Ch. 4 Asymptotic Theory

From the discussion of last Chapter it is obvious that determining the distribution of $h(X_1, X_2, ..., X_T)$ is by no means a trivial exercise.[1] It turns out that more often than not we cannot determine the distribution exactly. Because of the importance of the problem, however, we are forced to develop approximations; the subject of this Chapter.

This Chapter will cover the limit theorem. The terms 'limit theorems' refers to several theorems in probability theory under the generic names, 'laws of large numbers' (LLN) and 'central limit theorem' (CLT). These limit theorem constitute one of the most important and elegant chapters of probability theory and play a crucial role in statistical inferences.

1 Consistency

In this section we introduce the concepts needed to analyzed the behaviors of a random variable indexed by the size of a sample, say $\hat{\theta}_T$, as $T \to \infty$.

1.1 Limits (Non-Stochastic)

Definition:

Let $\{b_T\}_1^T$, or just $\{b_T\}$ be a sequence of real numbers. If there exists a real number b and if for every $\delta > 0$ there exist an integer $N(\delta)$ such that for all $T \geq N(\delta)$,[2] $|b_T - b| < \delta$, then b is the limit of the sequence $\{b_T\}$.

In this definition the constant δ can take on any real value, but it is the very small values of δ that provide the definition with its impact. By choosing a very small δ, we ensure that b_T gets arbitrarily close to its limit b for all T that are sufficiently large. When a limit exists, we say that the sequence $\{b_T\}$ converges to b

[1] Here, $h(X_1, X_2, ..., X_T)$ may be an estimator or a test statistic. For an estimator, we always are interested in his convergence in probability, while in the case of a test statistics, we are interested in his convergence in distribution.

[2] Here, N must have opposite direction with δ to capture the idea that as N getting larger (so does T), b_T and b get closer.

as T tends to infinity, written as $b_T \to b$ as $T \to \infty$. We also write $b = \lim_{T\to\infty} b_T$. When no ambiguity is possible, we simply write $b_T \to b$ or $b = \lim\ b_T$.

Example:
Let

$$a_T = \frac{2^T - (-1)^T}{2^T}.$$

Here $1 = \lim_{T\to\infty} a_T$, for

$$|a_T - 1| = \left| \frac{2^T - (-1)^T}{2^T} - 1 \right| = \frac{1}{2^T}.$$

Since by binomial theorem we have

$$2^T = (1+1)^T = 1 + T + \frac{T(T+1)}{2} \cdots + 1 > T.$$

Hence, if we choose $N = 1/\delta$ or large, we have, for $T > N$,

$$|a_T - 1| = \frac{1}{2^T} < \frac{1}{T} < \frac{1}{N} \leq \delta.$$

This complete the solution.

The concept of a limit extends directly to sequences of real vectors. Let $\mathbf{b}_T$ be a $k \times 1$ vector with real elements b_{Ti}, $i = 1, ..., k$. If $b_{Ti} \to b_i$, $\forall i = 1, ..., k$, then $\mathbf{b}_T \to \mathbf{b}$, where $\mathbf{b}$ has elements b_i, $i = 1, ..., k$. An analogous extensions applies to matrices.

Definition (Continuous):
Given $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^l$ $(k, l \in \mathbb{N})$ and $\mathbf{b} \in \mathbb{R}^{\mathbf{k}}$,
(a). the function $\mathbf{g}$ is continuous at $\mathbf{b}$ if for any sequence $\{\mathbf{b}_T\}$ such that $\mathbf{b}_T \to \mathbf{b}$, $\mathbf{g}(\mathbf{b}_T) \to \mathbf{g}(\mathbf{b})$; or equivalently
(b). the function $\mathbf{g}$ is continuous at $\mathbf{b}$ if for every $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that if $\mathbf{a} \in \mathbb{R}^k$ and $|a_i - b_i| < \delta(\varepsilon)$, $i = 1, ..., k$, then $|g_j(\mathbf{a}) - g_j(\mathbf{b})| < \varepsilon$, $j = 1, ..., l$.

Example:
From this it follows that if $\mathbf{a}_T \to \mathbf{a}$ and $\mathbf{b}_T \to \mathbf{b}$, then $\mathbf{a}_T + \mathbf{b}_T \to \mathbf{a} + \mathbf{b}$ and $\mathbf{a}_T \mathbf{b}'_T \to \mathbf{a}\mathbf{b}'$.

The following definition compares the behavior of a sequence $\{b_T\}$ with the behavior of a power of T, say T^λ, where λ is chosen so that $\{b_T\}$ and $\{T^\lambda\}$ behave similarly.

Definition:
(a). The sequence $\{b_T\}$ is at most of order T^λ, denoted $b_T = O(T^\lambda)$, if for some finite real number $\triangle > 0$, there exists a finite integer N such that for all $T \geq N$, $|T^{-\lambda} b_T| < \triangle$.
(b). The sequence $\{b_T\}$ is of order smaller than T^λ, denoted $b_T = o(T^\lambda)$, if for every real number $\delta > 0$, there exists a finite integer $N(\delta)$ such that for all $T \geq N(\delta)$, $|T^{-\lambda} b_T| < \delta$, *i.e.*, $T^{-\lambda} b_T \to 0$.

As we have defined these notations, $b_T = O(T^\lambda)$, if $\{T^{-\lambda} b_T\}$ is eventually bounded, whereas $b_T = o(T^\lambda)$ if $T^{-\lambda} b_T \to 0$. Obviously, if $b_T = o(T^\lambda)$, then $b_T = O(T^\lambda)$. Further, if $b_T = O(T^\lambda)$, then for every $\xi > 0$, $b_T = o(T^{\lambda+\xi})$. When $b_T = O(T^0)$, it is simply (eventually) bounded and may or may not have a limit. We often write $O(1)$ in place of $O(T^0)$. Similarly, $b_T = o(1)$ means $b_T \to 0$.

If each element of a vector or matrix is $O(T^\lambda)$ or $o(T^\lambda)$, then that vector or matrix is $O(T^\lambda)$ or $o(T^\lambda)$.

Proposition:
Let a_T and b_T be scalar.
(a). If $a_T = O(T^\lambda)$ and $b_T = O(T^\mu)$, then $a_T b_T = O(T^{\lambda+\mu})$ and $a_T + b_T = O(T^\kappa)$, where $\kappa = \max[\lambda, \mu]$.
(b). If $a_T = o(T^\lambda)$ and $b_T = o(T^\mu)$, then $a_T b_T = o(T^{\lambda+\mu})$ and $a_T + b_T = o(T^\kappa)$, where $\kappa = \max[\lambda, \mu]$.
(c). If $a_T = O(T^\lambda)$ and $b_T = o(T^\mu)$, then $a_T b_T = o(T^{\lambda+\mu})$ and $a_T + b_T = O(T^\kappa)$, where $\kappa = \max[\lambda, \mu]$.

1.2 Almost Sure Convergence

The stochastic convergence concept most closely related to the limit notations previously discussed is that of almost sure convergence. Recall our discussing a real-valued random variables b_T, we are in fact talking a mapping $b_T : S \to \mathbb{R}^1$.

we let s be a typical element of sample space S, and call the real number $b_T(s)$ a realization of the random variables.

Interest will often center on average such as

$$b_T(\cdot) = T^{-1}\sum_{t=1}^{T} Z_t(\cdot).$$

𝔇𝔢𝔣𝔦𝔫𝔦𝔱𝔦𝔬𝔫:

Let $\{b_T(\cdot)\}$ be a sequence of real-valued random variables. We say that $b_T(\cdot)$ converges almost surely to b, written $b_T(\cdot) \xrightarrow{a.s.} b$ if there exists a real number b such that $Pr\{s : b_T(s) \to b\} = 1$. When no ambiguity is possible, we may simply write $b_T \xrightarrow{a.s.} b$.

A sequence b_T converges almost surely if the probability of obtaining a realization of the sequence $\{Z_t\}$ for which convergence to b occurs is unity. Equivalently, the probability of observing a realization of $\{Z_t\}$ for which convergence to b does not occur is zero. Failure to converge is possible but will almost never happen under this definition.

As with nonstochastic limits, the almost sure convergence concept extends immediately to vectors and matrices of finite dimension. Almost sure convergence element by element suffices for almost sure convergence of vectors and matrices.

𝔓𝔯𝔬𝔭𝔬𝔰𝔦𝔱𝔦𝔬𝔫:

Given $\mathbf{g} : \mathbb{R}^{\mathbf{k}} \to \mathbb{R}^{\mathbf{l}}$ ($\mathbf{k}, \mathbf{l} \in \mathbb{N}$) and any sequence of random $k \times 1$ vector $\mathbf{b}_T$ such that $\mathbf{b}_T \xrightarrow{a.s.} \mathbf{b}$, where $\mathbf{b}$ is $k \times 1$, if $\mathbf{g}$ is continuous at $\mathbf{b}$, then $\mathbf{g}(\mathbf{b}_T) \xrightarrow{a.s.} \mathbf{g}(\mathbf{b})$.

This results is one of the most important in this Chapter, because consistency results for many of our estimators follows by simply applying this Proposition.

1.3 Convergence in Probability

A weaker stochastic convergence concept is that of convergence in probability.

𝔇𝔢𝔣𝔦𝔫𝔦𝔱𝔦𝔬𝔫:

Let $\{b_T\}$ be a sequence of real-valued random variables. If there exists a real number b such that for every $\delta > 0$, such that $Pr(s : |b_T(s) - b| < \delta) \to 1,\ as\ T \to \infty$, then b_T converge in probability to b, written as $b_T \xrightarrow{p} b$ or $plim\ b_T = b$.

Example:
Let $\bar{Z}_T \equiv T^{-1}\sum_{t=1}^{T} Z_t$, where $\{Z_t\}$ is a sequence of random variables such that $E(Z_t) = \mu$, $Var(Z_t) = \sigma^2 < \infty$ for all t and $Cov(Z_t, Z_\tau) = 0\ fort \neq \tau$. Then $\bar{Z}_T \xrightarrow{p} \mu$ by the Chebyshev weak law of large numbers. See the plot of Hamilton p.184.

When the *plim* of a sequence of estimator (such as $\{\bar{Z}_T\}_{T=1}^{\infty}$) is equal to the true population parameter (in this case, μ), the estimator is said to be *consistent*.

Convergence in probability is also referred as weak consistency, and since this has been the most familiar stochastic convergence concept in econometrics, the word "weak" is often simply dropped.

Theorem:
Let $\{b_T\}$ be a sequence of real-valued random variables. If $b_T \xrightarrow{a.s.} b$, then $b_T \xrightarrow{p} b$.

Vectors and matrices are said to converge in probability provided each element converges in probability.

Proposition:
Given $\mathbf{g} : \mathbb{R}^{\mathbf{k}} \to \mathbb{R}^{\mathbf{l}}$ $(\mathbf{k}, \mathbf{l} \in \mathbb{N})$ and any sequence of random $k \times 1$ vector $\mathbf{b}_T$ such that $\mathbf{b}_T \xrightarrow{p} \mathbf{b}$, where $\mathbf{b}$ is $k \times 1$, if g is continuous at $\mathbf{b}$, then $\mathbf{g}(\mathbf{b}_T) \xrightarrow{p} \mathbf{g}(\mathbf{b})$.

Example:
If $X_{1T} \xrightarrow{p} c_1$ and $X_{2T} \xrightarrow{p} c_2$, then $(X_{1T} + X_{2T}) \xrightarrow{p} (c_1 + c_2)$. This follows immediately, since $\mathbf{g}(X_{1T}, X_{2T}) \equiv (X_{1T} + X_{2T})$ is a continuous function of (X_{1T}, X_{2T}).

Example:
Consider an alternative estimator of the mean given by $\bar{Y}_T^* = [1/(T-1)]\sum_{t=1}^{T} Y_t$. This can be written as $c_{1T}\bar{Y}_T$, where $c_{1T} \equiv [T/(T-1)]$ and $\bar{Y}_T \equiv (1/T)\sum_{t=1}^{T} Y_t$. Under general condition, the sample mean is a consistent estimator of the population mean, implying that $\bar{Y}_T \xrightarrow{p} \mu$. It is also easy to verify that $c_{1T} \to 1$. Since

$c_{1T}\bar{Y}_T$ is a continuous function of c_{1T} and $\bar{Y}_T$, it follows that $c_{1T}\bar{Y}_T \xrightarrow{p} 1 \cdot \mu = \mu$. Thus $\bar{Y}_T^*$ is also a consistent estimator of μ.

Definition:

(a). The sequence $\{b_T\}$ is at most of order T^λ in probability, denoted $b_T = O_p(T^\lambda)$, if for every $\varepsilon > 0$ there exist a finite $\triangle_\varepsilon > 0$, and $N_\varepsilon \in \mathbb{N}$ such that for all $T \geq N_\varepsilon$, $Pr\{s : |T^{-\lambda} b_T(s)| > \triangle_\varepsilon\} < \varepsilon$.

(b). The sequence $\{b_T\}$ is of order smaller than T^λ in probability, denoted $b_T = o_p(T^\lambda)$, if $T^{-\lambda} b_T \xrightarrow{p} 0$.

Lemma (Product rule):

Let A_T be $l \times k$ and let b_T be $k \times 1$. If $A_T = o_p(1)$ and $b_T = O_p(1)$, then $A_T b_T = o_p(1)$.

Proof:

Each element of $A_T b_T$ is the sums of the product of $O_p(T^0) o_p(T^0) = o_p(T^{0+0}) = o_p(1)$ and therefore is $o_p(1)$.

1.4 Convergence in rth mean

The convergence notations of limits, almost sure limits, and probability limits are those most frequently encountered in econometrics, and most of the results in the literature are state in these terms. Another convergence concept often encountered in the context of time series data is that of convergence in the rth mean.

Definition:
Let $\{b_T\}$ be a sequence of real-valued random variables such that for some $r > 0$, $E|b_T|^r < \infty$. If there exists a real number b such that $E(|b_T - b|^r) \to 0$ as $T \to \infty$, then b_T converge in the rth mean to b, written as $b_T \xrightarrow{r.m.} b$.

The most commonly encountered situation is that of in which $r = 2$, in which case convergence is said to occur in *quadratic mean*, denoted $b_T \xrightarrow{q.m.} b$, or convergence in mean square, denoted $b_T \xrightarrow{m.s} b$.

Example:
Let $X_1, X_2, ..., X_T$ be *i.i.d.* random variables with mean μ and variance σ^2. Then $\bar{X}_T (= \frac{\sum_{t=1}^T X_t}{T}) \xrightarrow{m.s} \mu$ since $E|\bar{X}_T - \mu|^2 = E(\bar{X}_T - \mu)^2 = \sigma^2/T \to 0$.

Proposition (Generalized Chebyshev inequality):
Let Z be a random variable such that $E|Z|^r < \infty$, $r > 0$. Then for every $\varepsilon > 0$,

$$Pr(|Z| > \varepsilon) \leq \frac{E|Z|^r}{\varepsilon^r}.$$

When $r = 1$ we have Markov's inequality and when $r = 2$ we have the familiar Chebyshev inequality.

Theorem:
If $b_T \xrightarrow{r.m.} b$ for some $r > 0$, then $b_T \xrightarrow{p} b$.

Proof:
Since $E(|b_T - b|^r) \to 0$ as $T \to \infty$, $E(|b_T - b|^r) < \infty$ for all T sufficiently large. It follows from the Generalized Chebyshev inequality that, for every $\varepsilon > 0$,

$$Pr(s : |b_T(s) - b| > \varepsilon) \leq \frac{E|b_T - b|^r}{\varepsilon^r}.$$

Hence $Pr(s : |b_T(s) - b| < \varepsilon) \geq 1 - \frac{E|b_T - b|^r}{\varepsilon^r} \to 1$ as $T \to \infty$, since $b_T \xrightarrow{r.m.} b$. It follows that $b_T \xrightarrow{p} b$.

Without further conditions, no necessary relationship holds between convergence in the rth mean and almost sure convergence.

2 Convergence in Distribution

The most fundamental concept is that of convergence in distribution.

Definition:

Let $\{b_T\}$ be a sequence of scalar random variables with cumulative distribution function $\{F_T\}$. If $\lim_{T\to\infty} F_T(z) = F(z)$ for every continuity point z, where F is the (cumulative) distribution of a random variable Z, then b_T converge in distribution to the random variable Z, written as $b_T \xrightarrow{d} Z$.

When $b_T \xrightarrow{d} Z$, we also say that b_T converges in law to Z, written as $b_T \xrightarrow{L} Z$, or that b_T is asymptotically distributed as F, denoted as $b_T \overset{A}{\sim} F$. Then F is called the limiting distribution of b_T.

Example:

Let $\{Z_t\}$ be *i.i.d.* random variables with mean μ and finite variance $\sigma^2 > 0$. Define

$$b_T \equiv \frac{\bar{Z}_T - E(\bar{Z}_T)}{(Var(\bar{Z}_T))^{1/2}} = \frac{T^{-1/2}\sum_{t=1}^{T}(Z_t-\mu)}{\sigma} = \frac{\sqrt{T}(\bar{Z}_t-\mu)}{\sigma}.$$

Then by the Lindeberg-Levy central limit theorem, $b_T \overset{A}{\sim} N(0,1)$. See the plot of Hamilton p.185.

The above definition are unchanged if the scalar b_T is replaced with an $(k\times 1)$ vector $\mathbf{b}_T$. However it is noted that $\mathbf{b}_T \xrightarrow{L} \mathbf{z}$ implies $b_{iT} \xrightarrow{L} z_i, i = 1, 2, ..., k$, but the reverse is not true, that is pointwise convergence in distribution is **necessary** but not **sufficient** for joint convergence in distribution. One *can not* simply prove that $b_{iT} \xrightarrow{L} z_i$ for all $i = 1, 2, ..., k$ and say that $\mathbf{b}_T \xrightarrow{L} \mathbf{z}$. However, a simple way to verify convergence in distribution of a vector from random scalars is the following.

Proposition (Cramér-Wold device):

Let $\{\mathbf{b}_T\}$ be a sequence of random $k\times 1$ vector and suppose that for every real $k\times 1$ vector $\boldsymbol{\lambda}$ (such that $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$?), the scalar $\boldsymbol{\lambda}'\mathbf{b}_T \overset{A}{\sim} \boldsymbol{\lambda}'\mathbf{z}$ where $\mathbf{z}$ is a $k\times 1$ vector with joint (cumulative) distribution function F. Then the limiting distribution function of $\mathbf{b}_T$ exists and equals to F.

Lemma:
If $b_T \xrightarrow{L} Z$, then $b_T = O_p(1)$.

Lemma (Product rule):
Recall that if $A_T = o_p(1)$ and $b_T = O_p(1)$, then $A_T b_T = o_p(1)$. Hence, if $A_T \xrightarrow{p} 0$ and $b_T \xrightarrow{d} Z$, then $A_T b_T \xrightarrow{p} 0$.

Lemma (Asymptotic equivalence):
Let $\{a_T\}$ and $\{b_T\}$ be two sequence of random vectors. If $a_T - b_T \xrightarrow{p} 0$ and $b_T \xrightarrow{d} Z$, then $a_T \xrightarrow{d} Z$.

The results is helpful in situation in which we wish to find the asymptotic distribution of a_T but cannot do so directly. Often, however, it is easy to find a b_T that has a known asymptotic distribution and that satisfies $a_T - b_T \xrightarrow{p} 0$. This Lemma then ensures that a_T has the same limiting distribution as b_T and we say that a_T is "asymptotically equivalent" to b_T.

Lemma:
Given $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^l$ $(k, l \in \mathbb{N})$ and any sequence of random $k \times 1$ vector $\mathbf{b}_T$ such that $\mathbf{b}_T \xrightarrow{L} \mathbf{z}$, where $\mathbf{z}$ is $k \times 1$, if $\mathbf{g}$ is continuous (not dependent on T) at $\mathbf{z}$, then $\mathbf{g}(\mathbf{b}_T) \xrightarrow{L} \mathbf{g}(\mathbf{z})$.

Example:
Suppose that $X_T \xrightarrow{L} N(0,1)$ Then the square of X_T asymptotically behaves as the square of a $N(0,1)$ variables: $X_T^2 \xrightarrow{L} \chi^2_{(1)}$.

Lemma:
Let $\{\mathbf{x}_T\}$ be a sequence of random $(n \times 1)$ vector with $\mathbf{x}_T \xrightarrow{p} \mathbf{c}$, and let $\{\mathbf{y}_T\}$ be a sequence of random $(n \times 1)$ vector with $\mathbf{y}_T \xrightarrow{L} \mathbf{y}$. Then the sequence constructed from the sum $\{\mathbf{x}_T + \mathbf{y}_T\}$ converges in distribution to $\mathbf{c} + \mathbf{y}$ and the sequence constructed from the product $\{\mathbf{x}'_T \mathbf{y}_T\}$ converges in distribution to $\mathbf{c}'\mathbf{y}$.

Example:
Let $\{\mathbf{X}_T\}$ be a sequence of random $(m \times n)$ matrix with $\mathbf{X}_T \xrightarrow{p} \mathbf{C}$, and let $\{\mathbf{y}_T\}$ be a sequence of random $(n \times 1)$ vector with $\mathbf{y}_T \xrightarrow{L} \mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

Then the limiting distribution of $\mathbf{X}_T\mathbf{y}_T$ is the same as that of $\mathbf{Cy}$; that is $\mathbf{X}_T\mathbf{y}_T \xrightarrow{L} N(\mathbf{C}\boldsymbol{\mu}, \mathbf{C\Omega C}')$.

Lemma (Cramer δ):
Let $\{\mathbf{x}_T\}$ be a sequence of random $(n \times 1)$ vector such that

$$T^b(\mathbf{x}_T - \mathbf{a}) \xrightarrow{L} \mathbf{x}$$

for some $b > 0$. If $\mathbf{g}(\mathbf{x})$ is a real-valued function with gradient $\mathbf{g}'(\mathbf{a})(= \frac{\partial \mathbf{g}}{\partial \mathbf{x}'}\big|_{\mathbf{x}=\mathbf{a}})$, then

$$T^b(\mathbf{g}(\mathbf{x_T}) - \mathbf{g}(\mathbf{a})) \xrightarrow{L} \mathbf{g}'(\mathbf{a})\mathbf{x}.$$

Example:
Let $\{Y_1, Y_2, ..., Y_T\}$ be an *i.i.d.* sample of size T drawn from a distribution with mean $\mu \neq 0$ and variance σ^2. Consider the distribution of the reciprocal of the sample mean, $S_T = 1/\bar{Y}_T$, where $\bar{Y}_T = (1/T)\sum_{t=1}^T Y_t$. We know from the CLT that $\sqrt{T}(\bar{Y}_T - \mu) \xrightarrow{L} Y$, where $Y \sim N(0, \sigma^2)$. Also, $g(y) = 1/y$ is continuous at $y = \mu$. Let $g'(u)(= \partial g/\partial y|y = \mu) = (-1/\mu^2)$. Then $\sqrt{T}[S_T - (1/\mu)] \xrightarrow{L} g'(\mu)Y$; in other word, $\sqrt{T}[S_T - (1/\mu)] \xrightarrow{L} N(0, \sigma^2/\mu^4)$.

3 Martingales

The concept of conditional expectation provides us with an ideal link between the theory of random variables and that of stochastic processes. This is because the notion of conditional expectation enables us to formalize the temporal dependence in a stochastic process $\{X_t, t \in \mathcal{T}\}$ in terms of the conditional expectation of the process at time t, X_t (the present) given $X_{t-1}, X_{t-2}....$ (the past).[3] One important application of conditional expectation in such a context is in connection with a stochastic process with forms a martingales. Some very useful limit theorems pertain to *martingale sequence*.

Definition:

Let $\{X_t, t \in \mathcal{T}\}$ be a stochastic process defined on $(\mathcal{S}, \mathcal{F}, P(\cdot))$ and let $\{\mathcal{F}_t\}$ be a sequence of $\sigma - fields$ $\mathcal{F}_t \subset \mathcal{F}$ for all t (i.e.$\{\mathcal{F}_t\}$ is an increasing sequence of $\sigma - fields$) satisfying the following conditions:
(a). X_t is a random variable relatives to $\{\mathcal{F}_t\}$ for all $t \in \mathcal{T}$.
(b). $E(|X_t|) < \infty$ for all $t \in \mathcal{T}$.
(c). $E(X_t|\mathcal{F}_{t-1}) = X_{t-1}$, for all $t \in \mathcal{T}$.

Then $\{X_t, t \in \mathcal{T}\}$ is said to be a **martingale** with respect to $\{\mathcal{F}_t, t \in \mathcal{T}\}$.

Example (increasing sequence of $\sigma - fields$):

Define the function X—"the number of heads", then $X(\{HH\}) = 2$, $X(\{TH\}) = 1$, $X(\{HT\}) = 1$, and $X(\{TT\}) = 0$. Further we see that $X^{-1}(2) = \{(HH)\}$, $X^{-1}(1) = \{(TH), (HT)\}$ and $X^{-1}(0) = \{(TT)\}$. In fact, it can be shown that the $\sigma - field$ related to the random variables, X, so defined is

$$\begin{aligned}\mathcal{F} = \{S, \varnothing, \{(HH)\}, \{(TT)\}, \{(TH), (HT)\}, \{(HH), (TT)\},\\ \{(HT), (TH), (HH)\}, \{(HT), (TH), (TT)\}\}.\end{aligned}$$

We further define the function X_1—"at least one head", then $X_1(\{HH\}) = X_1(\{TH\}) = X_1(\{HT\}) = 1$, and $X_1(\{TT\}) = 0$. Further we see that $X_1^{-1}(1) = \{(HH), (TH), (HT)\} \in \mathcal{F}$ and $X^{-1}(0) = \{(TT)\} \in \mathcal{F}$. In fact, it can be shown that the $\sigma - field$ related to the random variables, X_1, so defined is

$$\mathcal{F}_1 = \{S, \varnothing, \{(HH), (TH), (HT)\}, \{(TT)\}\}.$$

[3] One of the more powerful economic theories is the theory of rational expectations.

Finally we define the function X_2—"two heads", then $X_2(\{HH\}) = 1$, $X_2(\{TH\}) = X_2(\{HT\}) = X_2(\{TT\}) = 0$. Further we see that $X_2^{-1}(1) = \{(HH)\} \in \mathcal{F}$, $X^{-1}(0) = \{(TH), (HT), (TT)\} \in \mathcal{F}$. In fact, it can be shown that the $\sigma - field$ related to the random variables, X_2, so defined is

$$\mathcal{F}_2 = \{S, \varnothing, \{(HH)\}, \{(HT), (TH), (TT)\}\}.$$

We see that $X = X_1 + X_2$ and find that $\mathcal{F}_1 \subset \mathcal{F}$.

The above example is a special case of general result where $X_1, X_2, ..., X_n$ are random variables on the same probability space $(\mathcal{S}, \mathcal{F}, P(\cdot))$ and we define the new random variables

$$Y_1 = X_1,\ Y_2 = X_1 + X_2,\ Y_3 = X_1 + X_2 + X_3, ...,\ Y_n = X_1 + X_2 + ... + X_n.$$

If $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ denote the minimal $\sigma - field$ generated by $Y_1, Y_2, ..., Y_n$ respectively, we can show that

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_n \subseteq \mathcal{F},$$

i.e. $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ form an increasing sequence of $\sigma - field$ in $\mathcal{F}$.

Several aspects of this definition need commenting on.
(a). A martingale is a relative concept; a stochastic process relative to an increasing sequence of $\sigma - field$. That is, $\sigma - field$ such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_t \subset ...$ and each X_t is a random variables relative to $\mathcal{F}_t, t \in \mathcal{T}$. A natural choice for such $\sigma - field$ will be $\mathcal{F}_t = \sigma(X_t, X_{t-1}, ..., X_1), t \in \mathcal{T}$.
(b). This stochastic process has constant mean because $E(X_t) = E[E(X_t|\mathcal{F}_{t-1})] = E(X_{t-1})$.
(c). It implies that $E(X_{t+\tau}|\mathcal{F}_{t-1}) = X_{t-1}$ for all $t \in \mathcal{T}$ and $\tau \geq 0$. That is the best predictor of $X_{t+\tau}$ given the information $\mathcal{F}_{t-1}$ is X_{t-1} for any $\tau \geq 0$.

The importance of martingales stem from the fact that they are general enough to include most forms of stochastic process of interest in economic modeling as special case, and restrictive enough so as to allow various limit theorem needed for their statistical analysis to go through, thus making probability models based on martingale largely operational.

Example:

Let $\{Z_t, t \in \mathcal{T}\}$ be a sequence of independent random variables such that $E(Z_t) = 0$ for all $t \in \mathcal{T}$. If we define X_t by[4]

$$X_t = \sum_{k=1}^{t} Z_k,$$

then $\{X_t, t \in \mathcal{T}\}$ is a martingale with $\mathcal{F}_t = \sigma(Z_t, Z_{t-1}, ..., Z_1) = \sigma(X_t, X_{t-1}, ..., X_1)$. This is because condition (a) and (b) are automatically satisfied and we can verify that

$$E(X_t|\mathcal{F}_{t-1}) = E[(X_{t-1} + Z_t)|\mathcal{F}_{t-1}] = X_{t-1},\ t \in \mathcal{T}.$$

Example:

Let $\{Z_t, t \in \mathcal{T}\}$ be an arbitrary stochastic process whose only restriction is $E(|Z_t|) < \infty$ for all $t \in \mathcal{T}$. If we define X_t by

$$X_t = \sum_{k=1}^{t} [Z_k - E(Z_k|\mathcal{F}_{k-1})],$$

where $\mathcal{F}_k = \sigma(Z_k, Z_{k-1}, ..., Z_1) = \sigma(X_k, X_{k-1}, ..., X_1)$, then $\{Z_t, t \in \mathcal{T}\}$ is a martingale. Condition (c) can verify by

$$\begin{aligned} E(X_t|\mathcal{F}_{t-1}) = E[(X_{t-1} + Z_t - E(Z_t|\mathcal{F}_{t-1}))|\mathcal{F}_{t-1}] &= X_{t-1} + E(Z_t|\mathcal{F}_{t-1}) - E(Z_t|\mathcal{F}_{t-1}) \\ &= X_{t-1},\ t \in \mathcal{T}. \end{aligned}$$

The above two examples illustrate the flexibility of martingales very well. As we can see, the main difference between then is that in the first example, X_t is a linear function of independent r.v.'s and in the second example as a linear function of dependent r.v's centred at their conditional means. A special case of example is that

$$Y_t = X_t - E(X_t|\mathcal{F}_{t-1}),\ \ t \in \mathcal{T}.$$

It can be easily verified that $\{Y_t,\ t \in \mathcal{T}\}$ defines what is known as a $martingale\ difference$ process relative to $\mathcal{F}_t$ because

$$E(Y_t|\mathcal{F}_{t-1}) = 0\ \ t \in \mathcal{T}.$$

[4] Therefore, here X_t is a pure random walk process, $I(1)$.

We can further deduce that for $t > k$

$$\begin{aligned} E(Y_tY_k) &= E[E(Y_tY_k|\mathcal{F}_{t-1})] \quad (since\ for\ t > k,\ E(Y_k|\mathcal{F}_{t-1}) = Y_k) & (1)\\ &= E[Y_kE(Y_t|\mathcal{F}_{t-1})] & (2)\\ &= E[Y_k \cdot 0] = 0. & (3) \end{aligned}$$

That is, a martingale difference $\{Y_t,\ t \in \mathcal{T}\}$ as an orthogonal sequence. (A special case of uncorrelateness, for their means are all zero).

Definition:

A stochastic process $\{Y_t, t \in \mathcal{T}\}$ is said to be a **martingale difference** process relative to the increasing sequence of $\sigma - fields$, $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_t \subset ...$ if
(a). Y_t is a random variable relatives to $\{\mathcal{F}_t\}$ for all $t \in \mathcal{T}$.
(b). $E(|Y_t|) < \infty$ for all $t \in \mathcal{T}$.
(c). $E(Y_t|\mathcal{F}_{t-1}) = 0$, for all $t \in \mathcal{T}$.

Note that condition (c) is stronger than the conditions that Y_t is serially uncorrelated as we can see that if Y_t is a martingale difference then it is uncorrelated from (3). From the point of forecasting, a serially uncorrelated sequence cannot be forecast on the basis of a linear function of its past value since the forecast error and the forecast are all linear functions .No function of past values, linear or nonlinear, can forecast a martingale difference sequence. While stronger than absence of serial correlation, the martingale difference condition is weaker than independence, since it doesn't rule out the possibility that higher moments such as $E(Y_t^2|Y_{t-1}, Y_{t-2}, ..., Y_1)$ might depend on past Y's.

Example:

If $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$, then $Y_t = \varepsilon_t\varepsilon_{t-1}$ is a martingale difference but not serially independent since

$$E(Y_t|\mathcal{F}_{t-1}) = E(\varepsilon_t\varepsilon_{t-1}|\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_1) = \varepsilon_{t-1}E(\varepsilon_t) = 0, \quad (martingale\ difference)$$

and

$$E(Y_t^2|\mathcal{F}_{t-1}) = E(\varepsilon_t^2\varepsilon_{t-1}^2|\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_1) = \varepsilon_{t-1}^2E(\varepsilon_t^2) = \varepsilon_{t-1}^2\sigma^2 \quad (a\ function\ of\ past\ value, so\ it\ is\ not\ independent)$$

𝔓𝔯𝔬𝔭𝔬𝔰𝔦𝔱𝔦𝔬𝔫:

Let X and Y be independent random variables and let $U = g(X)$ and $V = h(Y)$. Then U and V are also independent random variables.

3.1 Martingale Difference in a Regression Context

The martingale difference assumption often arises in regression context in the following way. Suppose we have observations on a scalar Y_t that we are interested in explaining or forecasting on the basis variables $\mathbf{z}_t$ as well as on the basis of the past value of Y_t. Let $\mathcal{F}_{t-1}$ be the σ-field containing the information used to explain or forecast Y_t, i.e., $\mathcal{F}_{t-1} = \sigma(..., (\mathbf{z}'_{t-1}, Y_{t-2})', (\mathbf{z}'_t, Y_{t-1})')$. Then

$$E(Y_t|\mathcal{F}_{t-1}) = g(..., (\mathbf{z}'_{t-1}, Y_{t-2})', (\mathbf{z}'_t, Y_{t-1})'),$$

where g is some function of current and past value of $\mathbf{z}_t$ and past values of Y_t. Let $\mathbf{x}_t$ contain a finite number of current and lagged value of $(\mathbf{z}'_t, Y_{t-1})$, e.g., $\mathbf{x}_t = ((\mathbf{z}'_{t-\tau}, Y_{t-\tau-1})', ..., (\mathbf{z}'_t, Y_{t-1})')$ for some $\tau < \infty$. Economic theory is then often used in an attempt to justify the assumption that for some $\boldsymbol{\beta}_0 < \infty$,

$$g(..., (\mathbf{z}'_{t-1}, Y_{t-2})', (\mathbf{z}'_t, Y_{t-1})') = \mathbf{x}'_t\boldsymbol{\beta}_0. \ (\textit{Linear Function})$$

If this is true, we then have

$$E(Y_t|\mathcal{F}_{t-1}) = \mathbf{x}'_t\boldsymbol{\beta}_0.$$

We find that $\{Y_t - E(Y_t|\mathcal{F}_{t-1}), \mathcal{F}_t\}$ is a martingale difference sequence since

$$E[Y_t - E(Y_t|\mathcal{F}_{t-1})|\mathcal{F}_{t-1}] = E(Y_t|\mathcal{F}_{t-1}) - E(Y_t|\mathcal{F}_{t-1}) = 0.$$

If we let

$$\varepsilon_t = Y_t - \mathbf{x}'_t\boldsymbol{\beta}_0$$

and it is true that $E(Y_t|\mathcal{F}_{t-1}) = \mathbf{x}'_t\boldsymbol{\beta}_0$, then $\varepsilon_t = Y_t - E(Y_t|\mathcal{F}_{t-1})$, so $\{\varepsilon_t, \mathcal{F}_t\}$ is a martingale difference sequence. Of direct importance for least squares estimation is that each sequence of cross products between regressors and error $\{X_{ti}\varepsilon_t, \mathcal{F}_t\}, i = 1, ..., k$ is also a martingale difference sequence,

$$E[X_{ti}\varepsilon_t|\mathcal{F}_{t-1}] = X_{ti}E(\varepsilon_t|\mathcal{F}_{t-1}) = 0.$$

4 Laws of Larger Numbers

In this section we study a familiar consistent estimator from the concept of strong consistency (which automatically imply weakly consistency, or convergence in probability).

The result that the sample mean is a consistent estimator of the population mean is known as the *law of large number*. The law of large number we consider are all of the following form.

Proposition:
Given restriction on the dependence, heterogeneity, and moments of a sequence of random variables (you may think this sequence as a sample of size T) $\{Z_t\}$,

$$\bar{Z}_T - \bar{\mu}_T \xrightarrow{a.s.} 0,$$

where

$$\bar{Z}_T \equiv \frac{1}{T}\sum_{t=1}^{T} Z_t \ \ and \ \ \bar{\mu}_T \equiv E(\bar{Z}_T).$$

As we shall see, there are sometimes trade-off among theses restrictions; for example, relaxing dependence or heterogeneity restrictions may require strengthening moment restriction.

4.1 Independent Identically distributed Observations

The simplest case is that of independent identically distributed (*i.i.d.*) random variables.

Theorem (Kolmogorov):
Let $\{Z_t\}$ be a sequence of *i.i.d.* random variables. Then

$$\bar{Z}_T \xrightarrow{a.s.} \mu$$

which implies

$$\bar{Z}_T \xrightarrow{p} \mu$$

if and only if $E|Z_t| < \infty$ and $E(Z_t) = \mu$.

Example:

We may make a stronger assumption that $Var(Z_t) = \sigma^2$, then

$$E(\bar{Z}_T - \mu)^2 = (1/T^2)Var(\sum_{t=1}^{T} Z_t) = (1/T^2)\sum_{t=1}^{T} Var(Y_t) = \sigma^2/T.$$

Since $\sigma^2/T \to 0$ as $T \to \infty$, the is mean that $\bar{Z}_T \xrightarrow{q.m.} \mu$, implying also $\bar{Z}_T \xrightarrow{p} \mu$.

4.2 Independent Heterogeneously distributed Observations

For cross-sectional data, it is often appropriate to assume that the observation are independent but not identically distributed. A law of large number useful in these situation is the following.

Theorem (Revised Markov):

Let $\{Z_t\}$ be a sequence of independent random variables such that $E|Z_t|^{1+\delta} < \triangle < \infty$ for some $\delta > 0$ and all t. Then

$$\bar{Z}_T \xrightarrow{a.s.} \bar{\mu}_T.$$

The above theorem impose slightly more in the way of moment restriction but allows the observations to be rather heterogeneous.

4.3 Dependent Identically Distributed Observations (such as a strongly stationary process)

The assumption of independence is inappropriate for economic time series, which typically exhibit considerable dependence. To cover this case, we need laws of large number that allow the random variables to be dependent. To be state below, we need an additional 'memory restriction' as we relax the independence assumption.

Definition:

Let $(S, \mathcal{F}, P(\cdot))$ be a probability space and $\mathcal{T}$ an index set of real numbers and define the function $X(\cdot,\cdot)$ by $X(\cdot,\cdot) : S \times \mathcal{T} \to \mathbb{R}$. The order sequence of random variables $\{X(\cdot,t),\ t\in\mathcal{T}\}$ is called a stochastic process.

Definition:

A stochastic process $\{X(\cdot,t),\ t\in\mathcal{T}\}$ is said to be (strongly) stationary if any subset $(t_1,t_2,...,t_T)$ of $\mathcal{T}$ and any τ, $F(X(t_1),...,X(t_T))=F(X(t_1+\tau),...,X(t_T+\tau))$.

In terms of the marginal distributions $F(X(t)), t\in\mathcal{T}$, stationary implies that $F(X(t))=F(X(t+\tau))$, and hence $F(X(t_1))=F(X(t_2))=...=F(X(t_T))$. That is stationarity implies that $X(t_1),...,X(t_T)$ are individual identically distributed.

Definition:

Let $(S,\mathcal{F},P(\cdot))$ be a probability space. Let $\{Z_t\}$ be a strongly stationary sequence and let K be the measure-preserving transformation function. Then $\{Z_t\}$ is ergodic if

$$\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}Pr(F\cap K^tG)=Pr(F)Pr(G),$$

for all events $F,G\in\mathcal{F}$, where K is defined on $(S,\mathcal{F},P(\cdot))$ such that $Z_1(s)=Z_1(s),\ Z_2(s)=Z_1(Ks),\ Z_3(s)=Z_1(K^2s),...,Z_T(s)=Z_1(K^{T-1}s)$ for all $s\in S$.

We can think of K^tG as being the event G shifted t periods into the future, and since $Pr(K^tG)=Pr(G)$ when K is measure preserving, this definition say that an ergodic process is one such that for any events F and G, F and K^tG are independent on average in the limit. Thus ergodicity can be thought of as a form of "average asymptotic independence".

Theorem (Ergodic Theorem):

Let $\{Z_t\}$ be a strongly stationary ergodic scalar random sequence with $E|Z_t|<\infty$. Then

$$\bar{Z}_T\xrightarrow{a.s.}\mu\equiv E(Z_t).$$

Lemma:

A stationary linear process is ergodic.

Example:

Let $X_t=\sum_{j=0}^{\infty}\varphi_j\varepsilon_{t-j},\quad t=1,2,...;\ \varepsilon_{t-j},\ j=0,1,...$ are *i.i.d.* random variables with $E(\varepsilon_{t-j}) = 0$ and $\{\varphi_j, j \geq 0\}$ is a sequence of square summable real number. Then X_t is ergodic. (see Wang et al. 2003 p.151)

In this example we see that by relaxing the assumption of X_t to be weakly stationary (or ε_t is a white noise sequence), we need the stronger conditions that $\{\varphi_j, j \geq 0\}$ is a sequence of absolute summable real number to make X_t be ergodic (See Hamilton, p.52).

4.4 Dependent Heterogeneously Distributed Observations

By replacing the ergodicity assumption with somewhat stronger conditions, we can apply the consistency results to dependent heterogeneously distributed observations.

Let $\mathcal{B}_1^t$ denote the $\sigma - field$ generated $X_1, , ..., X_t$ where $\{X_t, t \in \mathcal{T}\}$ is a stochastic process. A measure of the dependence among the elements of the stochastic process can be defined in terms of the events $B \in \mathcal{B}_{-\infty}^t$ and $A \in \mathcal{B}_{t+\tau}^{\infty}$ by

$$\alpha(\tau) = \sup_{\tau} |P(A \cap B) - P(A)P(B)|.$$

Definition:
A stochastic process $\{X_t, t \in \mathcal{T}\}$ is said to be strongly (or α) mixing if $\alpha(\tau) \rightarrow 0$ as $\tau \rightarrow \infty$.

A stronger form of mixing, called uniformly mixing, can be defined in terms of the following measure of dependence:

$$\phi(\tau) = \sup_{\tau} |P(A|B) - P(A)|, \quad P(B) > 0.$$

Definition:
A stochastic process $\{X_t, t \in \mathcal{T}\}$ is said to be uniformly (or ϕ) mixing if $\phi(\tau) \rightarrow 0$ as $\tau \rightarrow \infty$.

The notation of mixing is a stronger memory requirement than that of ergodicity for stationary sequences, since given stationarity, mixing implies ergodicity.

Proposition:
Let $\{Z_t\}$ be a stationary sequence. If $\alpha(\tau) \to 0$ as $\tau \to 0$, then $\{Z_t\}$ is ergodic.

Definition:
Let $a \in \mathbb{R}$.
(i). If $\alpha(\tau) = O(\tau^{-a-\varepsilon})$ for some $\varepsilon > 0$, then α is of size $-a$.
(ii). If $\phi(\tau) = O(\tau^{-a-\varepsilon})$ for some $\varepsilon > 0$, then ϕ is of size $-a$.

This definition allows precise statements about memory of a random sequence that we shall relate to moment condition expressed in terms of a. As a get smaller, the sequence exhibits more and more dependence.[5]

Theorem (Revised McLeish):
Let $\{Z_t\}$ be a random sequence with
(i) $E|Z_t|^{r+\delta} < \triangle < \infty$ for some $\delta > 0$ and all t, and
(ii) $\{Z_t\}$ is α-mixing with α of size $-r/(r-1)$, $r > 1$, or is a ϕ-mixing with ϕ of size $-r/(2r-1)$, $r \geq 1$. Then

$$\bar{Z}_T \xrightarrow{a.s.} \bar{\mu}_T.$$

For sequences with longer memories, r is greater $(r/(r-1) = 1+1/(r-1) = a)$, and the moment restrictions increase accordingly. Hence we have a clear trade-off between the amount of allowable dependence and the sufficient moment restrictions.

4.5 Asymptotic Uncorrelated Observations (such as a weakly stationary $ARMA$ process)

Although mixing is an appealing dependence concept, it shares with ergodicity the properties that it can be somewhat difficult to verify theoretically and is impossible to verify empirically. An alternative dependence concept that is easier to verify theoretically is a form of asymptotic non-correlation.

Theorem :
Let $\{Z_t\}$ is an asymptotically uncorrelated scalar sequence with means $\mu_t \equiv E(Z_t)$

[5] Think of $T^1 \cdot \frac{1}{T}$ and $T^2 \cdot \frac{1}{T^2}$.

and $\sigma_t^2 \equiv var(Z_t) < \infty$. Then

$$\bar{Z}_T \xrightarrow{a.s.} \bar{\mu}_T.$$

Compared with last Theorem, we have relaxed the dependence restriction from asymptotic independence (mixing) to asymptotic uncorrelation, but we have altered the moment requirements from restrictions on moments of order $r + \delta$ $(r \geq 1, \delta > 0)$ to second moments.

Example (Law of large numbers for a covariance-stationary process):
Let $(Y_1, Y_2, ..., Y_T)$ represent a sample of size T from a covariance-stationary process with

$$\begin{aligned} E(Y_t) &= \mu, \quad for\ all\ t \\ E(Y_t - \mu)(Y_{t-j} - \mu) &= \gamma_j, \quad for\ all\ t \\ \sum_{j=0}^{\infty} |\gamma_j| &< \infty. \end{aligned}$$

Then

$$\bar{Y}_T \xrightarrow{q.m} \mu.$$

Proof:

To see this, it suffices to show that $E(\bar{Y}_T-\mu)^2 \longrightarrow 0$. Since

$$
\begin{aligned}
&E(\bar{Y}_T-\mu)^2\\
&= E\left[(1/T)\sum_{t=1}^{T}(Y_t-\mu)\right]^2\\
&= (1/T^2)E\{(Y_1-\mu)[(Y_1-\mu)+(Y_2-\mu)+...+(Y_T-\mu)]\\
&\quad +(Y_2-\mu)[(Y_1-\mu)+(Y_2-\mu)+...+(Y_T-\mu)]\\
&\quad +(Y_3-\mu)[(Y_1-\mu)+(Y_2-\mu)+...+(Y_T-\mu)]\\
&\quad +...+(Y_T-\mu)[(Y_1-\mu)+(Y_2-\mu)+...+(Y_T-\mu)]\}\\
&= (1/T^2)\{[\gamma_0+\gamma_1+\gamma_2+\gamma_3+...+\gamma_{T-1}]\\
&\quad +[\gamma_1+\gamma_0+\gamma_1+\gamma_2+...+\gamma_{T-2}]\\
&\quad +[\gamma_2+\gamma_1+\gamma_0+\gamma_1+...+\gamma_{T-3}]\\
&\quad +...+[\gamma_{T-1}+\gamma_{T-2}+\gamma_{T-3}+...+\gamma_0]\}\\
&= (1/T^2)\{T\gamma_0+2(T-1)\gamma_1+2(T-2)\gamma_2+...+2\gamma_{T-1}\}\\
&= (1/T)\{\gamma_0+[(T-1)/T](2\gamma_1)+[(T-2)/T](2\gamma_2)+...+[1/T](2\gamma_{T-1})\}\\
&= (1/T)|\gamma_0+[(T-1)/T](2\gamma_1)+[(T-2)/T](2\gamma_2)+...+[1/T](2\gamma_{T-1})|,
\end{aligned}
$$

then

$$
\begin{aligned}
T\cdot E(\bar{Y}_T-\mu)^2 &= |\gamma_0+[(T-1)/T](2\gamma_1)+[(T-2)/T](2\gamma_2)+...+[1/T](2\gamma_{T-1})| \qquad (4)\\
&\le \{|\gamma_0|+[(T-1)/T]\cdot 2|\gamma_1|+[(T-2)/T]\cdot 2|\gamma_2|+...+[1/T]\cdot 2|\gamma_{T-1}|\}\\
&\le \{|\gamma_0|+2|\gamma_1|+2|\gamma_2|+...\}\\
&< \infty.
\end{aligned}
$$

So, $E(\bar{Y}_T-\mu)^2 \longrightarrow 0$.

4.6 Martingale Difference Sequences

A law of large numbers for martingale difference sequence is the following theorem.

Theorem (Revised Chow):
Let $\{Z_t,\mathcal{F}_t\}$ be a martingale difference sequence such that $E|Z_t|^{2r}<\triangle<\infty$ for some $r\ge 1$ and all t. Then

$$\bar{Z}_T \xrightarrow{a.s.} 0.$$

5 Central Limit Theory

In this section we study various form of central limit theorem (CLT) from the concept of convergence in distribution.

The central limit theorem we consider are all of the following form:

Proposition:
Given restriction on the dependence, heterogeneity, and moments of a sequence of random variables (you may think this sequence as a sample of size T) $\{Z_t\}$,

$$\frac{(\bar{Z}_T - \bar{\mu}_T)}{(\bar{\sigma}_T/\sqrt{T})} = \frac{\sqrt{T}(\bar{Z}_T - \bar{\mu}_T)}{\bar{\sigma}_T} \xrightarrow{L} N(0,1),$$

where[6]

$$\bar{Z}_T \equiv \frac{1}{T}\sum_{t=1}^{T} Z_t,\ \ \bar{\mu}_T \equiv E(\bar{Z}_T),\ \textit{and}\ \ \bar{\sigma}_T^2/T \equiv var(\bar{Z}_T)\ \ (\textit{that is}\ \bar{\sigma}_T^2 = \frac{var(\sum_{t=1}^T Z_t)}{T}).$$

As with the law of large numbers, there are natural trade-off among theses restrictions. Typically, greater dependence or heterogeneity restrictions is allowed at the expanse of requiring strengthening moment restriction.

5.1 Independent Identically distributed Observations

As with laws of large numbers, the case of *i.i.d.* observations is the simplest.

Theorem (Linderberg-Lévy).
Let $\{Z_t\}$ be a sequence of *i.i.d.* random scalars, with $\mu \equiv E(Z_t)$ and $\sigma^2 \equiv var(Z_t) < \infty$. If $\sigma^2 \neq 0$, then

$$\begin{aligned}\frac{\sqrt{T}(\bar{Z}_T - \bar{\mu}_T)}{\bar{\sigma}_T} &= \frac{\sqrt{T}(\bar{Z}_T - \mu)}{\sigma} \\ &= \frac{T^{-1/2}\sum_{t=1}^T (Z_t - \mu)}{\sigma} \xrightarrow{L} N(0,1).\end{aligned}$$

Compared with the law of large number for *i.i.d.* observations, we impose a single additional requirement, i.e., that $\sigma^2 \equiv var(Z_t) < \infty$. Note that this implies

[6] To see why this notation, notice that $Var(\bar{Z}_T) = \frac{Var(\sum Z_t)}{T^2} = \frac{Var(\sum Z_t)/T)}{T} = \frac{\bar{\sigma}_T^2}{T}$, that is, we assume $Var(\sum Z_t)$ is $O(T^1)$.

that $E|Z_t| < \infty$.

𝔓𝔯𝔬𝔭𝔬𝔰𝔦𝔱𝔦𝔬𝔫:

If the kth moment of a random variable exists, all moments of order less than k exist.

𝔓𝔯𝔬𝔬𝔣:

Let $f_X(x)$ be the pdf of X. $E(X^k)$ exists if and only if

$$\int_{-\infty}^{\infty} |x|^k \cdot f_X(x)dx < \infty.$$

Let $1 \leq j < k$, to prove the theorem we must show that

$$\int_{-\infty}^{\infty} |x|^j \cdot f_X(x)dx < \infty.$$

But

$$\begin{aligned}
\int_{-\infty}^{\infty} |x|^j \cdot f_X(x)dx &= \int_{-\infty}^{|x|\leq 1} |x|^j \cdot f_X(x)dx + \int_{|x|>1}^{\infty} |x|^j \cdot f_X(x)dx \\
&\leq \int_{-\infty}^{|x|\leq 1} f_X(x)dx + \int_{|x|>1}^{\infty} |x|^j \cdot f_X(x)dx \\
&\leq 1 + \int_{|x|>1}^{\infty} |x|^j \cdot f_X(x)dx \\
&\leq 1 + \int_{|x|>1}^{\infty} |x|^k \cdot f_X(x)dx < \infty.
\end{aligned}$$

5.2 Independent Heterogeneously Distributed Observations

Several different central limit theorems are available for the case in which our observations are not identically distributed.

𝔗𝔥𝔢𝔬𝔯𝔢𝔪 (Liapounov, revised Lindeberg-Feller)

Let $\{Z_t\}$ be a sequence of independent random variables such that $\mu_t \equiv E(Z_t)$, $\sigma_t^2 \equiv var(Z_t)$ and $E|Z_t - \mu_t|^{2+\delta} < \triangle < \infty$ for some $\delta > 0$ and all t. If $\bar{\sigma}_T^2 > \delta' > 0$ for all T sufficiently large, then

$$\frac{\sqrt{T}(\bar{Z}_T - \bar{\mu}_T)}{\bar{\sigma}_T} \xrightarrow{L} N(0,1).$$

Note that $E|Z_t|^{2+\delta} < \triangle$ also implies that $E|Z_t - \mu_t|^{2+\delta}$ is uniformly bounded. Note also the analogy with previous results there we obtained a law of large numbers for independent random variables by imposing a uniform bound on $E|Z_t|^{1+\delta}$ Now we can obtain a central limit theorem imposing a uniform bound on $E|Z_t|^{2+\delta}$.

5.3 Dependent Identically Distributed Observations

In the last two section we saw that obtaining central limit theorems for independent process typically required strengthening the moments restrictions beyond what was sufficient for obtaining laws of large numbers. In the class of stationary ergodic process, not only will we strengthen the moment requirements, but we will also impose stronger conditions on the memory of the process.

𝔗𝔥𝔢𝔬𝔯𝔢𝔪 (Scott):
Let $\{Z_t, \mathcal{F}_t\}$ be a stationary ergodic adapted mixingale with γ_m of size -1. Then $\bar{\sigma}_T^2 \equiv var(T^{-1/2}\sum_{t=1}^T Z_t) \to \bar{\sigma}^2 < \infty$ as $T \to \infty$ and if $\bar{\sigma}^2 > 0$, then $T^{-1/2}\bar{Z}_T/\bar{\sigma} \xrightarrow{L} N(0,1)$.

5.4 Dependent Heterogeneously distributed Observations

𝔗𝔥𝔢𝔬𝔯𝔢𝔪 (Wooldridge-White):
Let $\{Z_t\}$ be a scalar random sequence with $\mu_t \equiv E(Z_t)$ and $\sigma_t^2 \equiv var(Z_t)$ such that $E|Z_t|^r < \triangle < \infty$ for some $r \geq 2$ for all t and having mixing coefficients ϕ of size $-r/2(r-1)$ or α of size $-r/(r-2)$, $r > 2$. If $\bar{\sigma}_T^2 \equiv var(T^{-1/2}\sum_{t=1}^T Z_t) > \delta > 0$ for all T sufficiently large, then $\sqrt{T}(\bar{Z}_T - \bar{\mu}_T)/\bar{\sigma}_T \xrightarrow{L} N(0,1)$.

5.5 Asymptotic Uncorrelated Observations (such as a stationary *ARMA* process)

We now present a central limit theorem for a serial correlated sequence.

𝔗𝔥𝔢𝔬𝔯𝔢𝔪:

Let

$$Y_t = \mu + \sum_{j=0}^{\infty} \varphi_j \varepsilon_{t-j},$$

where $\{\varepsilon_t\}$ is a sequence of $i.i.d.$ random variables with $E(\varepsilon_t^2) < \infty$ and $\sum_{j=0}^{\infty} |\varphi_j| < \infty$. Then

$$\sqrt{T}(\bar{Y}_T - \mu) \xrightarrow{L} N(0, \sum_{j=-\infty}^{\infty} \gamma_j).$$

𝔓𝔯𝔬𝔬𝔣:

Given this theorem, it suffices to shows that $\bar{\sigma}_T^2 (= T \cdot var(\bar{Y}_T) = [var(\sum_{t=1}^{T} Y_t)]/T) = \sum_{j=-\infty}^{\infty} \gamma_j)$ from the general form of CLT.

Note that the assumption $\sum_{j=0}^{\infty} |\varphi_j| < \infty$ implies that $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ and means that for any $\epsilon > 0$ there exist a q such that

$$2|\gamma_{q+1}| + 2|\gamma_{q+2}| + 2|\gamma_{q+3}| + ... < \epsilon/2.$$

From (4) we have

$$\begin{aligned}
&\left| \sum_{j=-\infty}^{\infty} \gamma_j - T \cdot var(\bar{Y}_T) \right| \\
&\quad = |\{\gamma_0 + 2\gamma_1 + 2\gamma_2 + 2\gamma_3 + ...\} \\
&\quad\quad - \{\gamma_0 + [(T-1)/T]2\gamma_1 + [(T-2)/T]2\gamma_2 + ... + [1/T]2\gamma_{T-1}\}| \\
&\quad \leq (1/T) \cdot 2|\gamma_1| + (2/T) \cdot 2|\gamma_2| + (3/T) \cdot 2|\gamma_3| + ... \\
&\quad\quad + (q/T) \cdot 2|\gamma_q| + 2|\gamma_{q+1}| + 2|\gamma_{q+2}| + 2|\gamma_{q+3}| + ... \\
&\quad \leq (1/T) \cdot 2|\gamma_1| + (2/T) \cdot 2|\gamma_2| + (3/T) \cdot 2|\gamma_3| + ... \\
&\quad\quad + (q/T) \cdot 2|\gamma_q| + \epsilon/2.
\end{aligned}$$

Moreover, for this given q, we can find an N such that

$$(1/T) \cdot 2|\gamma_1| + (2/T) \cdot 2|\gamma_2| + (3/T) \cdot 2|\gamma_3| + ... + (q/T) \cdot 2|\gamma_q| < \epsilon/2$$

for all $T \geq N$, ensuring that

$$\left| \sum_{j=-\infty}^{\infty} \gamma_j - T \cdot var(\bar{Y}_T) \right| < \epsilon.$$

This completes the proof.

5.6 Martingale Difference Sequences

𝔗𝔥𝔢𝔬𝔯𝔢𝔪:

Let $\{Y_t\}$ be a scalar martingale difference sequence with $\bar{Y}_T = (1/T)\sum_{t=1}^{T} Y_t$. Suppose that

(a). $E(Y_t^2) = \sigma_t^2 > 0$ with $(1/T)\sum_{t=1}^{T}\sigma_t^2 \to \sigma^2 > 0$,[7]

(b). $E|Y_t|^r < \infty$ for some $r > 2$ and all t, and

(c). $(1/T)\sum_{t=1}^{T} Y_t^2 \xrightarrow{p} \sigma^2$,

then

$$\frac{\sqrt{T}(\bar{Y}_T - 0)}{\sigma} \xrightarrow{L} N(0,1).$$

[7] A martingale difference sequence is a serially uncorrelated sequence.