CHAPTER 4. LIMIT THEOREMS IN STATISTICS

4.1. SEQUENCES OF RANDOM VARIABLES

4.1.1. A great deal of econometrics uses relatively large data sets and methods of statistical inference that are justified by their desirable properties in large samples. The probabilistic foundations for these arguments are "laws of large numbers", sometimes called the "law of averages", and "central limit theorems". This chapter presents these foundations. It concentrates on the simplest versions of these results, but goes some way in covering more complicated versions that are needed for some econometric applications. For basic econometrics, the most critical materials are the limit concepts and their relationship covered in this section, and for independent and identically distributed (i.i.d.) random variables the first Weak Law of Large Numbers in Section 4.3 and the first Central Limit Theorem in Section 4.4. The reader may want to postpone other topics, and return to them as they are needed in later chapters.

4.1.2. Consider a sequence of random variables $Y_1,Y_2,Y_3,...$. These random variables are all functions $Y_k(s)$ of the <u>same</u> state of Nature s, but may depend on different parts of s. There are several possible concepts for the limit Y_o of a sequence of random variables Y_n. Since the Y_n are functions of states of nature, these limit concepts will correspond to different ways of defining limits of functions. Figure 4.1 will be used to discuss limit concepts. Panel (a) graphs Y_n and Y_o as functions of the state of Nature. Also graphed are curves denoted $Y_{o\pm}$ and defined by $Y_o \pm \varepsilon$ which for each state of Nature s delineate an ε-neighborhood of $Y_o(s)$. The set of states of Nature for which $|Y_o(s) - Y_n(s)| > \varepsilon$ is denoted $\mathbf{W}_n$. Panel (b) graphs the CDF's of Y_o and Y_n. For technical completeness, note that a random variable Y is a measurable real-valued function on a probability space $(S,\boldsymbol{F},P)$, where $\boldsymbol{F}$ is a σ-field of subsets of **S**, P is a probability on $\boldsymbol{F}$, and "measurable" means that F contains the inverse image of every set in the Borel σ-field of subsets of the real line. The CDF of a vector of random variables is then a measurable function with the properties given in 3.5.3.

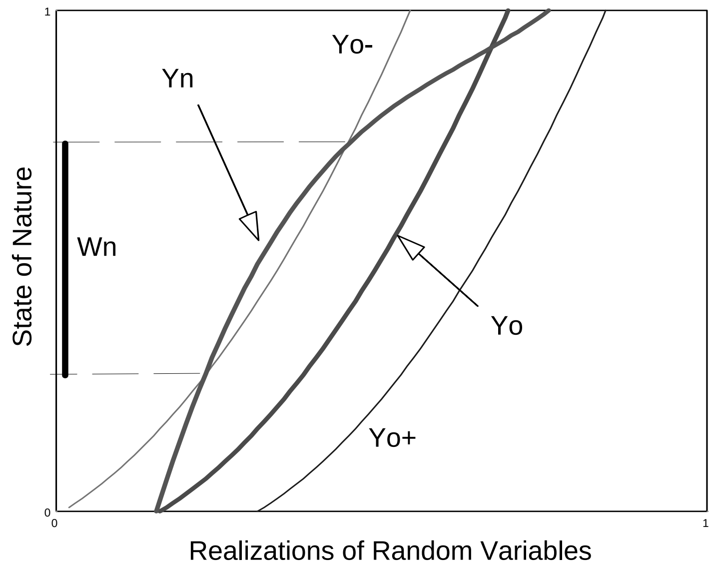
4.1.3. Y_n *converges in probability* to Y_o, if for each $\varepsilon > 0$, $\lim_{n\to\infty} \text{Prob}(|Y_n - Y_o| > \varepsilon) = 0$. Convergence in probability is denoted $Y_n \to_p Y_o$, or $\text{plim}_{n\to\infty} Y_n = Y_o$. With $\mathbf{W}_n$ defined as in Figure 4.1, $Y_n \to_p Y_o$ iff $\lim_{n\to\infty} \text{Prob}(\mathbf{W}_n) = 0$ for each $\varepsilon > 0$.

4.1.4. Y_n *converges almost surely* to Y_o, denoted $Y_n \to_{as} Y_o$, if for each $\varepsilon > 0$, $\lim_{n\to\infty} \text{Prob}(\sup_{m\ge n}|Y_m - Y_o| > \varepsilon) = 0$. For $\mathbf{W}_n$ defined in Figure 4.1, the set of states of nature for which $|Y_m(w) - Y_o(w)| > \varepsilon$ for <u>some</u> $m \ge n$ is $\cup_{m\ge n} \mathbf{W}_m$, and $Y_n \to_{as} Y_o$ iff $\text{Prob}(\cup_{m\ge n} \mathbf{W}_{n'}) \to 0$.

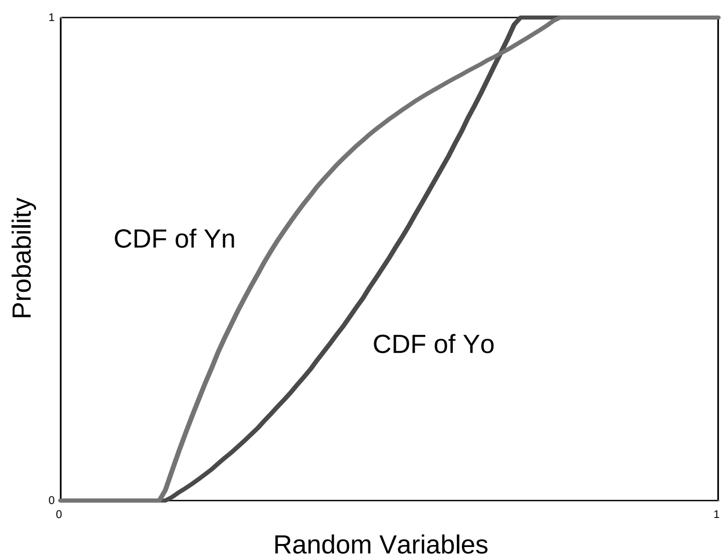
An implication of almost sure convergence is $\lim_{n\to\infty} Y_n(s) = Y_o(s)$ a.s. (i.e., except for a set of states of Nature of probability zero); this is <u>not</u> an implication of $Y_n \to_p Y_o$.

FIGURE 4.1. CONVERGENCE CONCEPTS FOR RANDOM VARIABLES

Panel (a)



Panel (b)



4.1.5. Y_n *converges in ρ-mean* (also called *convergence in* $\|\cdot\|_\rho$ *norm*, or *convergence in* $\mathbf{L}_\rho$ *space*) to Y_o if $\lim_{n\to\infty} \mathbf{E}|Y_n - Y_o|^\rho = 0$. For ρ = 2, this is called *convergence in quadratic mean*. The norm is defined as $\|Y\|_\rho = [\int_S |Y(s)|^\rho \cdot P(ds)]^{1/\rho} = [\mathbf{E}|Y|^\rho]^{1/\rho}$, and can be interpreted as a probability-weighted measure of the distance of Y from zero. The norm of a random variable is a moment. There are random variables for which the ρ-mean will not exist for any ρ > 0; for example, Y with CDF F(y) = 1 - 1/(log y) for $y \geq e$ has this property. However, in many applications moments such as variances exist, and the quadratic mean is a useful measure of distance.

4.1.6. Y_n *converges in distribution* to Y_o, denoted $Y_n \to_d Y_o$, if the CDF of Y_n converges to the CDF of Y_o at each continuity point of Y_o. In Figure 4.1(b), this means that F_n converges to the function F_o point by point for each argument on the horizontal axis, except possibly for points where F_o jumps. (Recall that distribution functions are always continuous from the right, and except at jumps are continuous from the left. Since each jump contains a distinct rational number and the rationals are countable, there are at most a countable number of jumps. Then the set of jump points has Lebesgue measure zero, and there are continuity points arbitrarily close to any jump point. Because of right-continuity, distribution functions are uniquely determined by their values at their continuity points.) If **A** is an open set, then $Y_n \to_d Y_o$ implies $\liminf_{n\to\infty} F_n(\mathbf{A}) \geq F_o(\mathbf{A})$; conversely, **A** closed implies $\limsup_{n\to\infty} F_n(\mathbf{A}) \leq F_o(\mathbf{A})$ see P. Billingsley (1968), Theorem 2.1. Convergence in distribution is also called *weak convergence* in the space of distribution functions.

4.1.7. The relationships between different types of convergence are summarized in Figure 4.2. In this table, "A $\models\!\!\Longrightarrow$ B" means that A implies B, but not vice versa, and "A $\Longleftrightarrow$ B" means that A and B are equivalent. Explanations and examples are given in Sections 4.1.8-4.1.18. On first reading, skim these sections and skip the proofs.

4.1.8. $Y_n \to_{as} Y_o$ implies $\text{Prob}(\mathbf{W}_n) \leq \text{Prob}(\cup_{m\geq n} \mathbf{W}_m) \to 0$, and hence $Y_n \to_p Y_o$. However, $\text{Prob}(\mathbf{W}_n) \to 0$ does not necessarily imply that the probability of $\cup_{m\geq n} \mathbf{W}_m$ is small, so $Y_n \to_p Y_o$ does not imply $Y_n \to_{as} Y_o$. For example, take the universe of states of nature to be the points on the unit circle with uniform probability, take the $\mathbf{W}_n$ to be successive arcs of length 2π/n, and take Y_n to be 1 on $\mathbf{W}_n$, 0 otherwise. Then $Y_n \to_p 0$ since $\Pr(Y_n \neq 0) = 1/n$, but Y_n fails to converge almost surely to zero since the successive arcs wrap around the circle an infinite number of times, and every s in the circle is in an infinite number of $\mathbf{W}_n$.

4.1.9. Suppose $Y_n \to_p Y_o$. It is a good exercise in manipulation of probabilities of events to show that $Y_n \to_d Y_o$. Given ε > 0, define $\mathbf{W}_n$ as before to be the set of states of Nature where $|Y_n(s) - Y_o(s)| > \varepsilon$. Given y, define $\mathbf{A}_n$, $\mathbf{B}_o$, and $\mathbf{C}_o$ to be, respectively, the states of Nature with $Y_n \leq y$, $Y_o \leq y - \varepsilon$, and $Y_o \le y + \varepsilon$. Then $\mathbf{B}_o \subseteq \mathbf{A}_n \cup \mathbf{W}_n$ (i.e., $Y_o(s) \le y - \varepsilon$ implies either $Y_n(s) \le y$ or $|Y_o(s) - Y_n(s)| > \varepsilon$) and $\mathbf{A}_n \subseteq \mathbf{C}_o \cup \mathbf{W}_n$ (i.e., $Y_n(s) \le y$ implies $Y_o(s) \le y + \varepsilon$ or $|Y_o(s) - Y_n(s)| > \varepsilon$). Hence, for n large enough so $\text{Prob}(\mathbf{W}_n) < \varepsilon$, $F_o(y-\varepsilon) \equiv \text{Prob}(\mathbf{B}_o) \le \text{Prob}(\mathbf{A}_n) + \text{Prob}(\mathbf{W}_n) < F_n(y) + \varepsilon$, and $F_n(y) \equiv \text{Prob}(\mathbf{A}_n) \le \text{Prob}(\mathbf{C}_o) + \text{Prob}(\mathbf{W}_n) < F_o(Y+\varepsilon) + \varepsilon$, implying $F_o(y-\varepsilon) - \varepsilon \le \lim_{n\to\infty} F_n(y) \le F_o(y+\varepsilon) + \varepsilon$. If y is a continuity point of Y_o, then $F_o(y-\varepsilon)$ and $F_o(y+\varepsilon)$ approach $F_o(y)$ as $\varepsilon \to 0$, implying $\lim_{n\to\infty} F_n(y) = F_o(y)$. This establishes that $Y_n \to_d Y_o$.

Convergence in distribution of Y_n to Y_o does not imply that Y_n and Y_o are close to each other. For example, if Y_n and Y_o are i.i.d. standard normal, then $Y_n \to_d Y_o$ trivially, but clearly not $Y_n \to_p Y_o$ since $Y_n - Y_o$ is normal with variance 2, and $|Y_n - Y_o| > \varepsilon$ with a positive, constant probability. However, there is a useful representation that is helpful in relating convergence in distribution and almost sure convergence; see P. Billingsley (1986), p.343.

Theorem 4.1. (Skorokhod) If $Y_n \to_d Y_o$, then there exist random variables Y_n' and Y_o' such that Y_n and Y_n' have the same CDF, as do Y_o and Y_o', and $Y_n' \to_{as} Y_o'$.

4.1.10. Convergence in distribution and convergence in probability to a <u>constant</u> are equivalent. If $Y_n \to_p c$ constant, then $Y_n \to_d c$ as a special case of 4.1.9 above. Conversely, $Y_n \to_d c$ constant means $F_n(y) \to F_o(y)$ at continuity points, where $F_c(y) = 0$ for $y < c$ and $F_c(y) = 1$ for $y \ge c$. Hence $\varepsilon > 0$ implies $\text{Prob}(|Y_n - c| > \varepsilon) = F_n(c-\varepsilon) + 1 - F_n(c+\varepsilon) \to 0$, so $Y_n \to_p c$. This result implies particularly that the statements $Y_n - Y_o \to_p 0$ and $Y_n - Y_o \to_d 0$ are equivalent. Then, $Y_n - Y_o \to_d 0$ implies $Y_n \to_d Y_o$, but the reverse implication does not hold.

4.1.11. The condition that convergence in distribution is equivalent to convergence of expectations of all bounded continuous functions is a fundamental mathematical result called the *Helly-Bray theorem*. Intuitively, the reason the theorem holds is that bounded continuous functions can be approximated closely by sums of continuous "almost-step" functions, and the expectations of "almost step" functions closely approximate points of CDF's. A proof by J. Davidson (1994), p. 352, employs the Skorokhod representation theorem 4.1.

4.1.12. A Chebyshev-like inequality is obtained by noting for a random variable Z with density f(z) that $\mathbf{E}|Z|^\rho = \int |z|^\rho f(z)dz \ge \int_{|z|\ge\varepsilon} \varepsilon^\rho f(z)dz = \varepsilon^\rho \text{Prob}(|Z| > \varepsilon)$, or $\text{Prob}(|Z| > \varepsilon) \le \mathbf{E}|Z|^\rho/\varepsilon^\rho$. (When $\rho = 2$, this is the conventional Chebyshev inequality. When $\rho = 1$, one has $\text{Prob}(|Z| > \varepsilon) \le \mathbf{E}|Z|/\varepsilon$.) Taking $Z = Y_n - Y_o$, one has $\lim_{n\to\infty} \text{Prob}(|Y_n - Y_o| > \varepsilon) \le \varepsilon^{-\rho}\cdot\lim_{n\to\infty} \mathbf{E}|Y_n - Y_o|^\rho$. Hence, convergence in ρ-mean (for any $\rho > 0$) implies convergence in probability. However, convergence almost surely or in probability does not necessarily imply convergence in ρ-mean. Suppose the sample space is the unit interval with uniform probability, and $Y_n(s) = e^{n'}$ for $s \le n^{-2}$, zero otherwise. Then $Y_n \to_{as} 0$ since $\text{Prob}(Y_m \ne 0 \text{ for any } m > n) \le n^{-2}$, but $\mathbf{E}|Y_n|^\rho = e^{\rho n}/n^2 \to +\infty$ for any $\rho > 0$.

FIGURE 4.2. RELATIONS BETWEEN STOCHASTIC LIMITS

(Section numbers for details are given in parentheses)

1	$Y_n \to_{as} Y_o \quad \overset{(1.8)}{\Longrightarrow} \quad Y_n \to_p Y_o \quad \overset{(1.9)}{\Longrightarrow} \quad Y_n \to_d Y_o$
	$\Uparrow\ (1.4)\ \Downarrow \qquad\qquad \Uparrow\ (1.3)\ \Downarrow \qquad\qquad \Uparrow\ (1.10)\ \not\Downarrow$
2	$Y_n - Y_o \to_{as} 0 \quad \underset{(1.8)}{\Longrightarrow} \quad Y_n - Y_o \to_p 0 \quad \underset{(1.10)}{\Longleftrightarrow} \quad Y_n - Y_o \to_d 0$
3	$Y_n \to_d$ c (a constant) $\Longleftrightarrow Y_n \to_p$ c (1.10)
4	$Y_n \to_d Y_o \Longleftrightarrow \mathbf{E}g(Y_n) \to \mathbf{E}g(Y_o)$ for all bounded continuous g (1.11)
5	$\|Y_n - Y_o\|_\rho \to 0$ for some $\rho > 0 \Longrightarrow Y_n \to_p Y_o$ (1.12)
6	$\|Y_n - Y_o\|_\rho \le M$ (all n) & $Y_n \to_p Y_o \Longrightarrow \|Y_n - Y_o\|_\lambda \to 0$ for $0 < \lambda < \rho$ (1.13)
7	$Y_n \to_p Y_o \Longrightarrow Y_{n_k} \to_{as} Y_o$ for some subsequence n_k, k = 1,2,... (1.14)
8	$\sum_{n=1}^{\infty} P(\vert Y_n - Y_o\vert > \varepsilon) < +\infty$ for each $\varepsilon > 0 \Longrightarrow Y_n \to_{as} Y_o$ (1.15)
9	$\sum_{n=1}^{\infty} \mathbf{E}\vert Y_n - Y_o\vert^\rho < +\infty$ (for some $\rho > 0$) $\Longrightarrow Y_n \to_{as} Y_o$ (1.15)
10	$Y_n \to_d Y_o$ & $Z_n - Y_n \to_p 0 \Longrightarrow Z_n \to_d Y_o$ (1.16)
11	$Y_n \to_p Y_o \Longrightarrow g(Y_n) \to_p g(Y_o)$ for all continuous g (1.17)
12	$Y_n \to_d Y_o \Longrightarrow g(Y_n) \to_d g(Y_o)$ for all continuous g (1.18)

4.1.13. Adding a condition of a uniformly bounded ρ-order mean $\mathbf{E}|Y_n|^\rho \le M$ to convergence in probability $Y_n \to_p Y_o$ yields the result that $\mathbf{E}|Y_o|^\lambda$ exists for $0 < \lambda \le \rho$, and $\mathbf{E}|Y_n|^\lambda \to \mathbf{E}|Y_o|^\lambda$ for $0 < \lambda < \rho$. This result can be restated as "the moments of the limit equal the limit of the moments" for moments of order λ less than ρ. Replacing Y_n by $Y_n - Y_o$ and Y_o by 0 gives the result in Figure 4.2.

To prove these results, we will find useful the property of moments that $\mathbf{E}|Y|^\lambda \le (\mathbf{E}|Y|^\rho)^{\lambda/\rho}$ for $0 < \lambda < \rho$. (This follows from Holder′s inequality (2.1.11), which states $\mathbf{E}|UV| \le (\mathbf{E}|U|^r)^{1/r}(\mathbf{E}|V|^s)^{1/s}$ for r,s > 0 and $r^{-1} + s^{-1} = 1$, by taking $U = |Y|^\lambda$, V = 1, and $r = \rho/\lambda$.) An immediate implication is $\mathbf{E}|Y_n|^\lambda \le M^{\lambda/\rho}$. Define $g(y,\lambda,k) = \min(|y|^\lambda,k^\lambda)$, and note that since it is continuous and bounded, the Healy-Bray theorem implies $\mathbf{E}g(Y_n,\lambda,k) \to \mathbf{E}g(Y_o,\lambda,k)$. Therefore,

$$M^{\lambda/\rho} \ge \mathbf{E}|Y_n|^\lambda \ge \mathbf{E}g(Y_n,\lambda,k) = \int_{-k}^{k} |y|^\lambda f_n(y)dy + k^\lambda\cdot \mathrm{Prob}(|Y_n| > k)$$

$$\to \int_{-k}^{k} |y|^\lambda f_o(y)dy + k^\lambda \mathrm{Prob}(|Y_o| > k).$$

Letting $k \to \infty$ establishes that $\mathbf{E}|Y_o|^\lambda$ exists for $0 < \lambda \le \rho$. Further, for $\lambda < \rho$,

$$0 \le \mathbf{E}|Y_n|^\lambda - \mathbf{E}g(Y_n,\lambda,k) \le \int_{|y|>k} |y|^\lambda f_n(y)dy \le k^{\lambda-\rho} \int_{|y|>k} |y|^\rho f_n(y)dy \le k^{\lambda-\rho}M.$$

Choose k sufficiently large so that $k^{\lambda-\rho}M < \varepsilon$. The same inequality holds for Y_o. Choose n sufficiently large so that $|\mathbf{E}g(Y_n,\lambda,k) \to \mathbf{E}g(Y_o,\lambda,k)| < \varepsilon$. Then

$$|\mathbf{E}|Y_n|^\lambda - \mathbf{E}|Y_o|^\lambda| \le |\mathbf{E}|Y_n|^\lambda - \mathbf{E}g(Y_n)| + |\mathbf{E}g(Y_n) - \mathbf{E}g(Y_o)| + |\mathbf{E}g(Y_o) - \mathbf{E}|Y_o|^\lambda| \le 3\varepsilon.$$

This proves that $\mathbf{E}|Y_n|^\lambda \to \mathbf{E}|Y_o|^\lambda$.

An example shows that $\mathbf{E}|Z_n|^\lambda \to 0$ for $\lambda < \rho$ does not imply $\mathbf{E}|Z_n|^\rho$ bounded. Take Z_n discrete with support {0,n} and probability log(n)/n at n. Then for $\lambda < 1$, $\mathbf{E}|Z_n|^\lambda = \log(n)/n^{1-\lambda} \to 0$, but $\mathbf{E}|Z_n|^1 = \log(n) \to +\infty$.

4.1.14. If $Y_n \to_p Y_o$, then $\mathrm{Prob}(\mathbf{W}_n) \to 0$. Choose a subsequence n_k such that $Prob(\boldsymbol{W}_{n_k}) \le 2^{-k}$.

Then $\mathrm{Prob}(\cup_{k'>k}\boldsymbol{W}_{n_{k'}}) \le \sum_{k'>k} Prob(\boldsymbol{W}_{n_{k'}}) \le \sum_{k'>k} 2^{-k'} = 2^{-k}$, implying $Y_{n_k} \to_{as} Y_o$.

4.1.15. Conditions for a.s. convergence follow from this basic probability theorem:

Theorem 4.2. (Borel-Cantelli) If $\mathbf{A}_i$ is any sequence of events in a probability space ($\mathbf{S}$,$\boldsymbol{F}$,P), $\sum_{n=1}^{\infty} P(\mathbf{A}_i) < +\infty$ implies that almost surely only a finite number of the events $\mathbf{A}_i$ occur. If $\mathbf{A}_i$ is a sequence of independent events, then $\sum_{n=1}^{\infty} P(\mathbf{A}_i) = +\infty$ implies that almost surely an infinite number of the events $\mathbf{A}_i$ occur.

Apply the Borel-Cantelli theorem to the events $\mathbf{A}_i = \{s\in\mathbf{S}|\ |Y_i - Y_o| > \varepsilon\}$ to conclude that $\sum_{n=1}^{\infty} P(\mathbf{A}_i) < +\infty$ implies that almost surely only a finite number of the events $\mathbf{A}_i$ occur, and hence $|Y_i - Y_o| \le \varepsilon$ for all I sufficiently large. Thus, $Y_n - Y_o \to_{as} 0$, or $Y_n \to_{as} Y_o$. For the next result in the table, use (1.12) to get $\text{Prob}(\cup_{m\ge n} \boldsymbol{W}_m) \le \sum_{m>n} \text{Prob}(\mathbf{W}_m) \le \varepsilon^{-\rho}\sum_{m>n} \mathbf{E}|Y_m - Y_o|^{\rho}$. Apply Theorem 4.2 to conclude that if this right-hand expression is finite, then $Y_n \to_{as} Y_o$. The example at the end of (1.12) shows that almost sure convergence does not imply convergence in ρ-mean. Also, the example mentioned in 1.8 which has convergence in probability but not almost sure convergence can be constructed to have ρ-mean convergence but not almost sure convergence.

4.1.16. A result which is very useful in applied work is that if two random variables Y_n and Z_n have a difference which converges in probability to zero, and if Y_n converges in distribution to Y_o, then $Z_n \to_d Y_o$ also. In this case, Y_n and Z_n are termed *asymptotically equivalent*. The argument demonstrating this result is similar to that for 4.1.9. Let F_n and G_n be the CDF's of Y_n and Z_n respectively. Let y be a continuity point of F_o and define the following events:

$$\mathbf{A}_n = \{s|Z_n(s) < y\},\ \mathbf{B}_n = \{s|Y_n(s) \le y - \varepsilon\},\ \mathbf{C}_n = \{s|Y_n(s) \le y + \varepsilon\},\ \mathbf{D}_n = \{s|\ |Y_n(s) - Z_n(s)| > \varepsilon\}.$$

Then $\mathbf{A}_n \subseteq \mathbf{C}_n\cup\mathbf{D}_n$ and $\mathbf{B}_n \subseteq \mathbf{A}_n\cup\mathbf{D}_n$, implying $F_n(y-\varepsilon) - \text{Prob}(\mathbf{D}_n) \le G_n(y) \le F_n(y+\varepsilon) + \text{Prob}(\mathbf{D}_n)$. Given $\delta > 0$, one can choose $\varepsilon > 0$ such that $y-\varepsilon$ and $y+\varepsilon$ are continuity points of F_n, and such that $F_o(y+\varepsilon) - F_o(y-\varepsilon) < \delta/3$. Then one can choose n sufficiently large so that $\text{Prob}(\mathbf{D}_n) < \delta/3$, $|F_n(y+\varepsilon) - F_o(y+\varepsilon)| < \delta/3$ and $|F_n(y+\varepsilon) - F_o(y+\varepsilon)| < \delta/3$. Then $|G_n(y) - F_o(y)| < \delta$.

4.1.17 A useful property of convergence in probability is the following result:

Theorem 4.3. If g(y) is a continuous function on an open set containing the support of Y_o, then $Y_n \to_p Y_o$ implies $g(Y_n) \to_p g(Y_o)$. The result also holds for vectors of random variables, and specializes to the rules that if $Y_{1n} \to_p Y_{10}$ and $Y_{2n} \to_p Y_{20}$, then (a) $Y_{1n} \cdot Y_{2n} \to_p Y_{10} \cdot Y_{20}$, (b) $Y_{1n} + Y_{2n} \to_p Y_{10} + Y_{20}$, and © If $\text{Prob}(|Y_{20}| < \varepsilon) = 0$ for some $\varepsilon > 0$, then $Y_{1n}/Y_{2n} \to_p Y_{10}/Y_{20}$. In these limits, Y_{10} and/or Y_{20} may be constants.

Proof: Given $\varepsilon > 0$, choose M such that $P(|Y_o| > M) < \varepsilon$. Let $\mathbf{A}_o$ be the set of y in the support of Y_o that satisfy $|y| \le M$. Then $\mathbf{A}_o$ is compact. Mathematical analysis can be used to show that there exists a nested sequence of sets $\mathbf{A}_o \subseteq \mathbf{A}_1 \subseteq \mathbf{A}_2 \subseteq \mathbf{A}_3$ with $\mathbf{A}_3$ an open neighborhood of $\mathbf{A}_o$ on which g is continuous, $\mathbf{A}_2$ compact, and $\mathbf{A}_1$ open. From 4.16, $\liminf_{n\to\infty} F_n(\mathbf{A}_1) \ge F_o(\mathbf{A}_1) \ge 1\text{-}\varepsilon$ implies there exists n_1 such that for $m > n_1$, $F_m(\mathbf{A}_1) \ge 1\text{-}2\varepsilon$. The continuity of g implies that for each $y \in \mathbf{A}_2$, there exists $\delta_y > 0$ such that $|y'\text{-}y| < \delta_y \Rightarrow |g(y') - g(y)| < \varepsilon$. These δ_y-neighborhoods cover $\mathbf{A}_2$. Then $\mathbf{A}_2$ has a finite subcover. Let δ be the smallest value of δ_y in this finite subcover. Then, g is uniformly continuous: $y \in \mathbf{A}_2$ and $|y'\text{-}y| < \delta$ imply $|g(y') - g(y)| < \varepsilon$. Choose $n > n_1$ such that for $m > n$, $P(|Y_m - Y_o| > \delta) < \varepsilon/2$. Then for $m > n$, $P(|g(Y_m) - g(Y_o)| > \varepsilon) \le P(|Y_n - Y_o| > \delta) + P(|Y_o| > M) + 1 - F_m(\mathbf{A}_1) \le 4\varepsilon$. □

4.1.18 The preceding result has an analog for convergence in distribution. This result establishes, for example, that if $Y_n \to_d Y_o$, with Y_o standard normal and $g(y) = y^2$, then Y_o is chi-squared, so that that Y_n^2 converges in distribution to a chi-squared random variable.

Theorem 4.4. If g(y) is a continuous function on an open set containing the support of Y_o, then $Y_n \to_d Y_o$ implies $g(Y_n) \to_d g(Y_o)$. The result also holds for vectors of random variables.

Proof: Construct the sets $\mathbf{A}_o \subseteq \mathbf{A}_1 \subseteq \mathbf{A}_2 \subseteq \mathbf{A}_3$ as in the proof of Theorem 4.3. A theorem from mathematical analysis (Urysohn) states that there exists a continuous function r with values between zero and one that satisfies $r(y) = 1$ for $y \in \mathbf{A}_1$ and $r(y) = 0$ for $y \notin \mathbf{A}_3$. Then $g^*(y) = g(y) \cdot r(y)$ is continuous everywhere. From the Healy-Bray theorem, $Y_n \to_d Y_o \Longleftrightarrow \mathbf{E}\, h(Y_n) \to \mathbf{E}\, h(Y_o)$ for all continuous bounded h $\Longrightarrow \mathbf{E}\, h(g^*(Y_n)) \to \mathbf{E}\, h(g^*(Y_o))$ for all continuous bounded h, since the composition of continuous bounded functions is continuous and bounded $\Longleftrightarrow g^*(Y_n) \to_d g^*(Y_o)$. But $P(g^*(Y_n) \ne g(Y_n)) \le P(Y_n \notin \mathbf{A}_1) \le 2\varepsilon$ for n sufficiently large, and $g^*(Y_o) = g(Y_o)$. Then, 4.1.16 and $g^*(Y_n) - g(Y_n) \to_p 0$ imply $g^*(Y_n) \to_d g^*(Y_o)$. □

4.1.19. Convergence properties are sometimes summarized in a notation called $O_p(\cdot)$ and $o_p(\cdot)$ which is very convenient for manipulation. (Sometimes too convenient; it is easy to get careless and make mistakes using this calculus.) The definition of $o_p(\cdot)$ is $Y_n \to_p Y_o \Rightarrow Y_n = Y_o + o_p(1)$, and more generally $n^{-\alpha}(Y_n - Y_o) \to_p 0 \Rightarrow Y_n - Y_o = o_p(n^{\alpha})$. Thus $o_p(\cdot)$ is a notation for convergence in probability to zero of a suitably normalized sequence of random variables. When two sequences of random variables Y_n and Z_n are asymptotically equivalent, or $Y_n - Z_n = o_p(1)$, so that they have a common limiting distribution, this is sometime denoted $Y_n \sim_a Z_n$.

The notation $Y_n = O_p(1)$ is defined to mean that given $\varepsilon > 0$, there exists a large M (not depending on n) such that $\text{Prob}(|Y_n| > M) < \varepsilon$ for all n. A sequence with this property is called *stochastically bounded*. More generally, $Y_n = O_p(n^\alpha)$ means $\text{Prob}(|Y_n| > M \cdot n^\alpha) < \varepsilon$ for all n. An abbreviated list of rules for o_p and O_p is given in Figure 4.3.

A sequence that is convergent in distribution is stochastically bounded: If $Y_n \to_d Y_o$, then one can find M and n_o such that $\pm$ M are continuity points of Y_o, $\text{Prob}(|Y_o| \le M) > 1-\varepsilon/2$, $|F_n(M) - F_o(M)| < \varepsilon/4$ and $|F_n(-M) - F_o(-M)| < \varepsilon/4$ for $n > n_o$. Then $\text{Prob}(|Y_n| > M) < \varepsilon$ for $n > n_o$. This implies $Y_n = O_p(1)$. On the other hand, one can have $Y_n = 0_p(1)$ without having convergence to any distribution (e.g., consider $Y_n \equiv 0$ for n odd and Y_n standard normal for n even). The notation $Y_n = O_p(n^\alpha)$ means $n^{-\alpha}Y_n = O_p(1)$.

FIGURE 4.3. RULES FOR $O_p(\cdot)$ AND $o_p(\cdot)$

Definition: $Y_n = o_p(n^\alpha) \Longrightarrow \text{Prob}(|n^{-\alpha}Y_n| > \varepsilon) \to 0$ for each $\varepsilon > 0$.
Definition: $Y_n = O_p(n^\alpha) \Longrightarrow$ for each $\varepsilon > 0$, there exists $M > 0$ such that $\text{Prob}(|n^{-\alpha}Y_n| > M) < \varepsilon$ for all n

1	$Y_n = o_p(n^\alpha) \Longrightarrow Y_n = O_p(n^\alpha)$
2	$Y_n = o_p(n^\alpha)$ & $\beta > \alpha \Longrightarrow Y_n = o_p(n^\beta)$
3	$Y_n = O_p(n^\alpha)$ & $\beta > \alpha \Longrightarrow Y_n = o_p(n^\beta)$
4	$Y_n = o_p(n^\alpha)$ & $Z_n = o_p(n^\beta) \Longrightarrow Y_n \cdot Z_n = o_p(n^{\alpha+\beta})$
5	$Y_n = O_p(n^\alpha)$ & $Z_n = O_p(n^\beta) \Longrightarrow Y_n \cdot Z_n = O_p(n^{\alpha+\beta})$
6	$Y_n = O_p(n^\alpha)$ & $Z_n = o_p(n^\beta) \Longrightarrow Y_n \cdot Z_n = o_p(n^{\alpha+\beta})$
7	$Y_n = o_o(n^\alpha)$ & $Z_n = o_p(n^\beta)$ & $\beta \ge \alpha \Longrightarrow Y_n + Z_n = o_p(n^\beta)$
8	$Y_n = O_p(n^\alpha)$ & $Z_n = O_p(n^\beta)$ & $\beta \ge \alpha \Longrightarrow Y_n + Z_n = O_p(n^\beta)$
9	$Y_n = O_p(n^\alpha)$ & $Z_n = o_p(n^\beta)$ & $\beta > \alpha \Longrightarrow Y_n + Z_n = o_p(n^\beta)$
10	$Y_n = O_p(n^\alpha)$ & $Z_n = o_p(n^\beta)$ & $\beta < \alpha \Longrightarrow Y_n + Z_n = O_p(n^\alpha)$
11	$Y_n = O_p(n^\alpha)$ & $Z_n = o_p(n^\alpha) \Longrightarrow Y_n + Z_n = O_p(n^\alpha)$

We prove the very useful rule 6 in Figure 4.3: Given $\varepsilon > 0$, $Y_n = O_p(n^\alpha) \Longrightarrow \exists\ M > 0$ such that $\text{Prob}(|n^{-\alpha}Y_n| > M) < \varepsilon/2$. Next $Z_n = o_p(n^\beta)$ implies $\exists\ n_o$ such that for $n > n_o$, $\text{Prob}(|n^{-\beta}Z_n| > \varepsilon/M) < \varepsilon/2$. Hence $\text{Prob}(|n^{-\alpha-\beta}Y_nZ_n| > \varepsilon) \le \text{Prob}(|n^{-\alpha}Y_n| > M) + \text{Prob}(|n^{-\beta}Z_n| > \varepsilon/M) < \varepsilon$. Demonstration of the remaining rules is left as an exercise.

4.2. INDEPENDENT AND DEPENDENT RANDOM SEQUENCES

4.2.1. Consider a sequence of random variables $Y_1,Y_2,Y_3,...$. The *joint distribution* (CDF) of a finite subsequence $(Y_1,...,Y_n)$, denoted $F_{1,...,n}(y_1,...,y_n)$, is defined as the probability of a state of Nature such that all of the inequalities $Y_1 \le y_1,...,Y_n \le y_n$ hold. The random variables in the sequence are *mutually statistically independent* if for every finite subsequence $Y_1...,Y_n$, the joint CDF factors:

$$F_{1,...,n}(y_1,...,y_n) \equiv F_1(y_1)\cdot...\cdot F_n(y_n).$$

The variables are *independent and identically distributed* (i.i.d.) if in addition they have a common univariate CDF $F_1(y)$. The case of i.i.d. random variables leads to the simplest theory of stochastic limits, and provides the foundation needed for much of basic econometrics. However, there are many applications, particularly in analysis of economic time series, where i.i.d. assumptions are not plausible, and a limit theory is needed for dependent random variables. We will define two types of dependence, martingale and mixing, that will cover a variety of econometric time series applications and require a modest number of tools from probability theory. We have introduced a few of the needed tools in Chapter 3, notably the idea of information contained in σ-fields of events, with the evolution of information captured by refinements of these σ-fields, and the definitions of measurable functions, product σ-fields, and compatability conditions for probabilities defined on product spaces. There are treatments of more general forms of dependence than martingale or mixing, but these require a more comprehensive development of the theory of stochastic processes.

4.2.2. Consider a sequence of random variables Y_k with k interpreted as an index of (discrete) time. One can think of k as the infinite sequence $k \in \mathbf{K} = \{1,2,...\}$, or as a doubly infinite sequence, extending back in time as well as forward, $k \in \mathbf{K} = \{...,-2,-1,0,1,2,...\}$. The set of states of Nature can be defined as the product space $\mathbf{S} = \times_{i\in\mathbf{K}} \mathbb{R}$, or $\mathbf{S} = \mathbb{R}^{\mathbf{K}}$, where $\mathbb{R}$ is the real line, and the "complete information" σ-field of subsets of **S** defined as $\boldsymbol{F}_{\mathbf{K}} = \otimes_{i\in\mathbf{K}} \boldsymbol{B}$, where $\boldsymbol{B}$ is the Borel σ-field of subsets of the real line; see 3.2. (The same apparatus, with **K** equal to the real line, can be used to consider continuous time. To avoid a variety of mathematical technicalities, we will not consider the continuous time case here.) 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 $\boldsymbol{G}_t = (\otimes_{i\le t} \boldsymbol{B})\otimes(\otimes_{i>t}\{\varphi,\mathbf{S}\})$ 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 sequence of random variables Y_t 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 ≤ t. The sequence $...,Y_{-1},Y_0,Y_1,Y_2,...$ adapted to $\boldsymbol{G}_t$ for k ∈ **K** is termed a *stochastic process*. One way of thinking of a stochastic process is to recall that random variables are functions of states of Nature, so that the process is a function $Y:S\times K \to \mathbb{R}$. Then Y(s,k) is the *realization* of the random variable in period k, Y(s,·) a realization or *time-path* of the stochastic process, and Y(·,k) the random variable in period k. Note that there may be more than one sequence of σ-fields in operation for a particular 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*.

Each subsequence $(Y_m,...,Y_{m+n})$ of the stochastic process has a multivariate CDF $F_{m,..,m+n}(y_m,...,y_{m+n})$. It is said to be *stationary* if for each n, this CDF is the same for every m. A stationary process has the obvious property that moments such as means, variances, and covariances between random variables a fixed number of time periods apart are the same for all times m. Referring to 4.2.1, a sequence i.i.d. random variables is always stationary.

4.2.3. 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.

Theorem 4.5. 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 $P(\max_{1\le k\le n} | \sum_{i=1}^{k} Y_i| > \varepsilon) \le \sum_{i=1}^{n} \sigma_i^2/\varepsilon^2$.

Proof: Let $S_k = \sum_{i=1}^{k} Y_i$. Let Z_k be a random variable that is one if $S_j \le \varepsilon$ for $j < k$ and $S_k > \varepsilon$, zero otherwise. Note that $\sum_{i=1}^{n} Z_i \le 1$ and $\mathbf{E}(\sum_{i=1}^{n} Z_i) = P(\max_{1\le k\le n} |\sum_{i=1}^{k} Y_i| > \varepsilon)$. The variables S_k and Z_k depend only on Y_i for $i \le k$. Then $E(S_n - S_k|S_k,Z_k) = 0$. Hence

$$\mathbf{E}s_n^2 \ge \sum_{k=1}^{n} \mathbf{E}S_n^2\cdot Z_k = \sum_{k=1}^{n} \mathbf{E}[S_k + (S_n - S_k)]^2\cdot Z_k \ge \sum_{k=1}^{n} \mathbf{E}S_k^2\cdot Z_k \ge \varepsilon^2 \sum_{k=1}^{n} \mathbf{E}Z_k. \ \square$$

4.2.4. As a practical matter, 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 avalance of new information (*shocks*). To formalize this, we introduce the concept of *stochastic mixing*. For a stochastic process Y_t, consider events $\mathbf{A} \in \boldsymbol{G}_t$ and $\mathbf{B} \in \boldsymbol{H}_{t+s}$; then **A** draws only on information up through period t and **B** draws only on information 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**, and 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 $|\mathbf{E}(U-\mathbf{E}U)(W-\mathbf{E}W)| \le \rho_t[(\mathbf{E}(U-\mathbf{E}U)^2)(\mathbf{E}(W-\mathbf{E}W)^2)]^{1/2}$ 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 imdependent standard normal. Then, $\mathrm{var}(Y_t) = 1/(1-\rho^2)$ and $\mathrm{cov}(Y_{t+s},Y_t) = \rho^s/(1-\rho^2)$, 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. We return to this subject in the chapter on time series analysis.

4.3. LAWS OF LARGE NUMBERS

4.3.1. Consider a sequence of random variables $Y_1,Y_2,...$ and a corresponding sequence of averages $X_n = n^{-1}\sum_{i=1}^{n} Y_i$ for n = 1,2,... . *Laws of large numbers* give conditions under which the averages X_n converge to a constant, either in probability (weak laws, or WLLN) or almost surely (strong laws, or SLLN). Laws of large numbers give formal content to the intuition that sample averages are accurate analogs of population averages when the samples are large, and are essential to establishing that statistical estimators for many problems have the sensible property that with sufficient data they are likely to be close to the population values they are trying to estimate. In econometrics, convergence in probability provided by a WLLN suffices for most purposes. However, the stronger result of almost sure convergence is occasionally useful, and is often attainable without additional assumptions.

4.3.2 Figure 4.4 lists a sequence of laws of large numbers. The case of independent identically distributed (i.i.d.) random variables yields the strongest result (Kolmogorov I). With additional conditions it is possible to get a laws of large numbers even for correlated variable provided the correlations of distant random variables approach zero sufficiently rapidly.

To show why WLLN work, I outline proofs of the first three laws in Figure 4.4.

Theorem 4.6. (Khinchine) If the Y_k are i.i.d., and $\mathbf{E}\, Y_k = \mu$, then $X_n \to_p \mu$.

Proof: The argument shows that the characteristic function (c.f.) of X_n converges pointwise to the c.f. for a constant random variable μ. Let $\psi(t)$ be the c.f. of Y_1. Then X_n has c.f. $\psi(t/n)^n$. Since $\mathbf{E}Y_1$ exists, ψ has a Taylor's expansion $\psi(t) = 1 + \psi'(\lambda t)t$, where $0 < \lambda < 1$ (see 3.5.12). Then $\psi(t/n)^n = [1 + (t/n)\,\psi'(\lambda t/n)]^n$. But $\psi'(\lambda t/n) \to \psi'(0) = \iota\mu$. A result from 2.1.10 states that if a sequence of scalars α_n has a limit, then $[1+\alpha_n/n]^n \to \exp(\lim \alpha_n)$. Then $\psi(t/n)^n \to e^{\iota\mu t}$. But this is the c.f. of a constant random variable μ, implying $X_n \to_d \mu$, and hence $X_n \to_p \mu$. □

FIGURE 4.4. LAWS OF LARGE NUMBERS FOR $X_n = n^{-1}\sum_{k=1}^{n} Y_k$

WEAK LAWS (WLLN)

1 (Khinchine) If the Y_k are i.i.d., and $\mathbf{E}\, Y_k = \mu$, then $X_n \to_p \mu$

2 (Chebyshev) If the Y_k are uncorrelated with $\mathbf{E}\, Y_k = \mu$ and $\mathbf{E}(Y_k - \mu)^2 = \sigma_k^2$ satisfying

$$\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty, \text{ then } X_n \to_p \mu$$

3 If the Y_k have $\mathbf{E}\, Y_k = \mu$, $\mathbf{E}(Y_k-\mu)^2 \equiv \sigma_k^2$, and $|\mathbf{E}(Y_k-\mu)(Y_m-\mu)| \le \rho_{km}\sigma_k\sigma_m$ with

$$\sum_{k=1}^{\infty} \sigma_k^2/k^{3/2} < +\infty \text{ and } \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\sum_{m=1}^{n} \rho_{km} < +\infty, \text{ then } X_n \to_p \mu$$

STRONG LAWS (SLLN)

1 (Kolmogorov I) If the Y_k are i.i.d., and $\mathbf{E}\, Y_k = \mu$, then $X_n \to_{as} \mu$

2 (Kolmogorov II) If the Y_k are independent, with $\mathbf{E}\, Y_k = \mu$, and $\mathbf{E}(Y_k-\mu)^2 = \sigma_k^2$ satisfying

$$\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty, \text{ then } X_n \to_{as} \mu$$

3 (Martingale) Y_k adapted to $\sigma(\ldots,Y_{k-1},Y_k)$ 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 $X_n \to_{as} 0$

4 (Serfling) If the Y_k have $\mathbf{E}\, Y_k = \mu$, $\mathbf{E}(Y_k-\mu)^2 = \sigma_k^2$, and $|\mathbf{E}(Y_k-\mu)(Y_m-\mu)| \le \rho_{|k-m|}\sigma_k\sigma_m$,

with $\sum_{k=1}^{\infty} (\log k)^2\sigma_k^2/k^2 < +\infty$ and $\sum_{k=1}^{\infty} \rho_{|k-m|} < +\infty$, then $X_n \to_{as} \mu$

.

Theorem 4.7. (Chebyshev) If the Y_k are uncorrelated with $\mathbf{E}\, Y_k = \mu$ and $\mathbf{E}(Y_k - \mu)^2 = \sigma_k^2$ satisfying $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$, then $X_n \to_p \mu$.

Proof: One has $\mathbf{E}(X_n-\mu)^2 = \sum_{k=1}^{n} \sigma_n^2/n^2$. Kronecker's Lemma (see 2.1.9) establishes that $\sum_{k=1}^{\infty} \sigma_k^2/k^2$ bounded implies $\mathbf{E}(X_n-\mu)^2 \to 0$. Then Chebyshev's inequality implies $X_n \to_p \mu$. □

The condition $\sum_{k=1}^{\infty} \sigma_k^2/k^2$ bounded in Theorem 4.7 is obviously satisfied if σ_k^2 is uniformly bounded, but is also satisfied if σ_k^2 grows modestly with k; e.g., it is sufficient to have $\sigma_k^2(\log K)/k$ bounded.

Theorem 4.8. (WLLN 3) If the Y_k have $\mathbf{E}\, Y_k = \mu$, $\mathbf{E}(Y_k-\mu)^2 \equiv \sigma_k^2$, and $|\mathbf{E}(Y_k-\mu)(Y_m-\mu)| \le \rho_{km}\sigma_k\sigma_m$ with $\sum_{k=1}^{\infty} \sigma_k^2/k^{3/2} < +\infty$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \sum_{m=1}^{n} \rho_{km} < +\infty$, then $X_n \to_p \mu$

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

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

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

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

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

The conditions for this result are obviously met if the σ_k^2 are uniformly bounded and the correlation coefficients decline at a sufficient rate with the distance between observations; examples are geometric decline with ρ_{km} bounded by a multiple of $\lambda^{|k-m|}$ for some $\lambda < 1$ and an arithmetic decline with ρ_{km} bounded by a multiple of $|k-m|^{-1}$.

The Kolmogorov SLLN 1 is a better result than the Kinchine WLLN, yielding a stronger conclusion from the same assumptions. Similarly, the Kolmogorov SLLN 2 is a better result than the Chebyshev WLLN. Proofs of these theorems can be found in C. R. Rao (1973), p. 114-115. The Serfling SLLN 4 is broadly comparable to WLLN 3, but Serfling gets the stronger almost sure conclusion with somewhat stronger assumptions on the correlations and somewhat weaker assumptions on the variances. If variances are uniformly bounded and correlation coefficients decline at least at a rate inversely proportional to the square of the time difference, this sufficient for either the WLLN 3 or SLLN 4 assumptions.

The SLLN 3 in the table applies to martingale difference sequences, and shows that Kolmogorov II actually holds for m.d. sequences.

Theorem 4.9. If Y_t adapted to $\sigma(\ldots,Y_{k-1},Y_k)$ is a martingale difference sequence with $\mathbf{E}Y_t^2 = \sigma_t^2$ and $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < +\infty$, then $X_n \to_{as} 0$.

Proof: The theorem is stated and proved by J. Davidson (1994), p. 314. To give an idea why SLLN work, I will 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 4.5) 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 $\sum_{m=1}^{\infty} P(\sup_k |X_k| > \delta) \le 36 \sum_{i=1}^{\infty} \sigma_i^2/i^{3/2}\delta^2 < +\infty$. Theorem 4.2 gives the result. □

4.4. CENTRAL LIMIT THEOREMS

4.4.1. Consider a sequence of random variables $Y_1,...,Y_n$ with zero means, and the associated sequence of scaled averages $Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i$. Central limit theorems (CLT) are concerned with conditions under which the Z_n, or variants with more generalized scaling, converge in distribution to a normal random variable Z_o. I will present several basic CLT, prove the simplest, and discuss the remainder. These results are summarized in Figure 4.5 near the end of this Section.

The most straighforward CLT is obtained for *independent and identically distributed* (i.i.d.) random variables, and requires only that the random variables have a finite variance. Note that the finite variance assumption is an additional condition needed for the CLT that was not needed for the SLLN for i.i.d. variables.

Theorem 4.10. (Lindeberg-Levy) If random variables Y_k are i.i.d. with mean zero and finite positive variance σ^2, then $Z_n \to_d Z_o \sim N(0,\sigma^2)$.

Proof: The approach is to show that the characteristic function of Z_n converges for each argument to the characteristic function of a normal. The CLT then follows from the limit properties of characteristic functions (see 3.5.12). Let $\psi(t)$ be the cf of Y_1. Then Z_n has cf $\psi(t\cdot n^{-1/2})^n$. Since $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^2 = \sigma^2$, $\psi(t)$ has a Taylor′s expansion $\psi(t) = [1 + \psi''(\lambda t)t^2/2]$, where $0 < \lambda < 1$ and ψ'' is continuous with $\psi''(0) = -\sigma^2$. Then $\psi(t\cdot n^{-1/2})^n = [1 + \psi''(\lambda t\cdot n^{-1/2})t^2/2n]^n$. Then the limit result 2.1.10 gives $\lim_{n\to\infty} [1 + \psi''(\lambda t\cdot n^{-1/2})t^2/2n]^n = \exp(-\sigma^2 t^2/2)$. Thus, the cf of Z_n converges for each t to the cf of $Z_o \sim N(0,\sigma^2)$. □

4.4.2. When the variables are independent but not identically distributed, an additional bound on the behavior of tails of the distributions of the random variables, called the *Lindeberg condition*, is needed. 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. The main result, stated next, allows more general scaling than by $n^{-1/2}$.

Theorem 4.11. (Lindeberg-Feller) Suppose random variables Y_k are independent with mean zero and positive finite variances σ_k^2. Define $c_n^2 = \sum_{k=1}^{n} \sigma_k^2$ and $U_n = \sum_{k=1}^{n} Y_k/c_n$. Then $c_n^2 \to \infty$, $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, and $U_n \to_d U_o \sim N(0,1)$ if and only if the Y_k satisfy the Lindeberg condition that for $\varepsilon > 0$, $\lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{E}\, Y_k^2\cdot\mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 = 0$.

FIGURE 5. CENTRAL LIMIT THEOREMS FOR $\mathbf{Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i}$

1	(Lindeberg-Levy) Y_k i.i.d., $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma^2$ positive and finite $\Longrightarrow Z_n \to_d Z_o \sim N(0,\sigma^2)$
2	(Lindeberg-Feller) If Y_k independent, $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma_k^2 \in (0,+\infty)$, $c_n^2 = \sum_{k=1}^n \sigma_k^2$, then $c_n^2 \to +\infty$, $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, and $U_n = \sum_{k=1}^n Y_k/c_n \to_d U_o \sim N(0,1)$ if and only if the *Lindeberg condition* holds: for each $\varepsilon > 0$, $\sum_{k=1}^n \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(\lvert Y_k\rvert > \varepsilon c_n)/c_n^2 \to 0$
3	If Y_k independent, $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma_k^2 \in (0,+\infty)$, $c_n^2 = \sum_{k=1}^n \sigma_k^2$ have $c_n^2 \to +\infty$ and $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, then each of the following conditions is sufficient for the Lindeberg condition: (i) For some $r > 2$, $\sum_{k=1}^n \mathbf{E}\,\lvert Y_k\rvert^r/c_n^r \to 0$. (ii) (Liapunov) For some $r > 2$, $\mathbf{E}\,\lvert Y_k/\sigma_k\rvert^r$ is bounded uniformly for all n. (iii) For some $r > 2$, $\mathbf{E}\,\lvert Y_k\rvert^r$ is bounded, and c_k^2/k is bounded positive, uniformly for all k.
4	Y_k a martingale difference sequence adapted to $\sigma(\ldots,Y_{k-1},Y_k)$ with $\lvert Y_k\rvert < M$ for all k and $\mathbf{E}y_k^2 = \sigma_k^2$ satisfying $n^{-1}\sum_{k=1}^n \sigma_k^2 \to \sigma_o^2 > 0 \Longrightarrow Z_n \to_d Z_o \sim N(0,\sigma_o^2)$
5	(Ibragimov-Linnik) Y_k stationary and strong mixing with $\mathbf{E}\, Y_k = 0$, $\mathbf{E}\, Y_k^2 = \sigma^2 \in (0,+\infty)$, $\mathbf{E}Y_{k+s}Y_k = \sigma^2\rho_s$, and for some $r > 2$, $\mathbf{E}\lvert Y_n\rvert^r < +\infty$ and $\sum_{k=1}^\infty \alpha(k)^{1-2/r} < +\infty \Longrightarrow$ $\sum_{s=1}^\infty \lvert\rho_s\rvert < +\infty$ and $Z_n \to_d Z_o \sim N(0,\sigma^2(1+2\sum_{s=1}^\infty \rho_s))$

A proof of Theorem 4.11 can be found, for example, in P. Billingsley (1986), p. 369-375. It involves an analysis of the characteristic functions, with detailed analysis of the remainder terms in their Taylor's expansion. To understand the theorem, it is useful to first specialize it to the case that the σ_k^2 are all the same. Then $c_n^2 = n\sigma_1^2$, the conditions $c_n^2 \to \infty$ and $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$ are met automatically, and in the terminology at the start of this section, $U_n = Z_n/\sigma_1$. The theorem then says $U_n \to_d U_o \sim N(0,1)$ if and only if the sample average of $\mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon n^{1/2})$ converges to zero for each $\varepsilon > 0$. The last condition limits the possibility that the deviations in a single random variable could be as large in magnitude as the sum, so that the shape of the distribution of this variable makes a significant contribution to the shape of the distribution of the sum. An example shows how the Lindeberg condition bites. Consider independent random variables Y_k that equal $\pm k^r$ with probability $1/2k^{2r}$, and zero otherwise, where r is a positive scalar. The Y_k have mean zero and variance one, and $\mathbf{1}(|Y_k| > \varepsilon n^{1/2}) = 1$ if $k^r > \varepsilon n^{1/2}$, implying $n^{-1}\sum_{i=1}^{n} \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon n^{1/2}) = \max(0, 1-\varepsilon^{1/r} n^{(1-2r)/2r})$. This converges to zero, so the Lindeberg condition is satisfied iff $r < 1/2$. Thus, the tails of the sequence of random variables cannot "fatten" too rapidly.

The Lindeberg condition allows the variances of the Y_k to vary within limits. For example, the variables $Y_k = \pm 2^k$ with probability 1/2 have σ_n/c_n bounded positive, so that the variances grow too rapidly and the condition fails. The variables $Y_k = \pm 2^{-k}$ with probability 1/2 have c_n bounded, so that σ_1/c_n is bounded positive, the variances shrink too rapidly, and the condition fails. The next result gives some easily checked sufficient conditions for the Lindeberg condition.

Theorem 4.12. Suppose random variables Y_k are independent with mean zero and positive finite variances σ_k^2 that satisfy $c_n^2 = \sum_{k=1}^{n} \sigma_k^2 \to \infty$ and $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$. Then, each of the following conditions is sufficient for the Lindeberg condition to hold:

(i) For some $r > 2$, $\sum_{k=1}^{n} \mathbf{E}\, |Y_k|^r/c_n^r \to 0$.

(ii) (Liapunov) For some $r > 2$, $\mathbf{E}\, |Y_k/\sigma_k|^r$ is bounded uniformly for all n.
(iii) For some $r > 2$, $\mathbf{E}\, |Y_k|^r$ is bounded, and c_k^2/k is bounded positive, uniformly for all k.

Proof: To show that (i) implies the Lindeberg condition, write

$$\sum_{k=1}^{n} \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 \le (\varepsilon c_n)^{2-r} \sum_{k=1}^{n} \mathbf{E}\, |Y_k|^r \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^2 \le \varepsilon^{2-r} \sum_{k=1}^{n} \mathbf{E}\, |Y_k/c_n|^r.$$

For (ii), the middle expression in the string of inequalities above satisfies

$$(\varepsilon c_n)^{2-r} \sum_{k=1}^{n} \mathbf{E}\,|Y_k|^r \cdot \mathbf{1}(|Y_k| > \varepsilon c_n)/c_n^{\,2} \le \varepsilon^{2-r}(\max_{k\le n} \mathbf{E}|Y_k/\sigma_k|^r)\cdot \sum_{k=1}^{n} \sigma_k^{\,r}/c_n^{\,r}$$

$$\le \varepsilon^{2-r}(\max_{k\le n} \mathbf{E}|Y_k/\sigma_k|^r)\cdot \sum_{k=1}^{n} (\sigma_k^{\,2}/c_n^{\,2})\cdot(\max_{k\le n} (\sigma_k/c_n)^{r-2}),$$

and the assumptions ensure that $\max_{k\le n} \mathbf{E}|Y_k/\sigma_k|^r$ is bounded and $\max_{k\le n} (\sigma_k/c_n)^{r-2} \to 0$.

Finally, if (iii), then continuing the first string of inequalities,

$$\sum_{i=1}^{n} \mathbf{E}\,|Y_k|^r/c_n^{\,r} \le c_n^{\,2-r} n\cdot(\sup_k \mathbf{E}\,|Y_k|^r)/n\cdot(\inf_n c_n^{\,2}/n),$$

and the right-hand-side is proportional to $c_n^{\,2-r}$, which goes to zero. □

4.4.3. The following theorem establishes a CLT for the scaled sum $Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i$ of martingale differences; or $Z_n = n^{-1/2}(X_n - X_o)$. 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. Martingale differences can display dependence that corresponds to important economic applications, such as conditional variances that depend systematically on history.

Theorem 4.13. Suppose Y_k is a martingale difference adapted to $\sigma(\ldots,Y_{k-1},Y_k)$, and Y_k satisfies a uniform bound $|Y_k| < M$. Let $\mathbf{E}Y_k^{\,2} = \sigma_k^{\,2}$, and assume that $n^{-1}\sum_{k=1}^{n} \sigma_k^{\,2} \to \sigma_o^{\,2} > 0$. Then $Z_n \to_d Z_o \sim N(0,\sigma_o^{\,2})$.

4.4.4. Intuitively, the CLT results that hold for independent or martingale difference random variables should continue to hold if the degree of dependence between variables is 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 like the one given in the last example.

Theorem 4.14. (Ibragimov-Linnik) Suppose Y_k is stationary and strong mixing with mean zero, 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 $Z_n \to_d Z_o \sim N(0,\sigma^2(1+2\sum_{s=1}^{\infty} \rho_s))$.

The figure below summarizes the CLT's stated in this section.

FIGURE 4.5. CENTRAL LIMIT THEOREMS FOR $\mathbf{Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i}$

1	(Lindeberg-Levy) Y_k i.i.d., $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma^2$ positive and finite $\Longrightarrow Z_n \to_d Z_o \sim N(0,\sigma^2)$
2	(Lindeberg-Feller) If Y_k independent, $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma_k^2 \in (0,+\infty)$, $c_n^2 = \sum_{k=1}^{n} \sigma_k^2$, then $c_n^2 \to +\infty$, $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, and $U_n = \sum_{k=1}^{n} Y_k/c_n \to_d U_o \sim N(0,1)$ if and only if the *Lindeberg condition* holds: for each $\varepsilon > 0$, $\sum_{k=1}^{n} \mathbf{E}\, Y_k^2 \cdot \mathbf{1}(\lvert Y_k\rvert > \varepsilon c_n)/c_n^2 \to 0$
3	If Y_k independent, $\mathbf{E}Y_k = 0$, $\mathbf{E}Y_k^2 = \sigma_k^2 \in (0,+\infty)$, $c_n^2 = \sum_{k=1}^{n} \sigma_k^2$ have $c_n^2 \to +\infty$ and $\lim_{n\to\infty} \max_{1\le k\le n} \sigma_k/c_n = 0$, then each of the following conditions is sufficient for the Lindeberg condition: (i) For some $r > 2$, $\sum_{k=1}^{n} \mathbf{E}\, \lvert Y_k\rvert^r/c_n^r \to 0$. (ii) (Liapunov) For some $r > 2$, $\mathbf{E}\, \lvert Y_k/\sigma_k\rvert^r$ is bounded uniformly for all n. (iii) For some $r > 2$, $\mathbf{E}\, \lvert Y_k\rvert^r$ is bounded, and c_k^2/k is bounded positive, uniformly for all k.
4	Y_k a martingale difference sequence adapted to $\sigma(\ldots,Y_{k-1},Y_k)$ with $\lvert Y_k\rvert < M$ for all t and $\mathbf{E}Y_k^2 = \sigma_k^2$ satisfying $n^{-1}\sum_{k=1}^{n} \sigma_k^2 \to \sigma_o^2 > 0 \Longrightarrow Z_n \to_d Z_o \sim N(0,\sigma_o^2)$
5	(Ibragimov-Linnik) Y_k stationary and strong mixing with $\mathbf{E}\, Y_k = 0$, $\mathbf{E}\, Y_k^2 = \sigma^2 \in (0,+\infty)$, $\mathbf{E}Y_{k+s}Y_k = \sigma^2\rho_s$, and for some $r > 2$, $\mathbf{E}\lvert Y_n\rvert^r < +\infty$ and $\sum_{k=1}^{\infty} \alpha(k)^{1-2/r} < +\infty \Longrightarrow \sum_{s=1}^{\infty} \lvert\rho_s\rvert < +\infty$ and $Z_n \to_d Z_o \sim N(0,\sigma^2(1+2\sum_{s=1}^{\infty} \rho_s))$

4.5. EXTENSIONS OF LIMIT THEOREMS

4.5.1. Limit theorems can be extended in several directions: (1) obtaining results for "triangular arrays" that include weighted sums of random variables, (2) sharpening the rate of convergence to the limit for "well-behaved" random variables, and (3) establishing "uniform" laws that apply to random functions. In addition, there are a variety of alternatives to the cases given above where independence assumptions are relaxed. The first extension gives limit theorems for random variables weighted by other (non-random) variables, a situation that occurs often in econometrics. The second extension provides tools that allow us to bound the probability of large deviations of random sums. This is of direct interest as a sharper version of a Chebychev-type inequality, and also useful in obtaining further results. To introduce uniform laws, first define a *random function* (or *stochastic process*) $y = Y(\theta,s)$ that maps a state of Nature s and a real variable (or vector of variables) θ into the real line. This may also be written, suppressing the dependence on s, as $Y(\theta)$. Note that $Y(\cdot,w)$ is a *realization* of the random function, and is itself an ordinary non-random function of θ. For each value of θ, $Y(\theta,\cdot)$ is an ordinary random variable. A uniform law is one that bounds sums of random functions uniformly for all arguments θ. For example, a uniform WLLN would say $\lim_{n\to\infty} P(\sup_\theta |\, n^{-1}\sum_{i=1}^n Y_i(\theta,\cdot)| > \varepsilon) = 0$. Uniform laws play an important role in establishing the properties of statistical estimators that are nonlinear functions of the data, such as maximum likelihood estimates.

4.5.2 Consider a doubly indexed array of constants a_{in} defined for $1 \le i \le n$ and $n = 1,2,...$, and weighted sums of the form $X_n = \sum_{i=1}^n a_{in}Y_i$. If the Y_i are i.i.d., what are the limiting properties of X_n? We next give a WLLN and a CLT for weighted sums. The way arrays like a_{in} typically arise is that there are some weighting constants c_i, and either $a_{in} = c_i/\sum_{i=1}^n c_j$ or $a_{in} = c_i/[\sum_{i=1}^n c_j]^{1/2}$. If $c_i = 1$ for all i, then $a_{in} = n^{-1}$or $n^{-1/2}$, respectively, leading to the standard scaling in limit theorems.

Theorem 4.15. Assume random variables Y_i are independently identically distributed with mean zero. If an array a_{in} satisfies $\lim_{n\to\infty} \sum_{i=1}^n |a_{jn}| = 0$ and $\lim_{n\to\infty} \max_{j\le n} |a_{jn}| = 0$, then $X_n \to_p 0$.

Proof: This is a weighted version of Khinchine's WLLN, and is proved in the same way. Let $\zeta(t)$ be the second characteristic function of Y_1. From the properties of characteristic functions we have $\zeta'(0) = 0$ and a Taylor's expansion $\zeta(t) = t\cdot\zeta'(\lambda t)$ for some $0 < \lambda < 1$. The second characteristic function of X_n is then $\gamma(t) = \sum_{i=1}^{n} a_{in}t\cdot\zeta'(\lambda_{in}a_{in}t)$, implying $|\gamma(t)| \le \sum_{i=1}^{n} |a_{in}t\cdot\zeta'(\lambda_{in}a_{in}t)| \le |t|\cdot(\max_{i\le n}|\zeta'(\lambda_{in}a_{in}t)|)\cdot \sum_{i=1}^{n} |a_{in}|$. Then $\lim \sum_{i=1}^{n} |a_{in}| < \infty$ and $\lim (\max_{i\le n}|a_{in}|) = 0$ imply $\gamma(t) \longrightarrow 0$ for each t, and hence X_n converges in distribution, hence in probability, to 0. □

Theorem 4.16. Assume random variables Y_i are i.i.d. with mean zero and variance $\sigma^2 \in (0,+\infty)$. If an array a_{in} satisfies $\lim_{n\to\infty} \max_{j\le n} |a_{jn}| = 0$ and $\lim_{n\to\infty} \sum_{i=1}^{n} a_{in}^2 = 1$, then $X_n \to_d X_o \sim N(0,\sigma^2)$.

Proof: The argument parallels the Lindeberg-Levy CLT proof. The second characteristic function of X_n has the Taylor's expansion $\gamma(t) = -(1/2)\sigma^2t^2a_{in} + [\zeta''(\lambda_{in}a_{in}t)+\sigma^2]\cdot a_{in}^2t^2/2$, where $\lambda_{in} \in (0,1)$. The limit assumptions imply $\gamma(t) + (1/2)\sigma^2t^2$ is bounded in magnitude by

$$\sum_{i=1}^{n} |\zeta''(\lambda_{in}a_{in}t)+\sigma^2|\cdot a_{in}t^2/2 \le [\sum_{i=1}^{n} a_{in}^2t^2/2]\cdot\max_{i\le n}|\zeta''(\lambda_{in}a_{in}t)+\sigma^2|.$$

This converges to zero for each t since $\lim_{n\to\infty}\max_{i\le n}|\zeta''(\lambda_{in}a_{in}t)+\sigma^2| \to 0$. Therefore, $\gamma(t)$ converges to the characteristic function of a normal with mean 0 and variance σ^2. □

4.5.3. The limit theorems 4.13 and 4.14 are special cases of a limit theory for what are called *triangular arrays* of random variables, Y_{nt} with t = 1,2,...,n and n = 1,2,3,... . (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. We first give limit theorems for Y_{nt} that are uncorrelated or independent within each row. These are by no means the strongest obtainable, but they have the merit of simple proofs.

Theorem 4.17. Assume random variables Y_{nt} for t = 1,2,...,n and n = 1,2,3,... are uncorrelated across t for each n, with $\mathbf{E}\, Y_{nt} = 0$, $\mathbf{E}\, Y_{nt}^2 = \sigma_{nt}^2$. Then, $\sum_{i=1}^{n} \sigma_{nt}^2 \to 0$ implies $\sum_{i=1}^{n} Y_{nt} \to_p 0$.

Proof: Apply Chebyshev's inequality. □

Theorem 4.18. Assume random variables Y_{nt} for t = 1,2,...,n and n = 1,2,3,... are independent across t for each n, with $\mathbf{E}\, Y_{nt} = 0$, $\mathbf{E}\, Y_{nt}^2 = \sigma_{nt}^2$, $\sum_{i=1}^n \sigma_{nt}^2 \to 1$, $\sum_{i=1}^n \mathbf{E}|Y_{nt}|^3 \to 0$, and $\sum_{i=1}^n \sigma_{nt}^4 \to 0$. Then $X_n = \sum_{i=1}^n Y_{nt} \to_d X_o \sim N(0,1)$.

Proof: From the properties of characteristic functions (see 3.5.12), the c.f. of Y_{nt} has a Taylor's expansion that satisfies $|\psi_{nt}(s) - 1 + s^2\sigma_{nt}^2/2| \le |s|^3\mathbf{E}|Y_{nt}|^3/6$. Therefore, the c.f. $\gamma_n(s)$ of X_n satisfies $\log \gamma_n(s) = \sum_{i=1}^n \log(1 - s^2\sigma_{nt}^2/2 + \lambda_{nt}|s|^3\mathbf{E}|Y_{nt}|^3/6)$, where $|\lambda_{nt}| \le 1$. From 2.1.10, we have the inequality that for $|a| < 1/3$ and $|b| < 1/3$, $|\mathrm{Log}(1+a+b) - a| < 4|b| + 3|a|^2$. Then, the assumptions guarantee that $|\log \gamma_n(s) + s^2 \sum_{i=1}^n \sigma_{nt}^2/2| \le 4|s|^3 \sum_{i=1}^n \mathbf{E}|Y_{nt}|^3/6 + 3\, s^4 \sum_{i=1}^n \sigma_{nt}^4/4$. The assumptions then imply that $\log \gamma_n(s) \to -s^2/2$, establishing the result. □

In the last theorem, note that if $Y_{nt} = n^{-1/2}Z_t$, then $E|Z_t|^3$ bounded is sufficient to satisfy all the assumptions. Another set of limit theorems can be stated for triangular arrays with the property that the random variables within each row form a martingale difference sequence. 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 $E\{Y_{nt}|\boldsymbol{G}_{n,t-1}\} = 0$; this is called a *martingale difference array*. A WLLN for this case is adapted from J. Davidson (1994), p. 299.

Theorem 4.19. If Y_{nt} and $\boldsymbol{G}_{nt}$ for t = 1,...,n and n = 1,2,3,... is an adapted martingale difference array with $|Y_{nt}| \le M$, $\mathbf{E}\, Y_{nt}^2 = \sigma_{nt}^2$, $\sum_{i=1}^n \sigma_{nt}$ uniformly bounded, and $\sum_{i=1}^n \sigma_{nt}^2 \to 0$, then $\sum_{i=1}^n Y_{nt} \to_p 0$.

The following CLT for martingale difference arrays is taken from D. Pollard (1984), p. 170-174.

Theorem 4.20. If 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}, $\sum_{i=1}^n \lambda_{nt}^2 \to_p \sigma^2 \in (0,+\infty)$, and if for each ε > 0, $\sum_{t=1}^n \mathbf{E}\, Y_{nt}^2\cdot\mathbf{1}(|Y_{nt}| > \varepsilon) \to 0$, then $X_n = \sum_{i=1}^n Y_{nt} \to_d X_o \sim N(0,\sigma^2)$.

4.5.4. Chebyshev's inequality gives an easy, but crude, bound on the probability in the tail of a density. For random variables with well behaved tails, there are sharper bounds that can be used to get sharper limit theorems. The following inequality, due to Hoeffding, is one of a series of results called *exponential inequalities* that are stated and proved in D. Pollard (1984), p. 191-193: If Y_n are independent random variables with zero means that satisfy the bounds $-a_n \le Y_n \le b_n$, then $P(n^{-1}\sum_{i=1}^n Y_i \ge \varepsilon) \le \exp(-2n^2\varepsilon^2/\sum_{i=1}^n (b_i+a_i)^2)$. Note that in Hoeffding's inequality, if $|Y_n| \le M$, then $P(|n^{-1}\sum_{i=1}^n Y_i| \ge \varepsilon) \le 2\cdot\exp(-n\varepsilon^2/2M^2)$. The next theorem gets a strong law of large numbers with weaker than usual scaling:

Theorem 4.21. If Y_n are independent random variables with zero means and $|Y_n| \le M$, then $X_n = n^{-1}\sum_{i=1}^n Y_i$ satisfies $X_k\cdot k^{1/2}/\log(k) \to_{as} 0$.

Proof: Hoeffding's inequality implies $\text{Prob}(k^{1/2}|X_k| > \varepsilon\cdot\log k) < 2\cdot\exp(-(\log k)\varepsilon^2/2M^2)$, and hence

$$\sum_{k=n+1}^{\infty} \text{Prob}(k^{1/2}|X_k| > \varepsilon\cdot\log k) \le \int_{z=n}^{\infty} 2\cdot\exp(-(\log z)^2\varepsilon^2/2M^2)dz$$

$$\le (6/\varepsilon)\cdot\exp(M^2/2\varepsilon^2)\cdot\Phi(-\varepsilon\cdot(\log n)/M + M/\varepsilon),$$

with the standard normal CDF Φ resulting from direct integration. Applying Theorem 4.2, this inequality implies $n^{1/2}|X_n|/\log n \to_{as} 0$. □

If the Y_i are not necessarily bounded, but have a proper moment generating function, one can get an exponential bound from the moment generating function.

Theorem 4.22. If i.i.d. mean-zero random variables Y_i have a proper moment generating function, then $X_n = n^{-1}\sum_{i=1}^n Y_i$ satisfies $P(X_n > \varepsilon) < \exp(-\tau\varepsilon n^{1/2}+\kappa)$, where τ and κ are positive constants determined by the distribution of Y_i.

Proof: $P(Z > \varepsilon) = \int_{z>\varepsilon} F(dz) \le \int_{z>\varepsilon} e^{(z-\varepsilon)t}F(dz) \le e^{-\varepsilon t}\mathbf{E}e^{Zt}$ for a random variable Z. Let m(t) be the moment generating function of Y_i and τ be a constant such that m(t) is finite for $|t| < 2\tau$. Then one has $m(t) = 1 + m''(\lambda t)t^2/2$ for some $|\lambda| < 1$, for each $|t| < 2\tau$, from the properties of mgf (see 3.5.12).

The mgf of X_n is $m(t/n)^n = (1 + m''(\lambda t/n)t^2/2n^2)^n$, finite for $|t|/n \le 2\tau$. Replace t/n by $\tau n^{-1/2}$ and observe that $m''(\lambda t/n) \le m''(\tau n^{-1/2})$ and $(1+m''(\tau n^{-1/2})\tau^2/2n)^n \le \exp(m''(\tau n^{-1/2})\,\tau^2/2)$. Substituting these expressions in the initial inequality gives $P(X_n > \varepsilon) \le \exp(-\tau\varepsilon n^{1/2} + m''(\tau n^{-1/2})\,\tau^2/2\,)$, and the result holds with $\kappa = m''(\tau)\tau^2/2$. □

Using the same argument as in the proof of Theorem 4.19 and the inequality $P(X_n > \varepsilon) < \exp(-\tau\varepsilon n^{1/2}+\kappa)$ from Theorem 4.20, one can show that $X_k\cdot k^{1/2}/(\log k)^2 \to_{as} 0$, a SLLN with weak scaling.

4.5.5. This section states a uniform SLLN for random functions on compact set Θ in a Euclidean space $\mathbb{R}^k$. Let $(\mathbf{S},\boldsymbol{F},P)$ denote a probability space. Define a *random function* as a mapping Y from $\Theta\times S$ into $\mathbb{R}$ with the property that for each $\theta \in \Theta$, $Y(\theta,\cdot)$ is measurable with respect to (S,F,P). Note that $Y(\theta,\cdot)$ is simply a random variable, and that $Y(\cdot,s)$ is simply a function of $\theta \in \Theta$. Usually, the dependence of Y on the state of nature is suppressed, and we simply write $Y(\theta)$. A random function is also called a *stochastic process*, and $Y(\cdot,s)$ is termed a *realization* of this process. A random function $Y(\theta,\cdot)$ is *almost surely continuous* at $\theta_o \in \Theta$ if for s in a set that occurs with probability one, $Y(\cdot,s)$ is continuous in θ at θ_o. It is useful to spell out this definition in more detail. For each $\varepsilon > 0$, define $\mathbf{A}_k(\varepsilon,\theta_o) = \left\{ s\in\boldsymbol{S} \,\middle|\, \sup_{|\theta-\theta_o|\le 1/k} |Y(\theta,s)-Y(\theta_o,s)|>\varepsilon \right\}$. Almost sure continuity states that these sets converge monotonically as $k\to\infty$ to a set $\mathbf{A}_o(\varepsilon,\theta_o)$ that has probability zero.

The condition of almost sure continuity allows the modulus of continuity to vary with s, so there is not necessarily a fixed neighborhood of θ_o independent of s on which the function varies by less than ε. For example, the function $Y(\theta,s) = \theta^s$ for $\theta \in [0,1]$ and s uniform on [0,1] is continuous at $\theta = 0$ for every s, but $\mathbf{A}_k(\varepsilon,0) = [0,(-\log\varepsilon)/(\log k))$ has positive probability for all k. The exceptional sets $\mathbf{A}_k(\varepsilon,\theta)$ can vary with θ, and there is no requirement that there be a set of s with probability one, or for that matter with positive probability, where $Y(\theta,s)$ is continuous for all θ. For example, assuming $\theta \in [0,1]$ and s uniform on [0,1], and defining $Y(\theta,s) = 1$ if $\theta \ge s$ and $Y(\theta,s) = 0$ otherwise gives a function that is almost surely continuous everywhere and always has a discontinuity.

Theorem 4.3 in Section 4.1 established that convergence in probability is preserved by continuous mappings. The next result extends this to almost surely continuous transformations; the result below is taken from Pollard (1984), p. 70.

Theorem 4.23. (Continuous Mapping**).** If $Y_n(\theta) \to_p Y_o(\theta)$ uniformly for θ in $\Theta \subseteq \mathbb{R}^k$, random vectors $\tau_o,\tau_n \in \Theta$ satisfy $\tau_n \to_p \tau_o$, and $Y_o(\theta)$ is almost surely continuous at τ_o, then $Y_n(\tau_n) \to_p Y_o(\tau_o)$.

Consider i.i.d. random functions $Y_i(\theta)$ that have a finite mean $\psi(\theta)$ for each θ, and consider the average $X_n(\theta) = n^{-1}\sum_{i=1}^n Y_i(\theta)$. Kolmogorov's SLLN I implies that pointwise, $X_n(\theta) \to_{as} \psi(\theta)$. However, we sometimes need in statistics a stronger result that $X_n(\theta)$ is uniformly close to $\psi(\theta)$ over the whole domain Θ. This is <u>not</u> guaranteed by pointwise convergence. For example, the random function $Y_n(s,\theta) = 1$ if $n^2\cdot|s - \theta| \le 1$, and $Y_n(s,\theta) = 0$ otherwise, where the sample space is the unit interval with uniform probability, has $P(Y_n(\cdot,\theta) > 0) \le 2/n^2$ for each θ. This is sufficient to give $Y_n(\cdot,\theta) \to_{as} 0$ pointwise. However, $P(\sup_\theta Y_n(\theta) > 0) = 1$.

Theorem 4.24. (Uniform SLLN**).** Assume $Y_i(\theta)$ are independent identically distributed random functions with a finite mean $\psi(\theta)$ for θ in a closed bounded set $\Theta \subseteq \mathbb{R}^k$. Assume $Y_i(\cdot)$ is almost surely continuous at each $\theta \in \Theta$. Assume that $Y_i(\cdot)$ is dominated; i.e., there exists a random variable Z with a finite mean that satisfies $Z \ge \sup_{\theta\in\Theta}|Y_1(\theta)|$. Then $\psi(\theta)$ is continuous in θ and $X_n(\theta) = \dfrac{1}{n}\sum_{i=1}^n Y_i(\theta)$ satisfies $\sup_{\theta\in\Theta}|X_n(\theta) - \psi(\theta)| \to_{as} 0$.

Proof: We follow an argument of Tauchen (1985). Let $(\mathbf{S},\boldsymbol{F},P)$ be the probability space, and write the random function $Y_i(\theta,s)$ to make its dependence on the state of Nature explicit. We have $\psi(\theta) = \int_S Y(\theta,s)P(ds)$. Define $u(\theta_o,s,k) = \sup_{|\theta-\theta_o|\le 1/k} |Y(\theta,s) - Y(\theta_o,s)|$. Let $\varepsilon > 0$ be given. Let $\mathbf{A}_k(\varepsilon/2,\theta_o)$ be the measurable set given in the definition of almost sure continuity, and note that for $k = k(\varepsilon/2,\theta_o)$ sufficiently large, the probability of $\mathbf{A}_k(\varepsilon/2,\theta_o)$ is less than $\varepsilon/(4\cdot\mathbf{E}\,Z)$. Then,

$$\mathbf{E}u(\theta_o,\cdot,k) \le \int_{A_k(\varepsilon/2,\theta_o)} u(\theta_o,s,k)P(ds) + \int_{A_k(\varepsilon/2,\theta_o)^c} u(\theta_o,s,k)P(ds)$$

$$\le \int_{A_k(\varepsilon/2,\theta_o)} 2\cdot Z(s)\cdot P(ds) + \int_{A_k(\varepsilon/2,\theta_o)^c} (\varepsilon/2)\cdot P(ds) \le \varepsilon\,.$$

Let $\mathbf{B}(\theta_o)$ be an open ball of radius $1/k(\varepsilon/2,\theta_o)$ about θ_o. These balls constructed for each $\theta_o \in \Theta$ cover the compact set Θ, and it is therefore possible to extract a finite subcovering of balls $\mathbf{B}(\theta_j)$ with centers at points θ_j for $j = 1,...,J$. Let $\mu_j = \mathbf{E}u(\theta_j,\cdot,k(\varepsilon/2,\theta_j)) \le \varepsilon$. For $\theta \in \mathbf{B}(\theta_j)$, $|\psi(\theta) - \psi(\theta_j)| \le \mu_j \le \varepsilon$. Then

$$\sup_{\theta\in B(\theta_j)} |X_n(\theta) - \psi(\theta)| \le |X_n(\theta) - X_n(\theta_j) - \mu_j| + \mu_j + |X_n(\theta_j) - \psi(\theta_j)| + |\psi(\theta_j) - \psi(\theta)|$$

$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} u(\theta_j,\cdot,k(\varepsilon/2,\theta_j)) - \mu_j \right| + \varepsilon + \left| X_n(\theta_j) - \psi(\theta_j) \right| + \varepsilon .$$

Apply Kolmogorov's SLLN to each of the first and third terms to determine a sample size n_j such that

$$P(\sup_{n \geq n_j} \left| n^{-1} \sum_{i=1}^{n} u(\theta_j,\cdot,k(\varepsilon/2,\theta_j)) - \mu_j \right| > \varepsilon) < \varepsilon/2J$$

and

$$P(\sup_{n \geq n_j} \left| X_n(\theta_j) - \psi(\theta_j) \right| > \varepsilon) < \varepsilon/2J .$$

With probability at least $1 - \varepsilon/J$, $\sup_{\theta \in B(\theta_j)} \left| X_n(\theta) - \psi(\theta) \right| \leq 4\varepsilon$. Then, with probability at least $1 - \varepsilon$,

$\sup_{\theta \in \Theta} \left| X_n(\theta) - \psi(\theta) \right| \leq 4\varepsilon$ for $n > n_o = \max(n_j)$. □

The construction in the proof of the theorem of a finite number of approximating points can be reinterpreted as the construction of a finite family of functions, the $Y(\theta_j,\cdot)$, with the approximation property that the expectation of the absolute difference between $Y(\theta,\cdot)$ for any θ and one of the members of this finite family is less than ε. Generalizations of the uniform SLLN above can be obtained by recognizing that it is this approximation property that is critical, with a limit on how rapidly the size of the approximating family can grow with sample size for a given ε, rather than continuity per se; see D. Pollard (1984).

REFERENCES

P. Billingsley (1968) Convergence of Probability Measures, Wiley.
P. Billingsley (1986) Probability and Measure, Wiley.
J. Davidson (1994) Stochastic Limit Theory, Oxford.
W. Feller (1966) An Introduction to Probability Theory and Its Applications, Wiley.
I. Ibragimov and Y. Linnik, Independent and Stationary sequences of Random Variables, Wolters-Noordhoff, 1971.
J. Neveu (1965) Mathematical Foundations of the Calculus of Probability, Holden-Day.
D. Pollard (1984) Convergence of Stochastic Processes, Springer-Verlag.
C. R. Rao (1973) Linear Statistical Inference and Its Applications, Wiley.
R. Serfling (1970) "Convergence Properties of S_n Under Moment Restrictions," *Annals of Mathematical Statistics*, 41, 1235-1248.
R. Serfling (1980) Approximation Theorems of Mathematical Statistics, Wiley.
G. Tauchen (1985)
H. White (1984) Asymptotic Theory for Econometricians, Academic Press.