DISCRETE TIME STOCHASTIC PROCESSES

1. Introduction

A discrete-time stochastic process is essentially a random vector with components indexed by time, and a time series observed in an economic application is one realization of this random vector. Then, a useful way to introduce stochastic processes is to return to the basic development of the theory of random variables and vectors, and use the analogy as a guide to the statistical analysis needed in the more general stochastic setting.

Let **S** denote the universe of states of Nature. A state $s \in \mathbf{S}$ describes everything that has happened and will happen. In particular, this description includes the outcomes of all probability and sampling experiments. The family of potentially observable events in **S** is denoted by $\boldsymbol{F}$, a σ-*field* (or *Boolean* σ-*algebra*) of subsets of **S** satisfying

(i) The "anything can happen" event **S** is in $\boldsymbol{F}$.

(ii) If event **A** is in $\boldsymbol{F}$, then the event "not **A**", denoted $\mathbf{A}^c$ or $\mathbf{S}\backslash\mathbf{A}$, is in $\boldsymbol{F}$.

(iii) If **A** and **B** are events in $\boldsymbol{F}$, then the event "both **A** and **B**", denoted $\mathbf{A}\cap\mathbf{B}$, is in $\boldsymbol{F}$.

(iv) If $\mathbf{A}_1,\mathbf{A}_2,...$ is a finite or countable sequence of events in $\boldsymbol{F}$, then the event "one or more of the events $\mathbf{A}_1$ or $\mathbf{A}_2$ or ...", denoted $\bigcup_{i=1}^{\infty} \mathbf{A}_i$, is in $\boldsymbol{F}$.

Implications of the definition of a σ-field are

(v) If $\mathbf{A}_1,\mathbf{A}_2,...$ is a finite or countable sequence of events in $\boldsymbol{F}$, then $\bigcap_{i=1}^{\infty} \boldsymbol{A}_i$ is also in $\boldsymbol{F}$.

(vi) If $\mathbf{A}_1,\mathbf{A}_2,...$ is a countable sequence of events in $\boldsymbol{F}$ that is *monotone decreasing* (i.e., $\mathbf{A}_1 \supseteq \mathbf{A}_2 \supseteq ...$), then its limit, denoted $A_i \searrow A_0$, is also in $\boldsymbol{F}$. Similarly, if a sequence in $\boldsymbol{F}$ is *monotone increasing* (i.e., $\mathbf{A}_1 \subseteq \mathbf{A}_2 \subseteq ...$), then its limit $\mathbf{A}_0 = \bigcup_{i=1}^{\infty} \boldsymbol{A}_i$ is also in $\boldsymbol{F}$

(vii) The empty event $\boldsymbol{\varphi}$ is in $\boldsymbol{F}$.

A σ-field of subsets of **S** is sometimes identified with the *information* available to an observer. One method of constructing a σ-field of subsets of **S** is to start from a specified family $\boldsymbol{A}$ of subsets of **S**, such as the open intervals when **S** is the real line, and define $\boldsymbol{F}$ to be the intersection of all σ-fields containing the specified family; $\boldsymbol{F}$ is then said to be the σ-field *generated* by $\boldsymbol{A}$, and is sometimes denoted $\sigma(\boldsymbol{A})$. The idea is that if the observer knows which of the events in $\boldsymbol{A}$ occur, then he can also determine which of the events in $\sigma(\boldsymbol{A})$ occur. There may be more than one σ-field of subsets of **S**; these may correspond to the information available to different observers. If $\boldsymbol{F}$ and $\boldsymbol{G}$ are both σ-fields, and $\boldsymbol{G} \subseteq \boldsymbol{F}$, then $\boldsymbol{G}$ is said to be a *sub-field* of $\boldsymbol{F}$, and $\boldsymbol{F}$ is said to *contain more information* or *refine* $\boldsymbol{G}$. It is possible that neither $\boldsymbol{F} \subseteq \boldsymbol{G}$ nor $\boldsymbol{G} \subseteq \boldsymbol{F}$. The intersection $\boldsymbol{F}\cap\boldsymbol{G}$ of two σ-fields is again a σ-field that contains the *common information* in $\boldsymbol{F}$ and $\boldsymbol{G}$. Further, the intersection of an arbitrary countable or uncountable collection of σ-fields is again a σ-field. It is this property that guarantees that there is always a smallest σ-field $\sigma(\boldsymbol{A})$ containing any family $\boldsymbol{A}$ of subsets of **S**. The union $\boldsymbol{F}\cup\boldsymbol{G}$ of two σ-fields is not necessarily a σ-field, but there is always a smallest σ-field that refines both $\boldsymbol{F}$ and $\boldsymbol{G}$, which is simply the σ-field $\sigma(\boldsymbol{F}\cup\boldsymbol{G})$ generated by the sets in the union of $\boldsymbol{F}$ and $\boldsymbol{G}$, or put another way, the intersection of all σ-fields that contain both $\boldsymbol{F}$ and $\boldsymbol{G}$.

A pair (**S**,$\boldsymbol{\mathcal{F}}$) consisting of a set **S** and a σ-field $\boldsymbol{\mathcal{F}}$ of subsets of **S** is called a *measurable space*, and the sets in $\boldsymbol{\mathcal{F}}$ are called the *measurable* events. Then the universe of states of Nature with its σ-field is a measurable space, but the definition of a measurable space will be used much more generally, including the set of time indices and the image space of random functions.

Finally, P is a probability defined on the measurable events in the universe of states of Nature **S**; i.e., a function P:$\boldsymbol{\mathcal{F}} \rightarrow [0,1]$ with the properties

(i) $P(\mathbf{A}) \geq 0$ for all $\mathbf{A} \in \boldsymbol{\mathcal{F}}$.
(ii) $P(\mathbf{S}) = 1$.
(iii) [Countable Additivity] If $\mathbf{A}_1, \mathbf{A}_2, \ldots$ is a finite or countable sequence of events in $\boldsymbol{\mathcal{F}}$ that are mutually exclusive (i.e., $\mathbf{A}_i \cap \mathbf{A}_j = \boldsymbol{\varphi}$ for all $i \neq j$), then $P(\bigcup_{i=1}^{\infty} \mathbf{A}_i) = \sum_{i=1}^{\infty} P(\mathbf{A}_i)$.

With conditions (i)-(iii), P has the following additional intuitive properties of a probability when **A** and **B** are events in $\boldsymbol{\mathcal{F}}$:

(iv) $P(\mathbf{A}) + P(\mathbf{A}^c) = 1$.
(v) $P(\boldsymbol{\varphi}) = 0$.
(vi) $P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$.
(vii) $P(\mathbf{A}) \geq P(\mathbf{B})$ when $\mathbf{B} \subseteq \mathbf{A}$, and $P(\mathbf{B} \backslash \mathbf{A}) = P(\mathbf{B}) - P(\mathbf{A})$.
(viii) If $\mathbf{A}_i$ in $\boldsymbol{\mathcal{F}}$ is monotone decreasing to $\boldsymbol{\varphi}$ (denoted $\mathbf{A}_i \searrow \boldsymbol{\varphi}$), then $P(\mathbf{A}_i) \rightarrow 0$.
(ix) If $\mathbf{A}_i \in \boldsymbol{\mathcal{F}}$, not necessarily disjoint, then $P(\bigcup_{i=1}^{\infty} \mathbf{A}_i) \leq \sum_{i=1}^{\infty} P(\mathbf{A}_i)$.
(x) If $\{\mathbf{A}_i\}$ is a finite or countable *partition* of **S** (i.e., the events $\mathbf{A}_i \in \boldsymbol{\mathcal{F}}$ are mutually exclusive and exhaustive, or $\mathbf{A}_i \cap \mathbf{A}_j = \boldsymbol{\varphi}$ for all $i \neq j$ and $\bigcup_{i=1}^{\infty} \mathbf{A}_i = \mathbf{S}$), then $P(\mathbf{B}) = \sum_{i=1}^{\infty} P(\mathbf{B} \cap \mathbf{A}_i)$.

The triplet (**S**,$\boldsymbol{\mathcal{F}}$,P) consisting of a measurable space (**S**,$\boldsymbol{\mathcal{F}}$) and a probability measure P is called a *probability space*. If $\mathbf{A} \in \boldsymbol{\mathcal{F}}$ has $P(\mathbf{A}) = 1$, then **A** is said to occur *almost surely* (a.s.), or *with probability one* (w.p.1). If $\mathbf{A} \in \boldsymbol{\mathcal{F}}$ has $P(\mathbf{A}) = 0$, then **A** is said to occur with *probability zero* (w.p.0). Finite or countable intersections of events that occur almost surely again occur almost surely, and finite or countable unions of events that occur with probability zero again occur with probability zero. We noted earlier that σ-fields of subsets of **S** formalize the concept of the information known to an observer. There is an additional technical reason for introducing σ-fields. Starting from a probability with reasonable properties on selected events in **S**, one can always extend it to a probability on the σ-field generated by these selected events, but an extension to all the subsets of **S** may be impossible due to the existence of non-measurable sets whose probabilities are not well-defined. It is often useful for analysis of a probability space (**S**,$\boldsymbol{\mathcal{F}}$,P) to work with the *completion* $\boldsymbol{\mathcal{G}}$ the σ-field $\boldsymbol{\mathcal{F}}$; we obtain $\boldsymbol{\mathcal{G}}$ by augmenting $\boldsymbol{\mathcal{F}}$ with all sets **B** satisfying $\mathbf{B} \subseteq \mathbf{A}$ for some set $\mathbf{A} \in \boldsymbol{\mathcal{F}}$ that has $P(\mathbf{A}) = 0$. Since the additional sets **B** unambiguously have probability zero, their addition causes no problem with the definition of the probability.

Measurable spaces (**S**,$\boldsymbol{\mathcal{F}}$) will usually have an associated metric that measures distance between points in **S**, and open neighborhoods consisting of all the points in **S** less than a given distance from

each specified point. Often, the σ-field $\boldsymbol{F}$ will be defined as the smallest σ-field that contains all the open neighborhoods, and will be said to be *generated* by these neighborhoods. A space **S** with a distance metric is *complete* if every Cauchy sequence (i.e., a sequence of points in S such that the distance between any two points sufficiently far out in the sequence converges to zero) has a limit contained in **S**. Note that this is a different use of the word "complete" than the completion of a probability space. A metric space **S** is *separable* if it contains a countable dense subset; i.e., there is a countable subset of **S** such that each point in **S** can be obtained as a limit of points in the subset. An important measurable space is $(\mathbb{R},\boldsymbol{B})$, where $\mathbb{R}$ is the real line and $\boldsymbol{B}$ is the Borel σ-*field*, the smallest σ-*field* of subsets of $\mathbb{R}$ that contains all the open intervals in $\mathbb{R}$. The real line with the Euclidean distance metric is a complete separable metric space.

Often a measurable space $(\mathbf{S},\boldsymbol{F})$ will have an associated *measure* ν that is a countably additive function from $\boldsymbol{F}$ into the nonnegative real line; i.e., $\nu(\bigcup_{i=1}^{\infty} \mathbf{A}_i) = \sum_{i=1}^{\infty} \nu(\mathbf{A}_i)$ for any sequence of disjoint $\mathbf{A}_i \in \boldsymbol{F}$. The measure is *non-negative* if $\nu(\mathbf{A}) \geq 0$ for all $\mathbf{A} \in \boldsymbol{F}$; we will consider only non-negative measures. The measure ν is *finite* if $|\nu(\mathbf{S})| \leq M$, and σ-*finite* if $\boldsymbol{F}$ contains a countable partition $\{\mathbf{A}_i\}$ of **S** such that the measure of each partition set is finite. The measure ν may be a probability, but often is a measure of "length" or "volume". For example, it is common when **S** is the countable set of positive integers to define ν to be *counting measure* with $\nu(\mathbf{A})$ equal to the number of points in **A**. For the real line $(\mathbb{R},\boldsymbol{B})$, it is common to define ν to be *Lebesgue measure*, with $\nu((a,b)) = b - a$ for any open interval (a,b). Both counting measure and Lebesgue measure are non-negative σ-finite measures. A set **A** is said to be of *ν-measure zero* if $\nu(\mathbf{A}) = 0$. A property that holds except on a set of measure zero is said to hold *almost everywhere* (a.e.). A set $\mathbf{A} \in \boldsymbol{F}$ is called an *atom* if every set $\mathbf{B} \in \boldsymbol{F}$ with $\mathbf{B} \subset \mathbf{A}$ has either $\nu(\mathbf{B}) = 0$ or $\nu(\mathbf{B}) = \nu(\mathbf{A})$. Every non-negative measure can be decomposed into a countable sum of atoms plus a non-negative atomless measure. Given two measures ν and λ on $\boldsymbol{F}$, one says that λ is *continuous with respect to* ν if $\nu(A) = 0$ implies $\lambda(A) = 0$. Statistical analysis often uses σ-finite measure spaces $(\mathbf{S},\boldsymbol{F},\mu)$ where μ is non-negative and σ-finite and may either be a probability measure or a more general counting or length measure such as Lebesgue measure. Often a measurable space $(\mathbf{S},\boldsymbol{F})$ will have both a probability measure *P* and a "length" measure ν.

Suppose f is a real-valued function on a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$. This function is *measurable* if $f^{-1}(\mathbf{C}) \in \boldsymbol{F}$ for each open set **C** in the real line. A measurable function is *integrable* on a set $\mathbf{A} \in \boldsymbol{F}$ if $\int_{\mathbf{A}} |f(s)| \cdot \mu(ds)$, defined as the limit of $\sum_{k=0}^{n^2} (k/n) \cdot \mu(\{s \in \mathbf{A} | k/n \leq |f(s)| < (k+1)/n\})$, exists and is finite. If it is integrable, the integral $\int_{\mathbf{A}} f(s)\mu(ds)$ is sometimes denoted $\int_{\mathbf{A}} f(s)d\mu$, or in the case of Lebesgue measure, $\int_{\mathbf{A}} f(s)ds$. When $\mathbf{A} = \mathbf{S}$, the integral is often written as $\int f(s)\mu(ds)$. In general, the measure μ can have point masses (at atoms), or continuous measure, or both, so that the notation for integration with respect to μ includes sums and mixed cases. If a probability measure *P* is continuous with respect to Lebesgue measure ν, then there exists a measurable non-negative function g on the real line, called the *density* of *P*, such that $\int f(s)P(ds) = \int f(s)g(s)ds$; this result is called the Radon-Nikodym theorem (see Chap. 3, Theorem 3.2).

For a σ-finite measure space $(\mathbf{S},\boldsymbol{F},\mu)$, define $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $1 \le q < +\infty$ to be the set of measurable real-valued functions on **S** with the property that $|f|^q$ is integrable, and define the *norm* of f to be $\|f\|_q = [\int |f(s)|^q \, \mu(ds)]^{1/q}$. Then, $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ is a linear space, since linear combinations of integrable functions are again integrable. This space has many, but not all, of familiar properties of finite-dimensional Euclidean space. The set of all linear functions on the space $\mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ for $q > 1$ is the space $\mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu)$, where $1/r = 1 - 1/q$. This follows from an application of Holder's inequality, which generalizes from finite vector spaces to the condition

$f \in \mathbf{L}_q(\mathbf{S},\boldsymbol{F},\mu)$ and $g \in \mathbf{L}_r(\mathbf{S},\boldsymbol{F},\mu)$ with $q^{-1} + r^{-1} = 1$ imply $\int |f(s)\cdot g(s)| \, \mu(ds) \le \|f\|_q \cdot \|g\|_r$.

The case $q = r = 2$ gives the Cauchy-Schwartz inequality in general form. The space $\mathbf{L}_2(\mathbf{S},\boldsymbol{F},\mu)$, called *Hilbert space*, occurs often in statistics.

A *random variable* X is a measurable real-valued function on the probability space $(\mathbf{S},\boldsymbol{F},P)$, or $X:S \to \mathbb{R}$. Then each state of Nature s determines a value X(s) of the random variable, termed its *realization* in state s. When the functional nature of the random variable is to be emphasized, it is denoted X(·), or simply X. When its values or realizations are used, they are denoted X(s) or x. For each set $\mathbf{B} \in \boldsymbol{B}$, the probability of the event that the realization of X is contained in **B** is well-defined and equals $P'(\mathbf{B}) \equiv P(X^{-1}(\mathbf{B}))$, where *P'* is termed the probability *induced* on $\mathbb{R}$ by the random variable X. The probability that X is contained in a half-line $(-\infty,x]$ defines the CDF of X, which in turn characterizes *P'*:

$$F_X(x) = P'((-\infty,x]) \equiv P(X^{-1}((-\infty,x])) \equiv P(\{s\in S|X(s)\in(-\infty,x]\}).$$

Multiplying a random variable by a scalar, or adding random variables, results in another random variable. Then, the family of random variables forms a *linear vector space*. In addition, products of random variables are again random variables, so that the family of random variables forms an *Abelian group under multiplication*. The family of random variables is also closed under majorization, so that $Z:S \to \mathbb{R}$ defined by $Z(s) = \max(X(s),Y(s))$ for random variables X and Y is again a random variable. Then, the family of random variables forms a *lattice* with respect to the partial order $X \le Y$ (i.e., $X(s) \le Y(s)$ almost surely).

Most econometric applications deal with random variables which can be assumed to have finite variances. The space of these random variables is $\mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$, the space of random variables X for which $\mathbf{E}\, X^2 = \int_S X(x)^2 P(ds) < +\infty$. The space $\mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$ is also termed the space of *square-integrable functions*. The norm in this space is root-mean-square, $\|X\|_2 = [\int_S X(s)^2 \, P(ds)]^{½}$. Implications of $X \in \mathbf{L}_2(\mathbf{S},\boldsymbol{F},P)$ are $\mathbf{E}\,|X| \le \int_S \max(X(s),1)P(ds) \le \int_S (X(s)^2+1)P(ds) = \|X\|_2^2 + 1 < +\infty$ and $\mathbf{E}\,(X - \mathbf{E}X)^2 = \|X\|_2^2 - (\mathbf{E}\,|X|)^2 \le \|X\|_2^2 < +\infty$, so that X has a well-defined, finite mean and variance.

If T random variables are formed into a vector, $X(\cdot) = (X(\cdot,1),...,X(\cdot,T))$, the result is termed a *random vector*. For each $s \in \mathbf{S}$, the realization of the random vector is a point $(X(s,1),...,X(s,T))$ in $\mathbb{R}^T$, and the random vector has an induced probability on $\mathbb{R}^T$ which is characterized by its multivariate CDF, $F_X(x_1,...,x_T) = P(\{s\in\mathbf{S}|X(s,1)\le x_1,...,X(s,T)\le x_T\})$. Note that all the components of a random vector are functions of the *same* state of Nature s, and the random vector can be written as a function

$X:S \rightarrow \mathbb{R}^T$. The measurability of X requires $X^{-1}(\mathbf{C}) \in S$ for each open rectangle $\mathbf{C}$ in $\mathbb{R}^T$. The independence or dependence of the components of X is determined by the fine structure of P on $\mathbf{S}$.

Another way to write a random vector X is to define an index set $\mathbf{T} = \{1,...,T\}$, and then define X as a real-valued function on $\mathbf{S}$ and $\mathbf{T}$, $X:\mathbf{S}\times\mathbf{T}\rightarrow\mathbb{R}$. Then, $X(\cdot,t)$ is a simple random variable for each $t \in \mathbf{T}$, and $X(s,\cdot)$ is a real vector that is a realization of X for each $s \in \mathbf{S}$. The measurability requirement on X is the same as before, but can be written in a different form as requiring that the inverse image of each open interval in $\mathbb{R}$ be contained in $\boldsymbol{F}\otimes\boldsymbol{T}$, where $\boldsymbol{T}$ is a σ-field of subsets of $\mathbf{T}$ that can be taken to be the family of all subsets of $\mathbf{T}$ and "⊗" denotes the operation that forms the smallest σ-field containing all sets $\mathbf{A}\times\mathbf{B}$ with $\mathbf{A} \in \boldsymbol{F}$ and $\mathbf{B} \in \boldsymbol{T}$. There is then a complete duality between random vectors in a T-dimensional linear space and random functions on a T-dimensional index set. This duality between vectors and functions will generalize and provide useful insights into statistical applications in which $\mathbf{T}$ is a more general set indexing time.

2. Stochastic Processes

Consider a measurable space $(\mathbf{T},\boldsymbol{T})$ consisting of an ordered set $\mathbf{T}$ indexing time, and a σ-field $\boldsymbol{T}$ of subsets of $\mathbf{T}$. The set $\mathbf{T}$ is assumed to be a complete separable metric space. The most important cases are $\mathbf{T}$ discrete, either $\mathbf{T} = \{1,2,3,...\}$ with initial time t = 1, or $\mathbf{T} = \{...,-2,-1,0,1,2,...\}$ with an indefinite past, and $\mathbf{T}$ continuous, either $\mathbf{T} = [0,+\infty)$ with an initial time t = 0, or $\mathbf{T} = (-\infty,+\infty)$ with an indefinite past. In the discrete case, $\boldsymbol{T}$ is usually assumed to be the smallest σ-field containing all sets of the form $\{...,t-2,t-1,t\}$ for each integer t; this also guarantees that $\boldsymbol{T}$ contains all finite subsets of $\mathbf{T}$. In the continuous case, $\boldsymbol{T}$ is usually assumed to be the Borel σ-field relative to $\mathbf{T}$, so that it contains all sets of the form $(-\infty,t)$, $(-\infty,t]$, $[0,t)$, and $[0,t]$.

A real-valued function $X:\mathbf{S}\times\mathbf{T} \rightarrow \mathbb{R}$ is called a *stochastic process*. When $\mathbf{T}$ is discrete, X is called a discrete-time stochastic process, and when $\mathbf{T}$ is continuous, X is called a continuous-time stochastic process. For each $s \in S$, $X(s,\cdot)$ is a real-valued function on $\mathbf{T}$. and is termed a *realization*, *trajectory*, or *sample path* of the stochastic process, and for each $t \in \mathbf{T}$, $X(\cdot,t)$ is a random variable. Alternately one can think of a stochastic process as a function $X:\mathbf{S} \rightarrow \mathbb{R}^{\mathbf{T}}$ that maps states of Nature into random vectors in the *product space* $\mathbb{R}^{\mathbf{T}}$ whose coordinates are indexed by $\mathbf{T}$, and think of a trajectory as a vector $X(s)$ in this linear space with components $X(s,t)$. There are substantial technical differences between finite and infinite dimensional vector spaces, but it will nevertheless be the case that much of the geometric intuition for manipulation of vectors in finite-dimensional space will carry over to the manipulation of stochastic processes, even in the case of continuous time.

For a finite subset $\mathbf{T}_1 = \{t_1,....,t_n\}$ of $\mathbf{T}$, the random vector $(X(\cdot,t_1),...,X(\cdot,t_n))$ is termed the *restriction* of X to $\mathbf{T}_1$. Conversely, X is termed an *extension* of this finite-dimensional random vector. A *rectangle* in $\mathbb{R}^{\mathbf{T}}$ is a set of the form $\prod_{t\in T}\mathbf{A}_t$, where $\mathbf{A}_t$ is a set in the Borel σ-field of subsets of $\mathbb{R}$ for each $t \in \mathbf{T}$, and $\mathbf{A}_t = \mathbb{R}$ except for t in a finite set $\mathbf{T}_1 \subseteq \mathbf{T}$. The σ-field of subsets of $\mathbb{R}^{\mathbf{T}}$ generated by the rectangles, called the product σ-field, is denoted $\bigotimes_{t\in T}\boldsymbol{B}$; this is the smallest σ-field with the property that if X is measurable with respect to $\bigotimes_{t\in T}\boldsymbol{B}$ (i.e., for each restriction Y of X to a finite subset $\mathbf{T}_1 = \{t_1,....,t_n\}$ of $\mathbf{T}$ and open rectangle $\mathbf{C} \subseteq \mathbb{R}^n$, the set $\{s\in\mathbf{S}|Y(s) \in \mathbf{C}\}$ is contained in $\boldsymbol{F}$), then all the restrictions of X to finite-dimensional coordinate subspaces are measurable. A family of random vectors defined on the finite subsets of $\mathbf{T}$ are *compatible* if for each pair of finite subsets $\mathbf{T}_1$ and $\mathbf{T}_2$ of $\mathbf{T}$ satisfying $\mathbf{T}_1 \subseteq \mathbf{T}_2$, the restriction to $\mathbf{T}_1$ of the random vector assigned to $\mathbf{T}_2$

coincides almost surely with the random vector assigned to $\mathbf{T}_1$. An important result called the *Kolmogorov existence theorem* states that the finite-dimensional restrictions of a stochastic process on **T** are a compatible family of random vectors, and conversely any compatible family of random vectors defined on the finite subsets of a complete separable metric space **T** have an almost surely unique extension to $(\mathbf{T},\bigotimes_{t\in T}\boldsymbol{B})$. These definitions were encountered previously in the discussion of repeated trials and independence in Chapter 3. There **T** was countable, but the existence theorem holds for uncountable **T** as well. One implication of this result is that the probability distribution of a stochastic process on **T** is characterized by the finite-dimensional distributions associated with its restrictions to finite-dimensional subsets of **T**. Another implication of this result is that if one starts with the family of random vectors $(X(\cdot,t_1),...,X(\cdot,t_n))$ on finite subsets $\mathbf{T}_1 = \{t_1,....,t_n\}$ of $[0,+\infty)$ that are multivariate normal with mean zero and covariances $E\,(X(\cdot,t_i)X(\cdot,t_j) = \min(t_i,t_j)$, then these random vectors are compatible, and hence are restrictions of a stochastic process X on $[0,+\infty)$ that has $X(\cdot,t)$ normal with mean zero and variance t for each t, has $E\,X(\cdot,t)[X(\cdot,t') - X(\cdot,t)] = 0$ for $t' > t$, and has trajectories that are almost surely continuous. The process X is called the *Weiner process*; it plays a fundamental role in the statistical theory of continuous time stochastic processes.

For analysis, one usually needs to restrict further the family of stochastic processes on the index space $(\mathbf{T},\boldsymbol{T})$. One way to do this is to specify a general space of stochastic processes satisfying some conditions, such as continuity and integrability conditions, and then argue that the time series encountered in an econometric application are sample paths from stochastic processes meeting these definitions. A second way is to generate specific families of stochastic processes, say with a particular dynamic structure such as finite-order autoregression, and study the properties of the spaces of stochastic processes generated by these specific families. We will begin with the first approach, which is often quite useful for establishing statistical limit theory for stochastic processes. The second is more commonly used in time series econometrics, which for many purposes does not have to use the linear space characterizations of the stochastic processes it treats.

For most econometric time series and problems, it is natural to treat observations as coming from discrete time stochastic processes. The discrete time analysis is also technically simpler. Consequently, it is the best starting point for a course in econometric time series analysis. However, there are some econometric and finance applications where continuous time processes are required, so that it is useful to identify a few important cases and connections to the discrete time treatment of time series that is the main subject of this course. A stochastic process on continuous **T** is said to be *separabl*e if **T** contains a countable dense subset $\mathbf{T}_0$, and almost surely the graph of the process on **T** is contained in the closure of the graph of the process on $\mathbf{T}_0$; i.e., for s in a subset of **S** that occurs with probability one, the set $\{(t,x)|x=X(s,t)$ for $t\in\mathbf{T}\}$ is contained in the closure of $\{(t,x)|x=X(s,t)$ for $t\in\mathbf{T}_0\}$. The concept of separability of a stochastic process is different than the concept of a separable metric space, although the two are closely related. A separable stochastic process has the property that its behavior is effectively determined on a countable set of times, similarly to a discrete time stochastic process. This permits many of the properties of a separable continuous time stochastic process to be deduced by limiting arguments from the properties of discrete time stochastic processes.

One leading vector space for stochastic processes in continuous time is $C(\mathbf{T})$, the space of continuous real-valued functions $X(s,\cdot)$ on **T**. Then, a continuous time stochastic process with a.s.

continuous trajectories will be a measurable function from (**S**,***F***,*P*) into C(**T**). A second common vector space is D(**T**), the space of real-valued functions X(s,·) on **T** with the property that the sample paths are almost surely right-continuous and have left-hand limits. When **T** is the real line with the Borel σ-field, both C(**T**) and D(**T**) are spaces of separable stochastic processes. An important limit theorem that is the basis for other limit results for continuous time processes, due to Donsker, states that if Y_i are i.i.d. random variables with mean zero and variance one, T = [0,1], and one constructs the sequence of stochastic process in C[T] that satisfy

$$X_n(s,t) = n^{-1/2}\{Y_1(s) + Y_2(s) + ...+ Y_{[nt]}(s) + (nt - [nt])Y_{[nt]+1}(s)\},$$

where [nt] is the largest integer that does not exceed nt, then X_n converges in distribution to the Weiner process (i.e., for every finite-dimensional subset of **T**, the restriction of X_n converges in distribution to the corresponding restriction of the Weiner process.)

Thinking of stochastic processes as random vectors and their trajectories as points in a linear vector space provides some insights that will be useful later. First, linear transformations in finite dimensional vector spaces are a familiar tool in econometrics, and linear operators in these spaces are identified with matrices. We give examples of linear operators that are important in time series analysis, defined here in a finite-dimensional setting in $\mathbb{R}^3$. They are the shift (or lag) operator **L**,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{L} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix};$$

the smoothing operator

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = (\tfrac{1}{2}\mathbf{I} + \tfrac{1}{2}\mathbf{L}) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix};$$

the differencing operator

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = (\mathbf{I} - \mathbf{L}) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix};$$

and projections onto linear subspaces. Operations like smoothing and differencing are sometimes called *filters*. These concepts carry over to infinite dimensions, where they are defined as linear operators in linear spaces of stochastic processes. When observed time series are finite dimensional, operations on these series will coincide with finite-dimensional restrictions of linear operators on stochastic processes. Specifically, lag, filter, and projection operations on stochastic processes have finite-dimensional matrix operator analogs that are used to manipulate data, and can be helpful in understanding what these operators are doing. This duality holds even in the case of continuous time stochastic processes, where linear operators take the form of stochastic integrals and integral equations. Despite the different look, these are again just infinite-dimensional extensions of finite-dimensional matrix transformations.

Another useful insight comes from considering different representations of vectors in finite-dimensional spaces, and extending these ideas to infinite-dimensional situations. To be specific, consider $\mathbb{R}^2$. When we express a function X on T = {1,2} as a point (X(1),X(2)) in this space, what we are really doing is defining two functions $Z_1 = (1,0)$ and $Z_2 = (0,1)$ with the property that Z_1 and Z_2 span the space, and then writing X as the linear combination $X = X(1)\cdot Z_1 + X(2)\cdot Z_2$. The pair of functions (points) Z_1 and Z_2 is called a *Hamel basis* for $\mathbb{R}^2$, and every point in the space has a unique representation in terms of this basis. However, there may be many different Hamel bases. For example, the unit function (1,1) and the function cos(πt) or (-1,1) also form a Hamel basis, and in terms of this basis X has the representation $X = \frac{1}{2}(X(1)+X(2))\cdot(1,1) + \frac{1}{2}(X(2)-X(1))\cdot(-1,1)$.

For countably infinite T = {1,2,...}, the coordinate functions $Z_i = (0,...,0,1,0,...)$ with a one in the i-th place form one Hamel basis. Another Hamel basis is the list of trigonometric functions $Z_{2i} = \cos(\pi t/2i)$ and $Z_{2i+1} = \sin(\pi t/2i)$ for i = 1,2,... . The representation of X in terms of this basis is called a *Fourier representation*. The reason to consider different bases is that they may reveal a simple or insightful characterization of a particular family of stochastic processes. For example, the coordinate basis representation above yields what is called the *time domain* representation of the process which is particularly helpful when the process has a simple dynamic structure. The Fourier representation yields what is called the *frequency domain* representation of the process which is particularly helpful when the process contains important cyclic components in different frequency ranges. Geometrically, these are just different representations of the same point in linear space, and one can switch from one to the other to gain insight into the structure of a particular economic time series.

3. Limit Theorems for Discrete-Time Stochastic Processes

The fundamental mechanisms that make statistics work are the data transformations that enhance "signal" relative to "noise" by filtering out noise components. For example, sample means can provide good estimates of location when the averaging process operates to dampen the influence of noise. Statistical limit theorems such as Laws of Large Numbers and Central Limit Theorems provide the analytic backbone for understanding how these filtering operations work, and also provide the basis for large-sample approximations that in many applications are an adequate foundation for statistical inference. In cross-section statistics, you get the statistical limit theorems to work through "ensemble" averages of independent observations obtained by random sampling. Then, asymptotic theory as random sample sizes go to infinity will usually be an adequate approximation for random samples of the sizes encountered in cross-section econometrics.

It is possible in principle to develop a statistical limit theory for time series by taking the ensemble approach. To make this work, one has to treat each observed time series as one observation, and consider the limit as the number of independent repeated draws of observed trajectories goes to infinity. This approach would be essentially the same as the usual cross-section analysis, and the relatively simple and tight asymptotic theory for cross-sections would carry over more or less directly. The problem is that most commonly in time series analysis, only one realization is observed, and the ensemble asymptotics is not relevant. It is not that one can not in principle envision "parallel universes" over which ensemble limits could be compiled, but rather that our inability to draw observations from parallel universes implies that the ensemble asymptotics are inadequate approximations to the actual finite sample observations. A partial exception to this dismissal of ensemble asymptotics is the analysis of panel data where the number of panel members is moderately large. For example, in study of time series for a panel of countries, counties, or firms, a limit theory based on the number of panel members may be adequate.

More commonly, the possibility of dampening noise and detecting signal in time series has to come through filtering of a single realization of the series, and the relevant limit theory has to come through letting the number of time series observations increase. This will lead us to consider time averages $T^{-1}\sum_{t=1}^{T} X_t$ of a stochastic process X (which itself may be the result of filtering). The stochastic process is said to be *ergodic* if the time average converges in probability to a limit. Then, for statistical analysis we will be interested in conditions under which time averages are ergodic, and under which suitably normalized time averages satisfy central limit theorems. Even before turning to statistical considerations, there are some common sense qualifications on what one can hope to determine from the available time-series data. First, a phenomenon that expresses itself globally, such as the location of a stochastic process that contains no trends, can be investigated through time averages. For a moderately broad class of stochastic processes, such time averages will have reasonable large sample properties that permit one to make useful statistical statements about the global phenomenon. By contrast, there are phenomenon that are intrinsically local, such as the exact timing of a regime shift, or the location of a stochastic process between regime shifts. The available data that can shed light on these phenomena are intrinsically finite, with information on timing limited by the periodicity of observations and information on location limited by the duration of the intra-regime shift interval. Statistical investigation of these phenomena must be limited to the finite samples that are relevant, and asymptotics are not helpful. In further discussion where we concentrate on the limiting properties of time averages, one should keep in mind that there may be time series phenomena of interest for which the behavior of such time averages is irrelevant. Then, to a greater degree than in cross-section econometrics, finite-sample statistis done analytically where possible, and with bootstrap techniques otherwise, will be a major analytic tool, and large sample approximations will be correspondingly less important.

A critical factor in the behavior of time averages is the relationship between a current observation and the history of the process. If history has too much influence, then there may be insufficient opportunity for recent noise to average against and diminish the impact of noise from the past. The analysis of the impact of history requires that we formalize the concept of historical information, and the conditional distribution of a stochastic process given historical information. We do this by identifying information with σ-fields of events.

Consider a stochastic process Y in discrete time, and let $Y_t = Y(\cdot,t)$ denote its component in period t. The dependence of Y_t on the state of Nature s will be suppressed. We will take T to be the infinite sequence $\mathbf{T} = \{1,2,...\}$, or as a doubly infinite sequence, extending back in time as well as forward, $\mathbf{T} = \{...,-2,-1,0,1,2,...\}$. The trajectories of Y are points in the product space $\mathbf{W} = \prod_{t\in\mathbf{T}} \mathbb{R}$ or $\mathbf{W} = \mathbb{R}^{\mathbf{T}}$, where $\mathbb{R}$ is the real line. The "complete information" σ-field of subsets of **W** is defined as $\boldsymbol{F}_{\mathbf{T}} = \otimes_{t\in\mathbf{T}} \boldsymbol{B}$, where $\boldsymbol{B}$ is the Borel σ-field of subsets of the real line. (The same apparatus, with **T** equal to the real line, can be used to consider continuous time.) Accumulation of information is described by a nondecreasing sequence of σ-fields $... \subseteq \boldsymbol{G}_{-1} \subseteq \boldsymbol{G}_0 \subseteq \boldsymbol{G}_1 \subseteq \boldsymbol{G}_2 \subseteq ...$ with $\boldsymbol{G}_t$ interpreted as the historical information available to an observer at time t and defined as $\boldsymbol{G}_t = (\otimes_{i\le t} \boldsymbol{B})\otimes(\otimes_{i>t}\{\varphi,\mathbf{S}\})$, capturing the idea that at time t the future is unknown. The monotone sequence of σ-fields $\boldsymbol{G}_t$, i = ...,-1,0,1,2,... is called a *filtration*. The stochastic process is *adapted* to the filtration if Y_t is measurable with respect to $\boldsymbol{G}_t$ for each t. Some authors use the notation $\sigma(...,Y_{t-2},Y_{t-1},Y_t)$ for $\boldsymbol{G}_t$ to emphasize that it is the σ-field generated by the information contained in Y_s for $s \le t$. Note that there may be more than one sequence of σ-fields in operation for a particular stochastic process. These might correspond, for example, to the information available to different economic agents. We will need in particular the sequence of σ-fields $\boldsymbol{H}_t = \sigma(Y_t,Y_{t+1},Y_{t+2},...)$ adapted to the process from time t forward; this is a nonincreasing sequence of σ-fields... $\supseteq \boldsymbol{H}_{t-1} \supseteq \boldsymbol{H}_t \supseteq \boldsymbol{H}_{t+1} \supseteq ...$. Sometimes $\boldsymbol{G}_t$ is termed the *natural upward filtration*, and $\boldsymbol{H}_t$ the *natural downward filtration*. If we start with a stochastic process Y that is measurable with respect to the product σ-field $\boldsymbol{F}_{\mathbf{T}} = \otimes_{t\in\mathbf{T}} \boldsymbol{B}$, then the natural upward and downward filtrations are subfields of $\boldsymbol{F}_{\mathbf{T}}$, and Y_t is adapted to the natural upward filtration. Conversely, if we construct a sequence of random variables Y_t that are adapted to the natural upward filtration for $t \in \mathbf{T}$, then they define a stochastic process on **T** that is measurable with respect to $\boldsymbol{F}_{\mathbf{T}}$.

Each subsequence $(Y_m,...,Y_{m+n})$ of the stochastic process has a multivariate CDF $F_{m,..,m+n}(y_m,...,y_{m+n})$. The process is said to be *strictly stationary*, or simply *stationary*, if for each n, this CDF is the same for every m. The process is said to be *covariance stationary* or *weakly stationary* if its means, variances, and covariances between random variables a fixed number of time periods apart are the same for all times m. A stationary process for which second moments exist is obviously always covariance stationary. A stochastic process composed of a sequence of i.i.d. random variables is always stationary. The concept of stationarity plays an important role in time series econometrics because it is considered a plausible property for many economic time series and because it excludes trends that complicate the definition and interpretation of time averages. Stationarity alone is neither necessary and sufficient for ergodicity. It is not necessary because time averages may converge in the presence of seasonality and other kinds of limited heterogeneity that are inconsistent with stationarity. Conversely, if statistical dependence across time is too strong, time averages may fail to converge even if the process is stationary.

Example 1: The stochastic process $Y_t = \varepsilon_t + \alpha\cdot\cos(\pi t/6)$, where ε_t is an i.i.d. standard normal disturbance and α is a non-zero constant, is non-stationary, having a seasonal cycle of length 12. The

time average $T^{-1}\sum_{t=1}^{T} Y_t$ is normally distributed with a mean that is bounded in magnitude by 2.37/T and a variance 1/T. Then, the process is ergodic even though it is not stationary.

Example 2: The stochastic process $Y_t = \varepsilon_1$, where ε_1 is a single standard normal draw, is stationary, but $T^{-1}\sum_{t=1}^{T} Y_t$ is standard normal for all T, so the process is not ergodic.

Example 3: The stochastic process $Y_t = \alpha t + \varepsilon_t$, where ε_t is an i.i.d. standard normal disturbance and α is a non-zero constant, is nonstationary due to the deterministic trend in its location. The time average $T^{-1}\sum_{t=1}^{T} Y_t$ is normal with mean (T+1)/2 and variance 1/T, so that the process is not ergodic. Thus, trends may cause ergodicity to fail. This is not true for all trends, however. If the previous process is modified to $Y_t = \alpha t/(1+t) + \varepsilon_t$, so that it has a bounded nonlinear trend, then it is ergodic.

Example 4: The non-stationary stochastic process $Y_t = \varepsilon_t + \alpha\cdot\cos(\pi t/6)$, where ε_t is a i.i.d. standard normal disturbance and α is a standard normal draw, has a stochastic seasonal cycle of length 12. The time average $T^{-1}\sum_{t=1}^{T} Y_t$ is normally distributed with mean zero and a variance 1/T, so that the process is ergodic. However, $\mathrm{cov}(Y_t, Y_{t+12k}) = \cos(\pi t/6)^2$ for all k, so that covariances do not all dampen out as the time interval between observations grows. This pattern of covariances is inconsistent with the usual conditions for weak stochastic dependence over time, and shows that these conditions are not necessary for ergodicity.

Example 5: The stochastic process $Y_t = \varepsilon_t$, where the ε_t are i.i.d. with a Cauchy distribution, is stationary, but not covariance stationary, and non-ergodic. Thus, stationarity and a strong form of weak dependence, namely independence, are not sufficient for ergodicity if the tails of the random variables in the stochastic process are too thick.

Keeping these examples in mind, we will consider several forms of weak dependence that in combination with other assumptions will lead to ergodicity. Specifically, we consider conditional moment restrictions, given history, that lead to what is called martingale limit theory, and restrictions on the influence of historical information from the remote past, that lead to what are called mixing conditions. These cases will cover most of the econometric time series models that generate stationary stochastic processes.

Martingale Limit Theory

One circumstance that arises in some economic time series is that while the successive random variables are not independent, they have the property that their expectation, given history, is zero. Changes in stock market prices, for example, will have this property if the market is efficient, with arbitragers finding and bidding away any component of change that is predictable from history. A sequence of random variables X_t adapted to $\boldsymbol{G}_t$ is a *martingale* if almost surely $\mathbf{E}\{X_t|\boldsymbol{G}_{t-1}) = X_{t-1}$. If X_t is a martingale, then $Y_t = X_t - X_{t-1}$ satisfies $\mathbf{E}\{Y_t|\boldsymbol{G}_{t-1}) = 0$, and is called a *martingale difference* (m.d.) *sequence*. Thus, stock price changes in an efficient market form a m.d. sequence. It is also useful to define a *supermartingale* (resp., *submartingale*) if almost surely $\mathbf{E}\{X_t|\boldsymbol{G}_{t-1}) \le X_{t-1}$ (resp,, $\mathbf{E}\{X_t|\boldsymbol{G}_{t-1}) \ge X_{t-1}$). The following result, called the *Kolmogorov maximal inequality*, is a useful property of martingale difference sequences. A proof can be found in Chapter 4.

Theorem 1. If random variables Y_k are have the property that $E(Y_k|Y_1.,,,.Y_{k-1}) = 0$, or more technically the property that Y_k adapted to $\sigma(...,Y_{k-1},Y_k)$ is a martingale difference sequence, and if $\mathbf{E}Y_k^2 = \sigma_k^2$, then $P(\max_{1\le k\le n} | \sum_{i=1}^{k} Y_i| > \varepsilon) \le \sum_{i=1}^{n} \sigma_i^2/\varepsilon^2$.

Mixing

Many economic time series exhibit correlation between different time periods, but these correlations dampen away as time differences increase. Bounds on correlations by themselves are typically not enough to give a satisfactory theory of stochastic limits, but a related idea is to postulate that the degree of statistical dependence between random variables approaches negligibility as the variables get further apart in time, because the influence of ancient history is buried in an avalanche of new information (*shocks*). To formalize this, we introduce the concept of *stochastic mixing*. For a stochastic process Y, consider events $\mathbf{A} \in \boldsymbol{G}_t$ and $\mathbf{B} \in \boldsymbol{H}_{t+s}$, where $\boldsymbol{G}_t$ is the natural upward filtration and $\boldsymbol{H}_{t+s}$ is the natural downward filtration. Then **A** draws only on information available up through period t and **B** draws only on information available from period t+s on. The idea is that when s is large, the information in **A** is too "stale" to be of much use in determining the probability of **B**, so that these events are nearly independent. Three definitions of mixing are given in the table below; they differ only in the manner in which they are normalized, but this changes their strength in terms of how broadly they hold and what their implications are. When the process is stationary, mixing depends only on time differences, not on time location.

Form of Mixing	**Coefficient**	**Definition** (for all $\mathbf{A} \in \boldsymbol{G}_t$ and $\mathbf{B} \in \boldsymbol{H}_{t+s}$, and all t)
Strong	$\alpha(s) \to 0$	$\lvert P(\mathbf{A}\cap\mathbf{B}) - P(\mathbf{A})\cdot P(\mathbf{B})\rvert \le \alpha(s)$
Uniform	$\varphi(s) \to 0$	$\lvert P(\mathbf{A}\cap\mathbf{B}) - P(\mathbf{A})\cdot P(\mathbf{B})\rvert \le \varphi(s)P(\mathbf{A})$
Strict	$\psi(s) \to 0$	$\lvert P(\mathbf{A}\cap\mathbf{B}) - P(\mathbf{A})\cdot P(\mathbf{B})\rvert \le \psi(s)P(\mathbf{A})\cdot P(\mathbf{B})$

There are links between the mixing conditions and bounds on correlations between events that are remote in time:

(1) Strict mixing $\Longrightarrow$ Uniform mixing $\Longrightarrow$ Strong mixing.
(2) (Serfling) If the Y_i are uniform mixing with $\mathbf{E}Y_i = 0$ and $\mathbf{E}Y_t^2 = \sigma_t^2 < +\infty$, then $|\mathbf{E}Y_tY_{t+s}| \le 2\varphi(s)^{1/2}\sigma_t\sigma_{t+s}$.
(3) (Ibragimov) If the Y_i are strong mixing with $\mathbf{E}Y_t = 0$ and $\mathbf{E}|Y_t|^d < +\infty$ for some $d > 2$, then $|\mathbf{E}Y_tY_{t+s}| \le 8\alpha(s)^{1-2/d}\sigma_t\sigma_{t+s}$.
(4) If there exists a sequence ρ_t with $\lim_{t\to\infty}\rho_t = 0$ such that $|\mathbf{E}(U-\mathbf{E}U)(W-\mathbf{E}W)| \le \rho_t[(\mathbf{E}(U-\mathbf{E}U)^2)(\mathbf{E}(W-\mathbf{E}W)^2)]^{½}$ for <u>all</u> bounded continuous functions $U = g(Y_1,...,Y_t)$ and $W = h(Y_{t+n},...,Y_{t+n+m})$ and <u>all</u> t, n, m, then the Y_t are strict mixing.

An example gives an indication of the restrictions on a dependent stochastic process that produce strong mixing at a specified rate. First, suppose a stationary stochastic process Y_t satisfies $Y_t = \rho Y_{t-1}$

+ Z_t, with the Z_t independent standard normal. Then, $\text{var}(Y_t) = 1/(1-\rho^2)$ and $\text{cov}(Y_{t+s},Y_t) = \rho^s/(1-\rho^2)$, and one can show with a little analysis that $|P(Y_{t+s} \le a, Y_t \le b) - P(Y_{t+s} \le a)\cdot P(Y_t \le b)| \le \rho^s/\pi(1-\rho^{2s})^{1/2}$. Hence, this process is strong mixing with a mixing coefficient that declines at a geometric rate. This is true more generally of processes that are formed by taking stationary linear transformations of independent processes.

Consider a stochastic process Y and the time averages $X_T = T^{-1}\sum_{i=1}^{T} Y_i$ for T = 1,2,.... *Laws of large numbers* give conditions under which the time averages X_T converge to a constant, either in probability (weak laws, or WLLN) or almost surely (strong laws, or SLLN). We give one WLLN and two SLLN that can be applied to time averages of discrete-time stochastic processes.

Theorem 2. (WLLN) If a stochastic process Y has $\lim_{T\to\infty} T^{-1}\sum_{t}^{T} \mathbf{E}\, Y_t = \mu$, $\mathbf{E}(Y_t-\mu)^2 \equiv \sigma_t^2$, and $|\text{cov}(Y_t,Y_s)| \le \rho_{ts}\sigma_t\sigma_s$ with $\sum_{t=}^{\infty} \sigma_t^2/t^{3/2} < +\infty$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{m=1}^{n} \rho_{km} < +\infty$, then the process is ergodic; i.e., $X_T \to_p \mu$.

Proof: Using Chebyshev's inequality, it is sufficient to show that $\mathbf{E}(X_n-\mathbf{E}X_n)^2$ converges to zero. The Cauchy-Schwartz inequality (see Chap 2.1.11) is applied first to establish

$$\left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m\rho_{km}\right)^2 \le \left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m^2\right)\left(\frac{1}{n}\sum_{m=1}^{n} \rho_{km}^2\right)$$

and then to establish that

$$\mathbf{E}(X_n-\mathbf{E}X_n)^2 = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{m=1}^{n} \sigma_k\sigma_m\rho_{km} = \frac{1}{n}\sum_{k=1}^{n} \sigma_k\left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m\rho_{km}\right)$$

$$\le \left(\frac{1}{n}\sum_{k=1}^{n} \sigma_k^2\right)^{1/2}\left[\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m\rho_{km}\right)^2\right]^{1/2} \le \left(\frac{1}{n}\sum_{k=1}^{n} \sigma_k^2\right)^{1/2}\left[\left(\frac{1}{n}\sum_{m=1}^{n} \sigma_m^2\right)\left(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{m=1}^{n} \rho_{km}^2\right)\right]^{1/2}$$

$$= \left(\frac{1}{n}\sum_{k=1}^{n} \sigma_k^2\right)\left(\frac{1}{n^2}\sum_{k=1}^{n}\sum_{m=1}^{n} \rho_{km}^2\right)^{1/2} = \left(\frac{1}{n^{3/2}}\sum_{k=1}^{n} \sigma_k^2\right)\left(\frac{1}{n}\sum_{k=1}^{n}\sum_{m=1}^{n} \rho_{km}^2\right)^{1/2}.$$

The last form and Kronecker's lemma (Chapter 2.1.11) give the result. □

Corollary 2.1: If the Theorem 2 stochastic process Y is covariance stationary and $\sum_{k=}^{\infty} |\text{cov}(Y_t,Y_{t+k})| < +\infty$, then the process is ergodic.

Corollary 2.2: If the Theorem 2 stochastic process is covariance stationary and strong mixing with $E\,|Y|^d < +\infty$ for some $d > 2$ and mixing coefficients that satisfy $\sum_{k=}^{\infty} \alpha(k)^{1-2/d} < +\infty$, then the process is ergodic.

Theorem 3. (Martingale SLLN) If a stochastic process Y_k adapted to $\sigma(...,Y_{k-1},Y_k)$ is a martingale difference sequence, $\mathbf{E}Y_t^2 = \sigma_t^2$, and $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$, then $X_T \rightarrow_{as} 0$.

Proof: The theorem is stated and proved by J. Davidson (1994), p. 314. To give an idea why SLLN work, we give a simplified proof when the assumption $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$ is strengthened to $\sum_{k=1}^{\infty} \sigma_k^2/k^{3/2} < +\infty$. Either assumption handles the case of constant variances with room to spare. Kolmogorov's maximal inequality (Theorem 1) with $n = (m+1)^2$ and $\varepsilon = \delta m^2$ implies that

$$P(\max_{m^2\le k\le (m+1)^2} |X_k| > \delta) \le P(\max_{1\le k\le n} |\sum_{i=1}^{k} Y_i| > \delta m^2) \le \sum_{i=1}^{(m+1)^2} \sigma_i^2/\delta^2 m^4.$$

The sum over m of the right-hand-side of this inequality satisfies

$$\sum_{m=1}^{\infty} \sum_{i=1}^{(m+1)^2} \sigma_i^2/\delta^2 m^4 = \sum_{i=1}^{\infty} \sum_{m\ge i^{1/2}.}^{\infty} \sigma_i^2/\delta^2 m^4 \le 36 \sum_{i=1}^{\infty} \sigma_i^2/i^{3/2}\delta^2.$$

Then $\sum_{m=1}^{\infty} P(\sup_k |X_k| > \delta) \le 36 \sum_{i=1}^{\infty} \sigma_i^2/i^{3/2}\delta^2 < +\infty$. The Borel-Cantelli theorem, which states that if $\mathbf{A}_i$ is any sequence of events in a probability space $(\mathbf{S},\boldsymbol{F},P)$, then $\sum_{n=1}^{\infty} P(\mathbf{A}_i) < +\infty$ implies that almost surely only a finite number of the events $\mathbf{A}_i$ occur, gives the result. □

Corollary 3.1: If the Theorem 3 stochastic process is covariance stationary with $\mathbf{E}y_1^2 < +\infty$, then $X_T \rightarrow_{as} 0$.

Theorem 4. (Serfling SLLN) If a stochastic process Y has $\lim_{T\rightarrow\infty} T^{-1} \sum_t^T \mathbf{E}\, Y_t = \mu$, $\mathbf{E}(Y_k - \mathbf{E}(Y_k))^2 = \sigma_k^2$, and $|\mathbf{cov}(Y_k,Y_m)| \le \rho_{|k-m|}\sigma_k\sigma_m$, with the bounds $\sum_{k=1}^{\infty} (\log k)^2\sigma_k^2/k^2 < +\infty$ and $\sum_{k=}^{\infty} \rho_k < +\infty$, then $X_n \rightarrow_{as} \mu$.

Corollary 4.1: If the Theorem 4 stochastic process is covariance stationary and the covariances satisfy $\sum_{k=}^{\infty} |\text{cov}(Y_1,Y_k)| < +\infty$, then $X_n \rightarrow_{as} \mu$.

Corollary 4.2: If the Theorem 4 stochastic process is covariance stationary and strong mixing with $E\,|Y|^d < +\infty$ for some $d > 2$ and $\sum_{k=}^{\infty} \alpha(k)^{1-1/d} < +\infty$, then $X_n \rightarrow_{as} \mu$.

We next consider central limit theorems for time averages of stochastic processes. For a mean zero stochastic process Y, consider the scaled time averages $Z_T = T^{-1/2} \sum_{i=1}^{T} Y_t$. Central limit theorems (CLT) are concerned with conditions under which the Z_T, or variants with more generalized scaling, converge in distribution to a normal random variable Z_o. We give CLT for martingale difference processes and for stationary processes with strong mixing. The proofs of these theorems are quite technical, and are omitted.

When random variables are not identically distributed, a bound on the behavior of tails of their distributions, called the *Lindeberg condition*, is needed to get a CLT. This condition ensures that sources of relatively large deviations are spread fairly evenly through the series, and not concentrated in a limited number of observations. The Lindeberg condition can be difficult to interpret and check, but there are a number of sufficient conditions that are useful in applications. For example, a condition that random variables are uniformly bounded is always sufficient, and a condition that some moment higher than two is uniformly bounded is usually enough.

The next theorem establishes a CLT for martingale differences. The uniform boundedness assumption in this theorem is a strong restriction, but it can be relaxed to a Lindeberg condition or to a "uniform integratability" condition; see P. Billingsley (1984), p. 498-501, or J. Davidson (1994), p. 385.

Theorem 5. Suppose Y_k is a martingale difference sequence adapted to $\sigma(...,Y_{k-1},Y_k)$, and Y_k satisfies a uniform bound $|Y_k| < M$. Let $\mathbf{E}Y_k^2 = \sigma_k^2$, and assume that $n^{-1}\sum_{k=1}^{n} \sigma_k^2 \rightarrow \sigma_o^2 > 0$. Then $Z_n \rightarrow_d Z_o \sim N(0,\sigma_o^2)$.

Intuitively, the CLT results that hold for independent or martingale difference processes should continue to hold if the degree of dependence between variables is asymptotically negligible. The following theorem from I. Ibragimov and Y. Linnik, 1971, gives a CLT for stationary strong mixing processes. This result will cover a variety of economic applications, including stationary linear transformations of independent processes.

Theorem 6. (Ibragimov-Linnik) Suppose Y_k is stationary and strong mixing with mean zero, positive finite variance σ^2, and covariances $\mathbf{E}\, Y_{k+s}Y_k = \sigma^2\rho_s$. Suppose that for some r > 2, $\mathbf{E}|Y_n|^r < +\infty$ and $\sum_{k=1}^{\infty} \alpha(k)^{1-2/r} < +\infty$. Then, $\sum_{s=1}^{\infty} |\rho_s| < +\infty$ and $Z_n \rightarrow_d Z_o \sim N(0,\sigma^2(1+2\sum_{s=1}^{\infty} \rho_s))$.

In some applications, one encounters what are called "triangular arrays" of random variables, indexed both by time period and by sample size. This could arise, for example, if one applies a linear operator or filter to a stochastic process and the filter used depends on sample size. It is convenient to have a limit theory for such arrays. Let Y_{nt} with t = 1,2,...,n and n = 1,2,3,... denote a triangular array of random variables. (One additional level of generality could be introduced by letting t range from 1 up to a function of n that increases to infinity, but this is not needed for most applications.) This setup will include simple cases like $Y_{nt} = Z_t/n$ or $Y_{nt} = Z_t/n^{1/2}$, and more general weightings like $Y_{nt} = a_{nt}Z_t$ with an array of constants a_{nt}, but can also cover more complicated cases. Assume that for each n, the random variables Y_{nt} for t = 1,...,n form a martingale difference sequence; this is called a *martingale difference array*. Formally, consider random variables Y_{nt} for t = 1,...,n and n = 1,2,3,... that are adapted to σ-fields $\boldsymbol{G}_{nt}$ that are a filtration in t for each n, with the property that $E\{Y_{nt}|\boldsymbol{G}_{n,t-1}\} = 0$; A WLLN for this case is adapted from J. Davidson (1994), p. 299. Note that in this result, normalizing by sample size to form averages is subsumed in the definition of the rows of the array, so that one works with sums rather than averages.

Theorem 7: If Y_{nt} and $\boldsymbol{G}_{nt}$ for t = 1,...,n and n = 1,2,3,... is an adapted martingale difference array with $|Y_{nt}| \leq M$, $\mathbf{E}\, Y_{nt}^2 = \sigma_{nt}^2$, $\sum_{t=1}^{n} \sigma_{nt}$ uniformly bounded, and $\sum_{t=1}^{n} \sigma_{nt}^2 \rightarrow 0$, then the array is ergodic; i.e., $\sum_{t=1}^{n} Y_{nt} \rightarrow_p 0$.

It is also possible to establish a CLT for martingale difference arrays. The following result is taken from D. Pollard (1984), p. 170-174. The last condition in this theorem is a Lindeberg condition which will certainly be satisfied if the Y_{nt} are constructed by multiplying uniformly bounded random variables by scalars that uniformly tend to zero with n.

Theorem 8. Suppose Y_{nt} and $\boldsymbol{G}_{nt}$ for t = 1,...,n and n = 1,2,3,... is an adapted martingale difference array, $\lambda_{nt}^2 = E(Y_{nt}^2|\boldsymbol{G}_{n,t-1})$ is the conditional variance of Y_{nt}, $\sum_{t=1}^{n} \lambda_{nt}^2 \rightarrow_p \sigma^2 \in (0,+\infty)$. Suppose for each $\varepsilon > 0$, $\sum_{t=1}^{n} \mathbf{E}\, Y_{nt}^2 \cdot \mathbf{1}(|Y_{nt}| > \varepsilon) \rightarrow 0$. Then $X_n = \sum_{t=1}^{n} Y_{nt} \rightarrow_d X_o \sim N(0,\sigma^2)$.