{P}(Y = 3) = 0.4$. Let $X_0 = 0$ and let
$$X_n = \max\{Y_1, \ldots, Y_n\}.$$
Show that $X_0, X_1, \ldots$ is a Markov chain and find the transition matrix.

3. Consider a two-state Markov chain with states $\mathcal{X} = \{1, 2\}$ and transition matrix
$$\mathbf{P} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$
where $0 < a < 1$ and $0 < b < 1$. Prove that
$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}.$$

4. Consider the chain from question 3 and set $a = .1$ and $b = .3$. Simulate the chain. Let
$$\begin{aligned} \widehat{p}_n(1) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i = 1) \\ \widehat{p}_n(2) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i = 2) \end{aligned}$$
be the proportion of times the chain is in state 1 and state 2. Plot $\widehat{p}_n(1)$ and $\widehat{p}_n(2)$ versus n and verify that they converge to the values predicted from the answer in the previous question.

5. An important Markov chain is the **branching process** which is used in biology, genetics, nuclear physics, and many other fields. Suppose that an animal has Y children. Let $p_k = \mathbb{P}(Y = k)$. Hence, $p_k \geq 0$ for all k and $\sum_{k=0}^{\infty} p_k = 1$. Assume each animal has the same lifespan and

that they produce offspring according to the distribution p_k. Let X_n be the number of animals in the n^{th} generation. Let $Y_1^{(n)}, \ldots, Y_{X_n}^{(n)}$ be the offspring produced in the n^{th} generation. Note that

$$X_{n+1} = Y_1^{(n)} + \cdots + Y_{X_n}^{(n)}.$$

Let $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \mathbb{V}(Y)$. Assume throughout this question that $X_0 = 1$. Let $M(n) = \mathbb{E}(X_n)$ and $V(n) = \mathbb{V}(X_n)$.

(a) Show that $M(n+1) = \mu M(n)$ and $V(n+1) = \sigma^2 M(n) + \mu^2 V(n)$.

(b) Show that $M(n) = \mu^n$ and that $V(n) = \sigma^2 \mu^{n-1}(1 + \mu + \cdots + \mu^{n-1})$.

(c) What happens to the variance if $\mu > 1$? What happens to the variance if $\mu = 1$? What happens to the variance if $\mu < 1$?

(d) The population goes extinct if $X_n = 0$ for some n. Let us thus define the extinction time N by

$$N = \min\{n : \ X_n = 0\}.$$

Let $F(n) = \mathbb{P}(N \leq n)$ be the CDF of the random variable N. Show that

$$F(n) = \sum_{k=0}^{\infty} p_k (F(n-1))^k, \quad n = 1, 2, \ldots$$

Hint: Note that the event $\{N \leq n\}$ is the same as event $\{X_n = 0\}$. Thus, $\mathbb{P}(\{N \leq n\}) = \mathbb{P}(\{X_n = 0\})$. Let k be the number of offspring of the original parent. The population becomes extinct at time n if and only if each of the k sub-populations generated from the k offspring goes extinct in $n-1$ generations.

(e) Suppose that $p_0 = 1/4$, $p_1 = 1/2$, $p_2 = 1/4$. Use the formula from (5d) to compute the CDF $F(n)$.

6. Let

$$\mathbf{P} = \begin{bmatrix} 0.40 & 0.50 & 0.10 \\ 0.05 & 0.70 & 0.25 \\ 0.05 & 0.50 & 0.45 \end{bmatrix}$$

Find the stationary distribution π.

7. Show that if i is a recurrent state and $i \leftrightarrow j$, then j is a recurrent state.

8. Let
$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$
Which states are transient? Which states are recurrent?

9. Let
$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
Show that $\pi = (1/2, 1/2)$ is a stationary distribution. Does this chain converge? Why/why not?

10. Let $0 < p < 1$ and $q = 1 - p$. Let
$$\mathbf{P} = \begin{bmatrix} q & p & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ q & 0 & 0 & p & 0 \\ q & 0 & 0 & 0 & p \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$
Find the limiting distribution of the chain.

11. Let $X(t)$ be an inhomogeneous Poisson process with intensity function $\lambda(t) > 0$. Let $\Lambda(t) = \int_0^t \lambda(u)du$. Define $Y(s) = X(t)$ where $s = \Lambda(t)$. Show that $Y(s)$ is a homogeneous Poisson process with intensity $\lambda = 1$.

12. Let $X(t)$ be a Poisson process with intensity λ. Find the conditional distribution of $X(t)$ given that $X(t + s) = n$.

13. Let $X(t)$ be a Poisson process with intensity λ. Find the probability that $X(t)$ is odd, i.e. $\mathbb{P}(X(t) = 1, 3, 5, \ldots)$.

14. Suppose that people logging in to the University computer system is described by a Poisson process $X(t)$ with intensity λ. Assume that a person stays logged in for some random time with CDF G. Assume these times are all independent. Let $Y(t)$ be the number of people on the system at time t. Find the distribution of $Y(t)$.

15. Let $X(t)$ be a Poisson process with intensity λ. Let $W_1, W_2, \ldots,$ be the waiting times. Let f be an arbitrary function. Show that
$$\mathbb{E}\left(\sum_{i=1}^{X(t)} f(W_i)\right) = \lambda \int_0^t f(w)dw.$$

16. A two-dimensional Poisson point process is a process of random points on the plane such that (i) for any set A, the number of points falling in A is Poisson with mean $\lambda\mu(A)$ where $\mu(A)$ is the area of A, (ii) the number of events in non-overlapping regions is independent. Consider an arbitrary point x_0 in the plane. Let X denote the distance from x_0 to the nearest random point. Show that

$$\mathbb{P}(X > t) = e^{-\lambda\pi t^2}$$

and

$$\mathbb{E}(X) = \frac{1}{2\sqrt{\lambda}}.$$