Laws of Large Numbers and Stochastic Processes

Contents

Download PDF here

1.1. Introduction

We shall interpret economic time series as temporally organized data whose source is a single random draw from a parameterized joint probability distribution. That joint distribution determines intertermporal probabilistic relationships among components of the data. We imagine a statistician who doesn't know values of the parameters that characterize the joint probability distribution and wants to use those data to infer them. A key ingredient in doing this successfully is somehow to form averages over time of some functions of the data and hope that they converge to something informative about model parameters. Some laws of large numbers (LLNs) can help us with this, but others can't. This chapter describes one that can.

Classic LLNs that are typically studied in entry level probability classes adopt a narrow perspective that is not applicable to economic time series. To help us appreciate why a plain vanilla LLN won't help us study time series, we'll start with a setting in which such a plain vanilla LLN is all that we need. Here the statistician views

a data set as a random draw from a particular joint probability distribution that is one among a family of models that he knows. Once again, from those data, our statistician wants to infer which member of the family is most plausible. The family could be finite or it could be represented more generally with an unknown parameter vector indexing alternative models. The parameter vector could be finite or infinite-dimensional. We call each joint probability distribution associated with a particular parameter vector a **statistical model**.

A plain vanilla LLN assumes that the data are a sequence of **independent** draws from the same (i.e., **identical**) probability distribution. This assumption immediately leads to a logical problem when it comes to thinking about **learning** which model has generated the data. If the parameter vector that pins down the probability distribution is **known** and draws are independent and identically distributed (IID), then past draws indicate nothing about future draws, meaning that there is nothing to learn. But if the parameters that pin down data generation are **not known**, then there is something to learn: a sequence of draws from that unknown distribution are not **IID** because past draws indicate something about probabilities of future draws.

Fig. 1.1 Top: a sequence of draws from a fair coin. Bottom: a sequence of draws from a possibly unfair brass tack.

To illustrate the issue, Fig. 1.1 shows IID draws from two Bernoulli distributions. The top sequence are flips of fair coins. Here the probability of a heads is one half for every coin. After observing the outcome of N coin

flips, a statistician would predict that the probability that the next flip will be one half, regardless of earlier outcomes. In this example, the statistician has no reason to use data to **learn** the "parameter vector of interest", i.e., the probability of a heads on a single draw of one coin, because he **knows** it. The bottom sequence is formed from successive tosses of brass tacks. We assume that the statistician does not know the probability that each brass tack will land on its head. The statistician wants to use this observed outcome of N tosses to predict outcomes of future tack tosses. A presumption that observations of past tosses of the tack contain information about future tosses is tantamount to assuming that the sequence is **not IID**. This is a context in which it is natural to assume that the way brass tacks are constructed is the same (i.e., **identical**) for all of the brass tacks, and thus that the probabilities of tacks landing on their head are the same for each toss. If we condition on the common probability that a tack lands on its head, it makes sense to view the hypothetical observations as **conditionally IID**.

To achieve a theory of statistical learning, we can relax the IID assumption and instead assume a property called exchangeability. For a sequence of random draws to be exchangeable, it must be true that the joint probability distribution of any finite segment does not change if we rearrange individuals' positions in the sequence. In other words the order of draws does not matter when forming joint probabilities. If prospective outcomes of tack tosses can plausibly be viewed as being exchangeable, this opens the door to learning about the probability of a heads as we accumulate more observations. Exchangeable sequences are conditionally IID, where we interpret conditioning as being on a statistical model.[1] Exchangeable sequences obey a conditional version of a Law of Large Numbers that we describe later in this chapter.

Because we are interested in economic time series, LLNs based on exchangeabilty are too restrictive because we are interested in probability models in which the temporal order in which observations arise matters. This motivates us to replace **exchangeability** with a notion of **stationarity** and to use an **ergodic decomposition theorem** for stationary processes. This LLN for stationary processes is conditional on what we call the statistical model. The resulting LLN allows us to study a statistician who considers several alternative statistical models simultaneously and who understands how long-run averages of some salient statistics depend on unknown parameters on which the LLN is conditioned.

A LLN can teach us about statistical challenges that we would face us even if we were to have time series of infinite length. While this is a good starting point, we'll have to do more. In later chapters, we will describe additional approaches to statistical inference that can help us to understand how much we can learn about model parameters from finite histories of data.

1.2. Stochastic Processes

We start with a probability space, namely, a triple $(\Omega, \mathfrak{F}, \Pr)$, where Ω is a set of sample points, $\mathfrak{F}$ is a collection of subsets of Ω called events (formally a sigma algebra) and $\Pr$ assigns probabilities to events. We refer to $\Pr$ as a probability measure. The following definition makes reference to Borel sets. Borel sets include open sets, closed sets, finite intersections, and countable unions of such sets.

Definition 1.1

X is an n-dimensional random vector if $X : \Omega \to \mathbb{R}^n$ has the property that for any Borel set $\mathfrak{b}$ in $\mathbb{R}^n$,

$$\{\omega \in \Omega : X(\omega) \in \mathfrak{b}\}$$

is in $\mathfrak{F}$.

A result from measure theory states that if $\{X \in \mathfrak{o}\} \stackrel{\text{def}}{=} \{\omega \in \Omega : X(\omega) \in \mathfrak{o}$ is an event in $\mathfrak{F}$ whenever $\mathfrak{o}$ is an open set in $\mathbb{R}^n$, then X is an n-dimensional random vector. In what follows, we will often omit the explicit reference to ω when it self evident.

This formal structure facilitates using mathematical analysis to formulate problems in probability theory. A random vector induces a probability distribution over the collection of Borel sets in which the probability assigned to set $\mathfrak{b}$ is given by

$$\Pr\{X \in \mathfrak{b}\}$$

By changing the set $\mathfrak{b}$, we trace out a probability distribution implied by the random vector X that is called the *induced distribution*. An induced distribution is what typically interests an applied worker. In practice, an induced distribution is just specified directly without constructing the foundations under study here. However, proceeding at a deeper level, as we have, by defining a random vector to be a function that satisfies particular measurable properties and imposing the probability measure $\Pr$ over the domain of that function has mathematical payoffs. We will exploit this mathematical formalism in various ways, among them being in construction of stochastic processes.

Definition 1.2

An n-dimensional stochastic process is an infinite sequence of n-dimensional random vectors $\{X_t : t = 0, 1, \dots\}$.

The measure $\Pr$ assigns probabilities to a rich and interesting collection of events. For example, consider a stacked random vector

$$X^{[\ell]}(\omega) \stackrel{\text{def}}{=} \begin{bmatrix} X_0(\omega) \\ X_1(\omega) \\ \vdots \\ X_\ell(\omega) \end{bmatrix}$$

and Borel sets $\mathfrak{b}$ in $\mathbb{R}^{n(\ell+1)}$. The joint distribution of $X^{[\ell]}$ induced by $\Pr$ over such Borel sets is

$$\Pr\{X^{[\ell]} \in \mathfrak{b}\}.$$

Since the choice of ℓ is arbitrary, $\Pr$ implies a distribution over a an entire sequence of random vectors $\{X_t(\omega) : t = 0, 1, \dots\}$.[2].

We may also go the other way. Given a probability distribution over infinite sequences of vectors, we can construct a probability space and a stochastic process that induce this distribution. Thus, the following way to construct a probability space is particularly enlightening.

Construction 1.1:

Let Ω be a collection of infinite sequences in $\mathbb{R}^n$ with a sample point $\omega \in \Omega$ being a sequence of vectors $\omega = (\mathbf{r}_0, \mathbf{r}_1, \dots)$, where $\mathbf{r}_t \in \mathbb{R}^n$. Let $\mathfrak{B}_t$ be the collection of Borel sets of $\mathbb{R}^{n(t+1)}$, and let $\mathfrak{F}$ be the smallest sigma-algebra that contains the Borel sets of $\mathbb{R}^{n(\ell+1)}$ for $\ell = 0, 1, 2.. ..$

For each integer $\ell \geq 0$, let $\Pr_\ell$ assign probabilities to the Borel sets of $\mathbb{R}^{n(\ell+1)}$. A Borel set in $\mathbb{R}^{n(\ell+1)}$ can also be viewed as a Borel set in $\mathbb{R}^{n(\ell+2)}$ with $\mathbf{r}_{n(\ell+1)}$ left unrestricted. Specifically, let $\mathfrak{b}_\ell$ be a Borel set in $\mathbb{R}^{n(\ell+1)}$. Then

$$\mathfrak{b}_\ell^{\ell+1} = \{(\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_\ell, \mathbf{r}_{\ell+1}) : (\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_\ell) \in \mathfrak{b}_\ell\}$$

is a Borel set in $\mathbb{R}^{n(\ell+2)}$. For probability measures $\{\Pr_\ell : \ell = 0, 1, \dots\}$ to be consistent, we require that the probability assigned by $\Pr_{\ell+1}$ satisfy

$$\Pr_\ell(\mathfrak{b}_\ell) = \Pr_{\ell+1}\left(\mathfrak{b}_\ell^{\ell+1}\right)$$

With this restriction, we can extend the probability $\Pr$ to the space $(\Omega, \mathfrak{F})$ that is itself consistent with the probability assigned by $\Pr_\ell$ for all nonnegative integers ℓ.[3]

Finally, we construct the stochastic process $\{X_t : t = 0, 1, \dots\}$ by letting

$$X_t(\omega) = \mathbf{r}_t$$

for $t = 0, 1, 2, \dots$. A convenient feature of this construction is that $\Pr_\ell$ is the probability induced by the random vector $\left[X_0', X_1', \dots, X_\ell'\right]'$.

We refer to this construction as *canonical*. While this is only one among other possible constructions of probability spaces, it illustrates the flexibility in building sequences of random vectors that induce alternative probabilities of interest.

The remainder of this chapter is devoted to studying Laws of Large Numbers. What is perhaps the most familiar Law of Large Numbers presumes that the stochastic process $\{X_t : t = 0, 1, \dots\}$ is IID. Then

$$\frac{1}{N}\sum_{t=1}^{N}\phi(X_t) \to E\phi(X_0)$$

for any (Borel measurable) function ϕ for which the expectation is well defined. Convergence holds in several senses that we state later. Notice that as we vary the function ϕ we can infer the (induced) probability distribution for X_0. In this sense, the outcome of the Law of Large Numbers under an IID sequence determines what we will call a *statistical model*.

For our purposes, an IID version of the Law of Large Numbers is too restrictive. First, we are interested in economic dynamics in which model outcomes are temporally dependent. Second, we want to put ourselves in the situation of a statistician who does not know *a priori* what the underlying data generating process is and therefore entertains multiple models. We will present a Law of Large Numbers that covers both settings.

1.3. Representing a Stochastic Process

We now generalize the canonical construction 1.1 of a stochastic process in a way that facilitates stating the Law of Large Numbers that interests us.

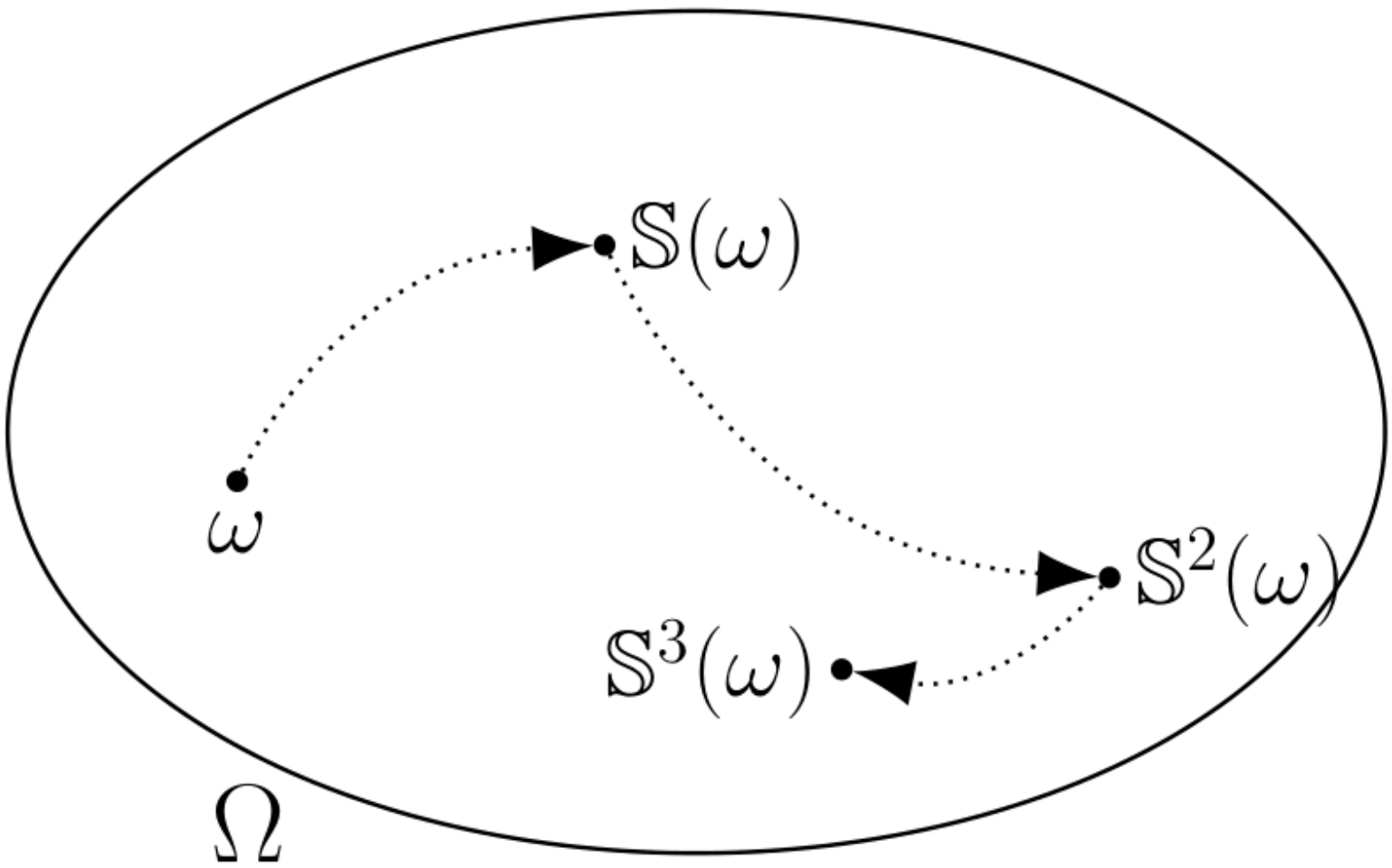

Fig. 1.2 The evolution of a sample point ω induced by successive applications of the transformation $\mathbb{S}$. The oval shaped region is the collection Ω of all sample points.

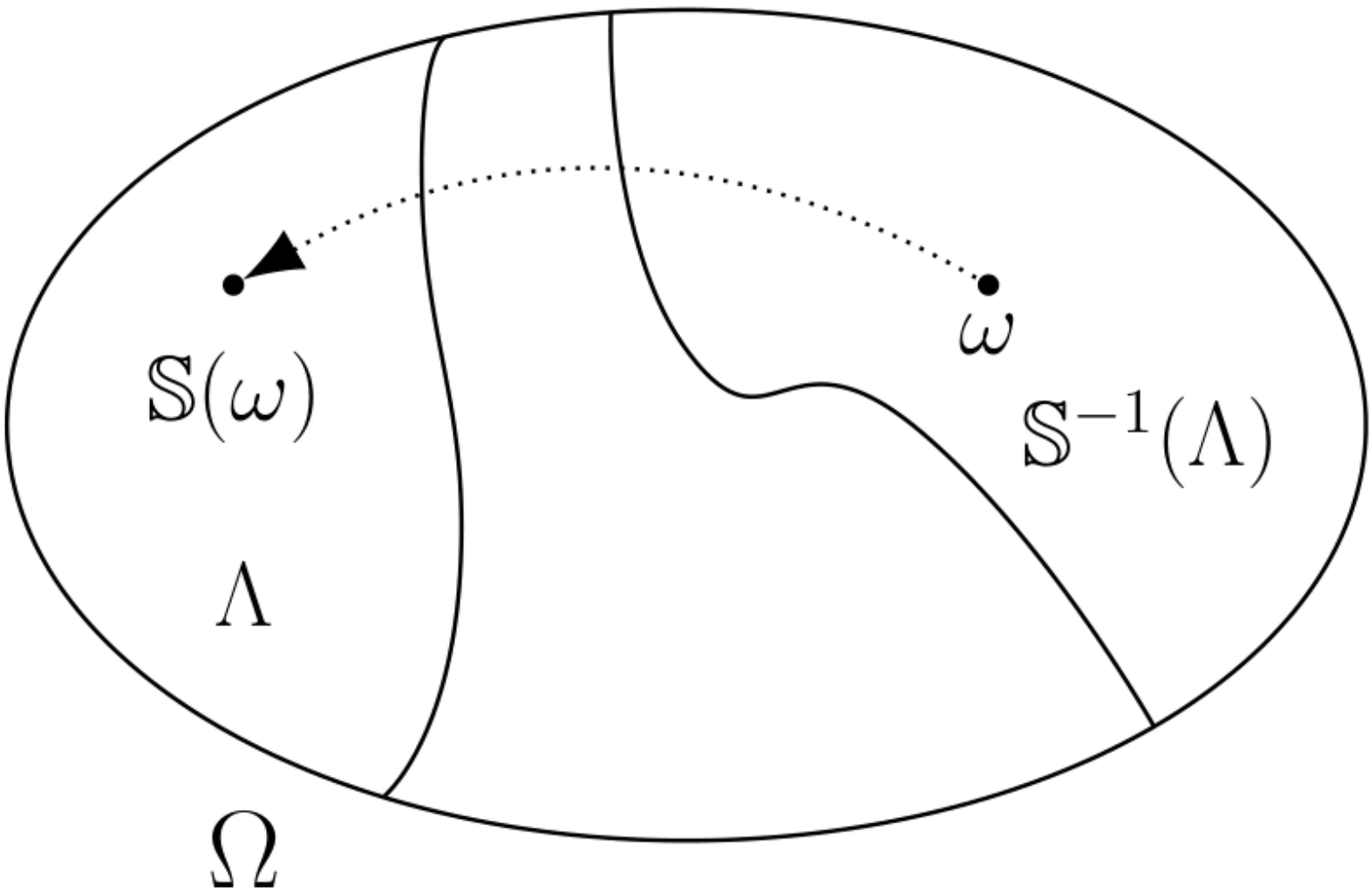

Fig. 1.3 An inverse image $\mathbb{S}^{-1}(\Lambda)$ of an event Λ is itself an event; $\omega \in \mathbb{S}^{-1}(\Lambda)$ implies that $\mathbb{S}(\omega) \in \Lambda$.

We use two objects.[4]

The first is a (measurable) transformation $\mathbb{S} : \Omega \to \Omega$ that describes the evolution of a sample point ω. See Fig. 1.2. Transformation $\mathbb{S}$ has the property that for any event $\Lambda \in \mathfrak{F}$,

$$\mathbb{S}^{-1}(\Lambda) = \{\omega \in \Omega : \mathbb{S}(\omega) \in \Lambda\}$$

is an event in $\mathfrak{F}$, as depicted in Fig. 1.3. The second object is an n-dimensional vector $X(\omega)$ that describes how observations depend on sample point ω.

We construct a stochastic process $\{X_t : t = 0, 1, \dots\}$ via the formula:

$$X_t(\omega) = X[\mathbb{S}^t(\omega)]$$

or

$$X_t = X \circ \mathbb{S}^t,$$

where we interpret $\mathbb{S}^0$ as the identity mapping asserting that $\omega_0 = \omega$.

Because a known function $\mathbb{S}$ maps a sample point $\omega \in \Omega$ today into a sample point $\mathbb{S}(\omega) \in \Omega$ tomorrow, the evolution of sample points is *deterministic*. For instance, $\omega_{t+j} = \mathbb{S}^{t+j}(\omega)$ for all $j \geq 1$ can be predicted perfectly if we know $\mathbb{S}$ and ω_t. But we typically do not observe ω_t at any t. Instead, we observe an $(n \times 1)$

vector $X(\omega)$ that contains incomplete information about ω. We assign probabilities $\Pr$ to collections of sample points ω called events, then use the functions $\mathbb{S}$ and X to induce a joint probability distribution over sequences of X's. The resulting stochastic process $\{X_t : t = 0, 1, 2, \dots\}$ is a sequence of n-dimensional random vectors.

This way of constructing a stochastic process might seem restrictive; but actually, it is more general than the canonical construction presented above.

Example 1.1

Consider again our canonical construction 1.1. Recall that the set of sample points Ω is the collection of infinite sequences of elements $\mathbf{r}_t \in \mathbb{R}^n$ so that $\omega = (\mathbf{r}_0, \mathbf{r}_1, \dots)$. For this example, $\mathbb{S}(\omega) = (\mathbf{r}_1, \mathbf{r}_2, \dots)$. This choice of $\mathbb{S}$ is called the *shift* transformation. Notice that the time t iterate is

$$\mathbb{S}^t(\omega) = (\mathbf{r}_t, \mathbf{r}_{t+1}, \dots)$$

Let the measurement function be: $X(\omega) = \mathbf{r}_0$ so that

$$X_t(\omega) = X\left[\mathbb{S}^t(\omega)\right] = \mathbf{r}_t$$

as posited in construction 1.1.

1.4. Stationary Stochastic Processes

We start with a probabilistic notion of invariance. We call a stochastic process *stationary* if for any finite integer ℓ, the joint probability distribution induced by the composite random vector $\left[X_t', X_{t+1}', \dots, X_{t+\ell}'\right]'$ is the same for all $t \geq 0$.[5] This notion of stationarity can be thought of as a stochastic version of a steady state of a dynamical system.

We now use the objects $(\mathbb{S}, X)$ to build a stationary stochastic process by restricting construction 1.1. Consider the set $\{\omega \in \Omega : X(\omega) \in \mathfrak{b}\} \overset{\text{def}}{=} \Lambda$ and its successors

$$\begin{aligned}\{\omega \in \Omega : X_1(\omega) \in \mathfrak{b}\} &= \{\omega \in \Omega : X\left[\mathbb{S}(\omega)\right] \in \mathfrak{b}\} = \mathbb{S}^{-1}(\Lambda) \\ \{\omega \in \Omega : X_t(\omega) \in \mathfrak{b}\} &= \{\omega \in \Omega : X\left[\mathbb{S}^t(\omega)\right] \in \mathfrak{b}\} = \mathbb{S}^{-t}(\Lambda).\end{aligned}$$

Evidently, if $\Pr(\Lambda) = \Pr[\mathbb{S}^{-1}(\Lambda)]$ for all $\Lambda \in \mathfrak{F}$, then the probability distribution induced by X_t equals the probability distribution of X for all t. This fact motivates the following definition and proposition.

Definition 1.3

The pair $(\mathbb{S}, \Pr)$ is said to be **measure-preserving** if

$$\Pr(\Lambda) = \Pr\{\mathbb{S}^{-1}(\Lambda)\}$$

for all $\Lambda \in \mathfrak{F}$.

Theorem 1.1

When $(\mathbb{S}, \Pr)$ is measure-preserving, probability distributions induced by the random vectors $X_t \stackrel{\text{def}}{=} X[\mathbb{S}^t(\omega)]$ are identical for all $t \geq 0$.

The measure-preserving property restricts the probability measure $\Pr$ for a given transformation $\mathbb{S}$. Some probability measures $\Pr$ used in conjunction with $\mathbb{S}$ will be measure-preserving and others not, a fact that will play an important role at several places below.

Suppose that $(\mathbb{S}, \Pr)$ is measure-preserving relative to probability measure $\Pr$. Given X and an integer $\ell > 1$, form a vector

$$X^{[\ell]}(\omega) \stackrel{\text{def}}{=} \begin{bmatrix} X_0(\omega) \\ X_1(\omega) \\ \cdots \\ X_\ell(\omega) \end{bmatrix}.$$

We can apply Theorem 1.1 to $X^{[\ell]}$ to conclude that the joint distribution function of $(X_t, X_{t+1}, \ldots, X_{t+\ell})$ is independent of t for $t = 0, 1, \ldots$. That this property holds for any choice of ℓ implies that the stochastic process $\{X_t : t = 1, 2, \ldots\}$ is stationary. Moreover, $f\left(X^{[\ell]}\right)$ where f is a Borel measurable function from $\mathbb{R}^{n(\ell+1)}$ into $\mathbb{R}$ is also a valid measurement function. Such f's include indicator functions of interesting events defined in terms of $X^{[\ell]}$.

For a given $\mathbb{S}$, we now present examples that illustrate how to construct a probability measure $\Pr$ that makes $\mathbb{S}$ measure-preserving and thereby brings stationarity.

Example 1.2

Suppose that Ω contains two points, $\Omega = \{\omega_1, \omega_2\}$. Consider a transformation $\mathbb{S}$ that maps ω_1 into ω_2 and ω_2 into ω_1: $\mathbb{S}(\omega_1) = \omega_2$ and $\mathbb{S}(\omega_2) = \omega_1$. Since $\mathbb{S}^{-1}(\{\omega_2\}) = \{\omega_1\}$ and $\mathbb{S}^{-1}(\{\omega_1\}) = \{\omega_2\}$, for $\mathbb{S}$ to be measure-preserving, we must have $\Pr(\{\omega_1\}) = \Pr(\{\omega_2\}) = 1/2$.

Example 1.3

Suppose that Ω contains two points, $\Omega = \{\omega_1, \omega_2\}$ and that $\mathbb{S}(\omega_1) = \omega_1$ and $\mathbb{S}(\omega_2) = \omega_2$. Since $\mathbb{S}^{-1}(\{\omega_2\}) = \{\omega_2\}$ and $\mathbb{S}^{-1}(\{\omega_1\}) = \{\omega_1\}$, $\mathbb{S}$ is measure-preserving for any $\Pr$ that satisfies $\Pr(\{\omega_1\}) \geq 0$ and $\Pr(\{\omega_2\}) = 1 - \Pr(\{\omega_1\})$.

Example 1.4

Suppose that Ω contains four points, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Moreover, $\mathbb{S}(\omega_1) = \omega_2$, $\mathbb{S}(\omega_2) = \omega_1$, $\mathbb{S}(\omega_3) = \omega_1$, and $\mathbb{S}(\omega_4) = \omega_2$. Notice that the sample points ω_3 and ω_4 are *transient*. Applying the $\mathbb{S}$ transformation does not allow for access to these points as they are not in image of $\mathbb{S}$. As a consequence, the only measure-preserving probability is the same one described in Example 1.2.

The next example illustrates how to represent an i.i.d. sequence of zeros and ones in terms of an Ω, $\Pr$ and an $\mathbb{S}$.

Example 1.5

Suppose that $\Omega = [0, 1)$ and that $\Pr$ is the uniform measure on $[0, 1)$. Let

$$\mathbb{S}(\omega) = \begin{cases} 2\omega & \omega \in [0, 1/2) \\ 2\omega - 1 & \omega \in [1/2, 1), \end{cases}$$

$$X(\omega) = \begin{cases} 1 & \omega \in [0, 1/2) \\ 0 & \omega \in [1/2, 1). \end{cases}$$

Calculate $\Pr\{X_1 = 1 | X_0 = 1\} = \Pr\{X_1 = 1 | X_0 = 0\} = \Pr\{X_1 = 1\} = 1/2$ and $\Pr\{X_1 = 0 | X_0 = 1\} = \Pr\{X_1 = 0 | X_0 = 0\} = \Pr\{X_1 = 0\} = 1/2$. So X_1 is statistically independent of X_0. By extending these calculations, it can be verified that $\{X_t : t = 0, 1, \dots\}$ is a sequence of independent random variables.[6] We can alter $\Pr$ to obtain other stationary distributions. For instance, suppose that $\Pr\{\frac{1}{3}\} = \Pr\{\frac{2}{3}\} = .5$. Then the process $\{X_t : t = 0, 1, \dots\}$ alternates in a deterministic fashion between zero and one. This provides a version of Example 1.2 in which $\omega_1 = \frac{1}{3}$ and $\omega_2 = \frac{2}{3}$.

1.5. Invariant Events and Conditional Expectations

In this section, we present a Law of Large Numbers that asserts that time series averages converge when $\mathbb{S}$ is measure-preserving relative to $\Pr$.

1.5.1. Invariant events

We use the concept of an invariant event to understand how limit points of time series averages relate to a conditional mathematical expectation.

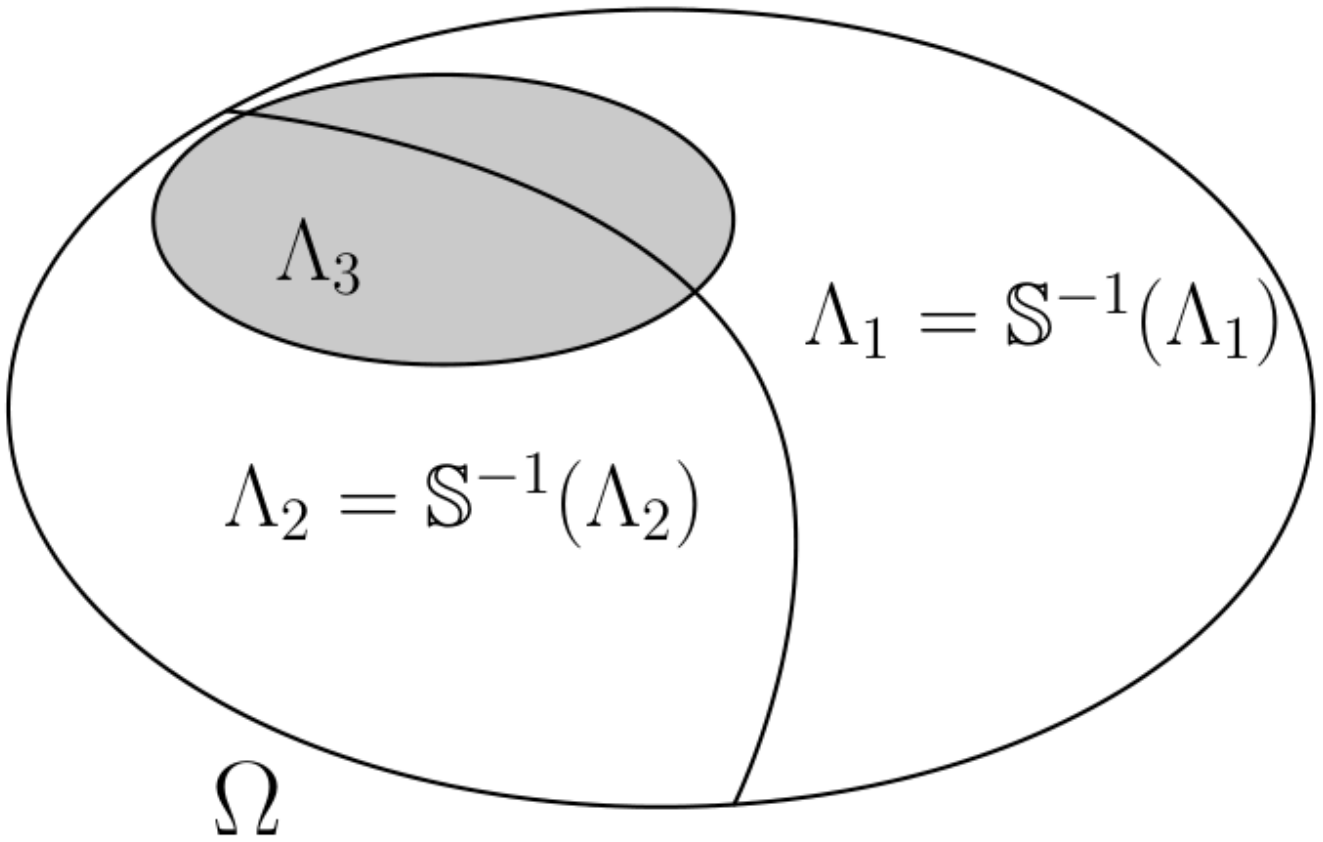

Fig. 1.4 Two invariant events Λ_1 and Λ_2 and an event Λ_3 that is not invariant.

Definition 1.4

An event Λ is **invariant** if $\Lambda = \mathbb{S}^{-1}(\Lambda)$.

Fig. 1.4 illustrates two invariant events in a space Ω. Notice that if Λ is an invariant event and $\omega \in \Lambda$, then $\mathbb{S}^t(\omega) \in \Lambda$ for $t = 0, 1, \ldots, \infty$. Thus under the transformation $\mathbb{S}$, sample points that are in Λ remain there. Furthermore, for each $\omega \in \Lambda$, there exists $\omega' \in \Lambda$ such that $\omega = \mathbb{S}(\omega')$.

Let $\mathfrak{I}$ denote the collection of invariant events. The entire space Ω and the null set $\varnothing$ are both invariant events. Like $\mathfrak{F}$, the collection of invariant events $\mathfrak{I}$ is a sigma algebra.

1.5.2. Conditional expectation

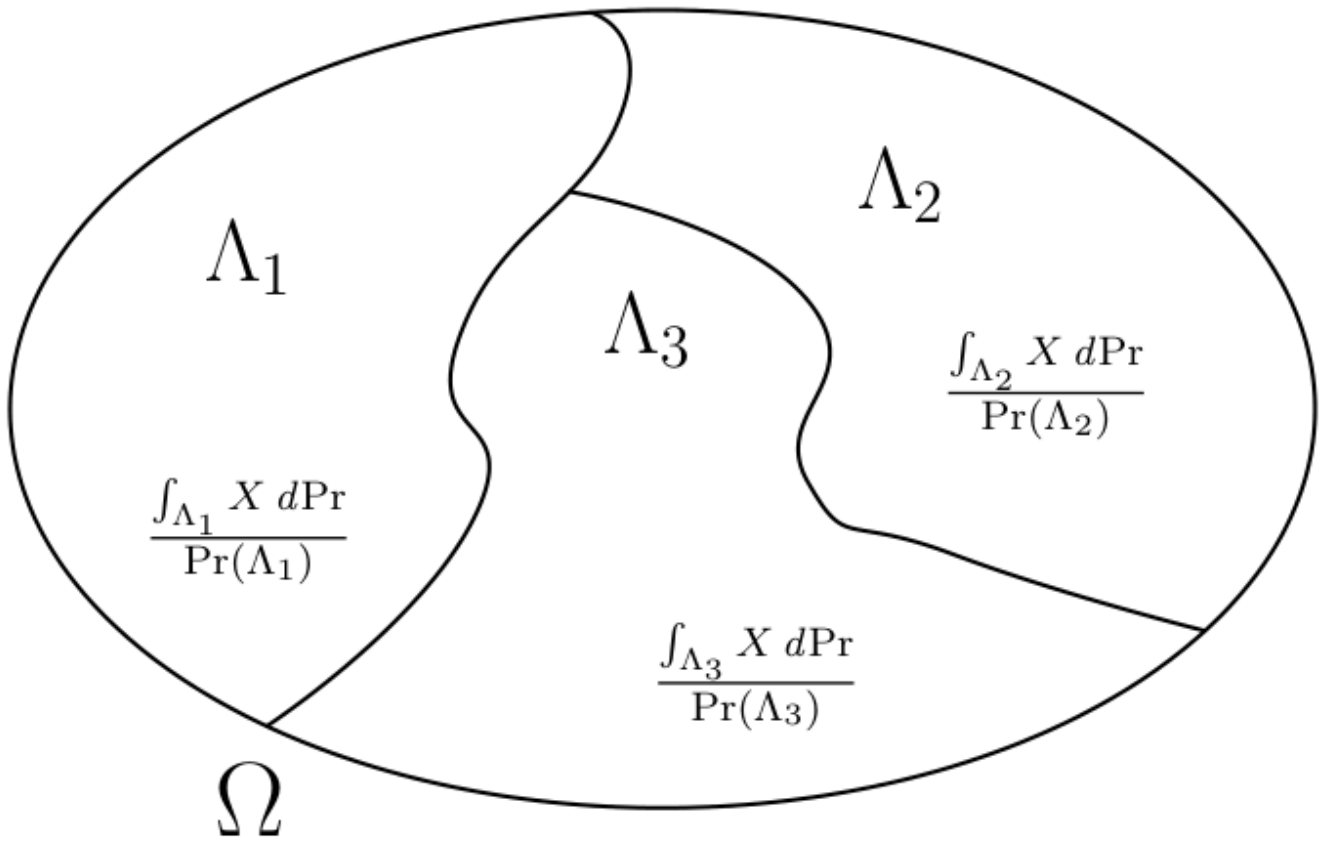

Fig. 1.5 A conditional expectation $E(X|\mathfrak{I})$ is constant for $\omega \in \Lambda_j = \mathbb{S}^{-1}(\Lambda_j)$.

We want to construct a random vector $E(X|\mathfrak{I})$ called the "mathematical expectation of X conditional on the collection $\mathfrak{I}$ of invariant events". We begin with a situation in which a conditional expectation is a discrete random vector as occurs when invariant events are unions of sets Λ_j belonging to a countable partition of Ω (together with the empty set). Later we'll extend the definition beyond this special setting.

A countable partition consists of a countable collection of nonempty events Λ_j such that $\Lambda_j \cap \Lambda_k = \varnothing$ for $j \neq k$ and such that the union of all Λ_j is Ω. Assume that each set Λ_j in the partition is itself an invariant event and has positive probability. Define the mathematical expectation conditioned on event Λ_j as

$$\frac{\int_{\Lambda_j} X d\Pr}{\Pr(\Lambda_j)}$$

when $\omega \in \Lambda_j$. To extend the definition of conditional expectation to all of $\mathfrak{I}$, take

$$E(X|\mathfrak{I})(\omega) = \frac{\int_{\Lambda_j} X d\Pr}{\Pr(\Lambda_j)} \text{ if } \omega \in \Lambda_j.$$

Thus, the conditional expectation $E(X|\mathfrak{I})$ is constant for $\omega \in \Lambda_j$ but varies across Λ_j's. Fig. 1.5 illustrates this characterization for a finite partition.

1.5.3. Least Squares

Now let X be a random vector with finite second moments $EXX' = \int X(\omega)X(\omega)'d\Pr(\omega)$. When a random vector X has finite second moments, a conditional expectation is a least squares projection. Let Z be an n-dimensional measurement function that is time-invariant and so satisfies

$$Z_t(\omega) = Z[\mathbb{S}^t(\omega)] = Z(\omega).$$

Let $\mathcal{Z}$ denote the collection of all such time-invariant random vectors. In the special case in which the invariant events can be constructed from a finite partition, Z can vary across sets Λ_j but must remain constant within Λ_j.[7] Consider the least squares problem

$$\min_{Z \in \mathcal{Z}} E[|X - Z|^2]. \tag{1.1}$$

Denote the minimizer in problem (1.1) by $\widetilde{X} = E(X|\mathfrak{I})$. Necessary conditions for the least squares minimizer $\widetilde{X} \in \mathcal{Z}$ imply that

$$E[(X - \widetilde{X})Z'] = 0$$

for Z in $\mathcal{Z}$ so that each entry of the vector $X - \widetilde{X}$ of regression errors is orthogonal to every vector Z in $\mathcal{Z}$.

A measure-theoretic approach constructs a conditional expectation by extending the orthogonality property of least squares. Provided that $E|X| < \infty$, $E(X|\mathfrak{I})(\omega)$ is the essentially unique random vector that, for any invariant event Λ, satisfies

$$E([X - E(X|\mathfrak{I})]\mathbf{1}_\Lambda) = 0,$$

where $\mathbf{1}_\Lambda$ is the indicator function that is equal to one on the set Λ and zero otherwise.

1.6. Law of Large Numbers

An elementary Law of Large Numbers asserts that the limit of an average over time of a sequence of independent and identically distributed random vectors equals the unconditional expectation of the random vector. We want a more general Law of Large Numbers that applies to averages over time of sequences of observations that are intertemporally dependent. To do this, we use a notion of probabilistic invariance that is expressed in terms of the measure-preserving restriction and that implies a Law of Large Numbers applicable to stochastic processes.

The following theorem asserts two senses in which averages of intertemporally dependent processes converge to mathematical expectations conditioned on invariant events.

Theorem 1.2

(Birkhoff) Suppose that $\mathbb{S}$ is measure-preserving relative to the probability space $(\Omega, \mathfrak{F}, \Pr)$.[8]

1. For any X such that $E|X| < \infty$,

$$\frac{1}{N}\sum_{t=1}^{N} X_t(\omega) \to E(X|\mathfrak{I})(\omega)$$

with probability one;

2. For any X such that $E|X|^2 < \infty$,

$$E\left[\left|\frac{1}{N}\sum_{t=1}^{N} X_t - E(X|\mathfrak{I})\right|^2\right] \to 0.$$

Part 1) asserts *almost-sure* convergence; part 2) asserts *mean-square* convergence.

We have ample flexibility to specify a measurement function $\phi\left(X^{\ell}\right)$, where ϕ is a Borel measurable function from $\mathbb{R}^{n(\ell+1)}$ into $\mathbb{R}$. In particular, an indicator functions for event $\Lambda = \{X^{\ell} \in \mathfrak{b}\}$ can be used as a measurement function where:

$$\phi(x^{\ell}) = \mathbf{1}_{\mathfrak{b}} = \begin{cases} 1 & \text{if } x^{\ell} \in \mathfrak{b} \\ 0 & \text{if } x^{\ell} \notin \mathfrak{b}. \end{cases}$$

where x^{ℓ} is a hypothetical realization of the random vector X^{ℓ}. The Law of Large Numbers applies to limits of

$$\frac{1}{N}\sum_{t=1}^{N} \phi\left[X_t^{\ell}\right]$$

for alternative ϕ's, so choosing ϕ's to be indicator functions shows how the Law of Large Numbers uncovers event probabilities of interest.

Definition 1.5

A transformation $\mathbb{S}$ that is measure-preserving relative to $\Pr$ is said to be **ergodic** under probability measure $\Pr$ if all invariant events have probability zero or one.

Thus, when a transformation $\mathbb{S}$ is *ergodic* under measure $\Pr$, the invariant events have either the same probability measure as the entire sample space Ω (whose probability measure is one), or the same probability measure as the empty set $\varnothing$ (whose probability measure is zero).

Proposition 1.1

Suppose that the measure-preserving transformation $\mathbb{S}$ is ergodic under measure $\Pr$. Then $E(X|\mathfrak{I}) = E(X)$.

Theorem 1.2 describes conditions for convergence in the general case that $\mathbb{S}$ is measure-preserving under $\Pr$, but in which $\mathbb{S}$ is not necessarily ergodic under $\Pr$. Proposition 1.1 describes a situation in which probabilities assigned to invariant events are degenerate in the sense that all invariant events have the same probability as either Ω (probability one) or the null set (probability zero). When $\mathbb{S}$ is *ergodic* under measure $\Pr$, limit points of time series averages equal corresponding unconditional expectations, an outcome we can call a *standard* Law of Large Numbers. When $\mathbb{S}$ is not ergodic under $\Pr$, limit points of time series averages equal expectations conditioned on invariant events.

The following examples remind us how ergodicity restricts $\mathbb{S}$ and $\Pr$.

Example 1.6

Consider Example 1.2 again.
Ω contains two points and $\mathbb{S}$ maps ω_1 into ω_2 and ω_2 into ω_1: $\mathbb{S}(\omega_1) = \omega_2$ and $\mathbb{S}(\omega_2) = \omega_1$.
Suppose that the measurement vector is

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \omega_1 \\ 0 & \text{if } \omega = \omega_2. \end{cases}$$

Then it follows directly from the specification of $\mathbb{S}$ that

$$\frac{1}{N}\sum_{t=1}^{N} X_t(\omega) \to \frac{1}{2}$$

for both values of ω. The limit point is the average across sample points.

Example 1.7

Return to Example 1.3. Ω contains two points, $\Omega = \{\omega_1, \omega_2\}$ and that $\mathbb{S}(\omega_1) = \omega_1$ and $\mathbb{S}(\omega_2) = \omega_2$. $X_t(\omega) = X(\omega)$ so that the sequence is time invariant and equal to its time-series average. A time-series average of $X_t(\omega)$ equals the average across sample points only when $\Pr$ assigns probability 1 to either ω_1 or ω_2.

1.7. Limiting Empirical Measures

Given a triple $(\Omega, \mathfrak{F}, \Pr)$ and a measure-preserving transformation $\mathbb{S}$, we can use Theorem 1.2 to construct *limiting empirical measures* on $\mathfrak{F}$. To start, we will analyze a setting with a countable partition of Ω consisting of invariant events $\{\Lambda_j : j = 1, 2, \dots\}$, each of which has strictly positive probability under $\Pr$. With the exception of the null set, we assume that all invariant events are unions of the members of this partitition. We consider a more general setting later. Given an event Λ in $\mathfrak{F}$ and for almost all $\omega \in \Lambda_j$, define the limiting empirical measure Qr_j as

$$\mathrm{Qr}_j(\Lambda)(\omega) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}_\Lambda \left[\mathbb{S}^t(\omega)\right] = \frac{\Pr(\Lambda \cap \Lambda_j)}{\Pr(\Lambda_j)}. \tag{1.2}$$

Thus, when $\omega \in \Lambda_j$, $\mathrm{Qr}_j(\Lambda)$ is the fraction of time $\mathbb{S}^t(\omega) \in \Lambda$ in very long samples. If we hold Λ_j fixed and let Λ be an arbitrary event in $\mathfrak{F}$, we can treat Qr_j as a probability measure on $(\Omega, \mathfrak{F})$. By doing this for each $\Lambda_j, j = 1, 2, \dots$, we can construct a countable set of probability measures $\{\mathrm{Qr}_j\}_{j=1}^{\infty}$. These comprise the set of all measures that can be recovered by applying the Law of Large Numbers. If nature draws an $\omega \in \Lambda_j$, then measure Qr_j describes outcomes.

So far, we started with a probability measure $\Pr$ and then constructed the set of possible limiting empirical measures Qr_j's. We now reverse the direction of the logic by starting with probability measures Qr_j and then finding measures $\Pr$ that are consistent with them. We do this because Qr_j's are the only measures that long time series can disclose through the Law of Large Numbers: each Qr_j defined by (1.2) uses the Law of Large Numbers to assign probabilities to events $\Lambda \in \mathfrak{F}$. However, because

$$\mathrm{Qr}_j(\Lambda) = \Pr(\Lambda \mid \Lambda_j) = \frac{\Pr(\Lambda \cap \Lambda_j)}{\Pr(\Lambda_j)} \text{ for } j = 1, 2, \dots,$$

are conditional probabilities, such Qr_j's are silent about the probabilities $\Pr(\Lambda_j)$ of the underlying invariant events Λ_j. There are multiple ways to assign probabilities $\Pr$ that imply identical probabilities conditioned on invariant events.

Because Qr_j is all that can ever be learned by "letting the data speak", we regard each probability measure Qr_j as a statistical model.[9]

Proposition 1.2

A *statistical model* is a probability measure that a Law of Large Numbers can disclose.

Probability measure Qr_j describes a statistical model associated with invariant set Λ_j.

Remark 1.1

For each j, $\mathbb{S}$ is measure-preserving and ergodic on $(\Omega, \mathfrak{F}, \mathrm{Qr}_j)$.
The second equality of definition (1.2) assures ergodicity by assigning probability one to the event Λ_j.

Relation (1.2) implies that probability Pr connects to probabilities Qr_j by

$$\mathrm{Pr}(\Lambda) = \sum_j \mathrm{Qr}_j(\Lambda)\mathrm{Pr}(\Lambda_j). \tag{1.3}$$

While decomposition (1.3) follows from definitions of the elementary objects that comprise a stochastic process and is "just mathematics", it is interesting because it tells how to construct alternative probability measures Pr for which $\mathbb{S}$ is measure-preserving. Because long data series disclose probabilities conditioned on invariant events to be Qr_j, to respect evidence from long time series we must hold the Qr_j's fixed, but we can freely assign probabilities Pr to invariant events Λ_j. In this way, we can create a family of probability measures for which $\mathbb{S}$ is measure-preserving.

1.8. Ergodic Decomposition

Up to now, we have represented invariant events with a countable partition. Dynkin [1978] deduced a more general version of decomposition (1.3) without assuming a countable partition. Thus, start with a pair $(\Omega, \mathfrak{F})$. Also, assume that there is a metric on Ω and that Ω is separable. We also assume that $\mathfrak{F}$ is the collection of Borel sets (the smallest sigma algebra containing the open sets). Given $(\Omega, \mathfrak{F})$, take a (measurable) transformation $\mathbb{S}$ and consider the set $\mathcal{P}$ of probability measures Pr for which $\mathbb{S}$ is measure-preserving. For some of these probability measures, $\mathbb{S}$ is ergodic, but for others, it is not. Let $\mathcal{Q}$ denote the set of probability measures for which $\mathbb{S}$ is ergodic. Under a nondegenerate convex combination of two probability measures in $\mathcal{Q}$, $\mathbb{S}$ is measure-preserving but *not* ergodic. Dynkin [1978] constructed limiting empirical measures Qr on $\mathcal{Q}$ and justified the following representation of the set $\mathcal{P}$ of probability measures Pr.

Proposition 1.3

For each probability measure $\widetilde{\mathrm{Pr}}$ in $\mathcal{P}$, there is a unique probability measure π over $\mathcal{Q}$ such that

$$\widetilde{\mathrm{Pr}}(\Lambda) = \int_{\mathcal{Q}} \mathrm{Qr}(\Lambda)\pi(d\mathrm{Qr}) \tag{1.4}$$

for all $\Lambda \in \mathfrak{F}$.[10]

Proposition 1.3 generalizes representation (1.3). It asserts a sense in which the set $\mathcal{P}$ of probabilities for which $\mathbb{S}$ is measure-preserving is convex. Extremal points of this set are in the smaller set $\mathcal{Q}$ of probability measures for which the transformation $\mathbb{S}$ is ergodic. Representation (1.4) shows that by forming "mixtures" (i.e., weighted averages or convex combinations) of probability measures under which $\mathbb{S}$ is ergodic, we can represent all probability specifications for which $\mathbb{S}$ is measure-preserving.

To add another perspective, a collection of invariant events $\mathfrak{I}$ is associated with a transformation $\mathbb{S}$. There exists a common conditional expectation operator $\mathbb{J} \equiv E(\cdot|\mathfrak{I})$ that assigns mathematical expectations to bounded measurable functions (mapping Ω into $\mathbb{R}$) conditioned on the set of invariant events $\mathfrak{I}$. The conditional expectation operator $\mathbb{J}$ characterizes limit points of time series averages of indicator functions of events of interest as well as other random vectors. Alternative probability measures $\Pr$ assign different probabilities to the invariant events.

1.9. Risk and Uncertainty

An applied researcher typically does not know which statistical model generated the data. This situation leads us to specifications of $\mathbb{S}$ that are consistent with a family $\mathcal{P}$ of probability models under which $\mathbb{S}$ is measure-preserving and a stochastic process is stationary. Representation (1.4) describes uncertainty about statistical models with a probability distribution π over the set of statistical models $\mathcal{Q}$.

For a Bayesian, π is a subjective prior probability distribution that pins down a convex combination of "statistical models."[11] A Bayesian expresses trust in that convex combination of statistical models used to construct a complete probability measure over outcomes[12] and uses it to compute expected utility. A Bayesian decision theory axiomatized by Savage makes no distinction between how decision makers respond to the probabilities described by the component statistical models and the π probabilities that he uses to mix them. All that matters to a Bayesian decision maker is the complete probability distribution over outcomes, not how it is attained as a π-mixture of component statistical models.

Some decision and control theorists challenge the complete confidence in a single prior probability assumed in a Bayesian approach.[13] They want to distinguish 'ambiguity', meaning not being able confidently to assign π, from 'risk', meaning prospective outcomes with probabilities reliably described by a statistical model. They imagine decision makers who want to evaluate decisions under alternative π's.[14] We explore these ideas in later chapters.

An important implication of the Law of Large Numbers is that for a given initial π, using Bayes' rule to update the π probabilities as data arrive will eventually concentrate posterior probability on the statistical model that generates the data. Even when a decision maker entertains a family of π's, the updated probabilities conditioned on the data may still concentrate on the statistical model that generates the data.

1.10. Estimating Vector Autoregressions

We now apply the Law of Large Numbers to the estimation of the equations in a vector autoregression

Let Y_{t+1} be one of the entries of X_{t+1}, and consider the regression equation:

$$Y_{t+1} = \beta \cdot X_t + U_{t+1},$$

where U_{t+1} is a least squares residual. By choosing Y_{t+1} to be alternative entries of X_{t+1}, we obtain the different equations in a VAR system. Our perspective in this discussion is that of an econometrician who fits such a regression system without taking a stand on the actual dynamic stochastic evolution of the $\{X_t : t = 0, 1, \dots\}$. To express subjective uncertainty about β, we allow it to be random but measurable in terms of the collection of invariant events $\mathfrak{I}$. As implied by least squares, we impose that the regression error, U_{t+1} is orthogonal to the vector X_t of regressors conditioned on $\mathfrak{I}$:

$$E\left(X_t U_{t+1} | \mathfrak{I}\right) = 0.$$

Then

$$E\left(X_t Y_{t+1} | \mathfrak{I}\right) = E\left[X_t (X_t)' | \mathfrak{I}\right] \beta, \tag{1.5}$$

which uniquely pins down the regression coefficient β provided that the matrix $E\left[X_t (X_t)' | \mathfrak{I}\right]$ is nonsingular with probability one. Notice that

$$\frac{1}{N} \sum_{t=1}^{N} X_t Y_{t+1} \to E\left(X_t Y_{t+1} | \mathfrak{I}\right)$$

$$\frac{1}{N} \sum_{t=1}^{N} X_t (X_t)' \to E\left(X_t (X_t)' | \mathfrak{I}\right),$$

where convergence is with probability one. Thus, from equation (1.5) it follows that a consistent estimator of β is a b_N that satisfies

$$\frac{1}{N} \sum_{t=1}^{N} X_t Y_{t+1} = \frac{1}{N} \sum_{t=1}^{N} X_t (X_t)' b_N.$$

Solving for b_N gives the familiar least squares formula:

$$b_N = \left[\sum_{t=1}^{N} X_t (X_t)'\right]^{-1} \sum_{t=1}^{N} X_t Y_{t+1}.$$

Note how statements about the consistency of b_N are conditioned on $\mathfrak{J}$. This conditioning is necessary when we do not know ex ante which among a family vector autoregressions generates the data.

1.11. Inventing an Infinite Past

When $\Pr$ is measure-preserving and the process $\{X_t : t = 0, 1, \dots\}$ is stationary, it can be useful to invent an infinite past. To accomplish this, we reason in terms of the (measurable) transformation $\mathbb{S} : \Omega \to \Omega$ that describes the evolution of a sample point ω. Until now we have assumed that $\mathbb{S}$ has the property that for any event $\Lambda \in \mathfrak{F}$,

$$\mathbb{S}^{-1}(\Lambda) = \{\omega \in \Omega : \mathbb{S}(\omega) \in \Lambda\}$$

is an event in $\mathfrak{F}$. In Section Stationary Increments, we want more. To prepare the way for that chapter, in this section we shall also assume that $\mathbb{S}$ is one-to-one and has the property that for any event $\Lambda \in \mathfrak{F}$,

$$\mathbb{S}(\Lambda) = \{\omega \in \Omega : \mathbb{S}^{-1}(\omega) \in \Lambda\} \in \mathfrak{F}. \tag{1.6}$$

Because

$$X_t(\omega) = X[\mathbb{S}^t(\omega)] = X_t = X \circ \mathbb{S}^t$$

is well defined for negative values of t, restrictions (1.6) allow us to construct a ``two-sided'' process that has both an infinite past and an infinite future.

Let $\mathfrak{A}$ be a subsigma algebra of $\mathfrak{F}$, and let

$$\mathfrak{A}_t = \left\{\Lambda_t \in \mathfrak{F} : \Lambda_t = \{\omega \in \Omega : \mathbb{S}^t(\omega) \in \Lambda\} \text{ for some } \Lambda \in \mathfrak{A}\right\}. \tag{1.7}$$

We assume that $\{\mathfrak{A}_t : -\infty < t < +\infty\}$ is nondecreasing sequence of subsigma algebras of $\mathfrak{F}$. The nondecreasing structure captures the information accumulation over time. If the original measurement function X is $\mathfrak{A}$-measurable, then X_t is $\mathfrak{A}_t$-measurable. Furthermore, X_{t-j} is in $\mathfrak{A}_t$ for all $j \geq 0$. The set $\mathfrak{A}_t$ depicts information available at date t, including past information. Invariant events in $\mathfrak{I}$ are contained in $\mathfrak{A}_t$ for all t.

We construct the following moving-average representation of a scalar process $\{X_t\}$ in terms of an infinite history of shocks.

Example 1.8

(Moving average) Suppose that $\{W_t : -\infty < t < \infty\}$ is a vector stationary process for which[15]

$$E\left(W_{t+1}|\mathfrak{A}_t\right) = 0$$

and that

$$E\left(W_t W_t{}'|\mathfrak{I}\right) = I$$

for all $-\infty < t < +\infty$.
Use a sequence of vectors $\{\alpha_j\}_{j=0}^{\infty}$ to construct

$$X_t = \sum_{j=0}^{\infty} \alpha_j \cdot W_{t-j} \tag{1.8}$$

where

$$\sum_{j=0}^{\infty} |\alpha_j|^2 < \infty. \tag{1.9}$$

Restriction (1.9) implies that X_t is well defined as a mean square limit. X_t is constructed from the infinite past $\{W_{t-j} : 0 \leq j < \infty\}$. The process $\{X_t : -\infty < t < \infty\}$ is stationary and is often called an infinite-order moving average process. The sequence $\{\alpha_j : j = 0, 1, \dots\}$ can depend on the invariant events.

Remark 1.2

Slutsky [1927] and Yule [1927] used probability models to analyze economic time series. Their models implied moving-average representations like the one in Example 1.8. Their idea was to view economic time series as responding linearly to current and past independent and identically distributed impulses or shocks. In distinct contributions, they showed how such models generate recurrent but aperiodic fluctuations that resemble business cycles and longer-term cycles as well. Yule [1927] and Slutsky [1927] came from different backgrounds and brought different perspectives. Yule [1927] was an eminent statistician who, among other important contributions, managed "effectively to invent modern time series analysis" in the words of Stigler [1986]. Yule constructed and estimated what we would now call a second-order autoregression and applied it to study sunspots. Yule's estimates implied α_j coefficients showed damped oscillations at the same periodicity as sunspots. In Russia in the 1920s, Slutsky [1927] wrote a seminal paper in Russian motivated by his interest in business cycles. Later an English version of his paper published in *Econometrica*. Even before that, it influenced economists including Ragnar Frisch. Indeed, Frisch was

> keenly aware of both Slutsky [1927] and Yule [1927] and generously acknowledged both of them in his seminal paper Frisch [1933] on the impulse and propagation problem. Building on insights of Slutsky [1927] and Yule [1927], Frisch [1933] pioneered impulse response functions. He aspired to provide explicit economic interpretations for how shocks alter economic time series intertemporally. [16]

1.12. Summary

For a fixed $\mathbb{S}$ there are often many possible probabilities $\Pr$ that are measure-preserving. A subset of these are ergodic. These ergodic probabilities can serve as building blocks for the other measure-preserving probabilities. Thus, each measure-preserving $\Pr$ can be expressed as a weighted average of the ergodic probabilities. We call the ergodic probabilities statistical models. The Law of Large Numbers applies to each of the ergodic building blocks with limit points that are unconditional expectations. As embodied in (1.3) and its generalization (1.4), this decomposition interests both frequentist and Bayesian statisticians.

---

[1] [de Finetti, 1937] established an important initial result showing that exchangeable sequences are conditionally IID, which is a central result in theory of subjective probability.

[2] By construction, probability by $X^{[\ell+1]}$ induced is consistent with that induced by $X^{[\ell]}$ in the formal sense of what is used in the well known Kolomogorov Extension Theorem.

[3] This essentially follows from the Kolomorov Extension Theorem or from Theorem 2.26 of Breiman [1968].

[4] Breiman [1968] is a good reference for these.

[5] Sometimes this property is called `strict stationarity' to distinguish it from weaker notions that require only that some moments of joint distributions be independent of time. What is variously called wide-sense or second-order or covariance stationarity requires only that first and second moments of joint distributions are independent of calendar time.

[6] This example is from Breiman [1968][p. 108].

[7] More generally, Z must be measurable with respect to $\mathfrak{I}$.

[8] See Breiman [1968] chapter 6 for extended discussions and proofs.

[9] Marschak [1953], Hurwicz [1962], Lucas [1976], and Sargent [1981] distinguished between structural econometric models and what we call statistical models. Structural econometric models are designed to forecast outcomes of hypothetical experiments that freeze some components of an economic environment and change others. A structural model accepts experiments that alter statistical models.

[10] Krylov and Bogolioubov [1937] provide an early statement of this result. Dynkin [1978] provides a more general formulation that nests this and other closely related results. His analysis includes a formalization of integration over the probability measures in $\mathcal{Q}$. Dynkin [1978] uses the resulting representation to draw connections between collections of invariant events and sets of sufficient statistics.

[11] This subsection is motivated in part by the intriguing discussions of von Plato [1982] and Cerreia-Vioglio *et al.* [2013].

[12] Here 'complete' can be taken to be synonymous with 'not conditioning on invariant events'.

[13] For example, see Hansen and Sargent [2008].

[14] This gives one way to formalize ideas of Knight [1921], who sought to distinguish risk from broader notions of uncertainty.

[15] An i.i.d.~sequence is just one example of such a $\{W_t : -\infty < t < \infty\}$ process.

[16] Sims [1980] and others advanced this idea by developing tractable multivariate time series methods and striving to isolate interpretable shocks in multivariate settings.