Chapter 4. Limit Theorems in Statistics

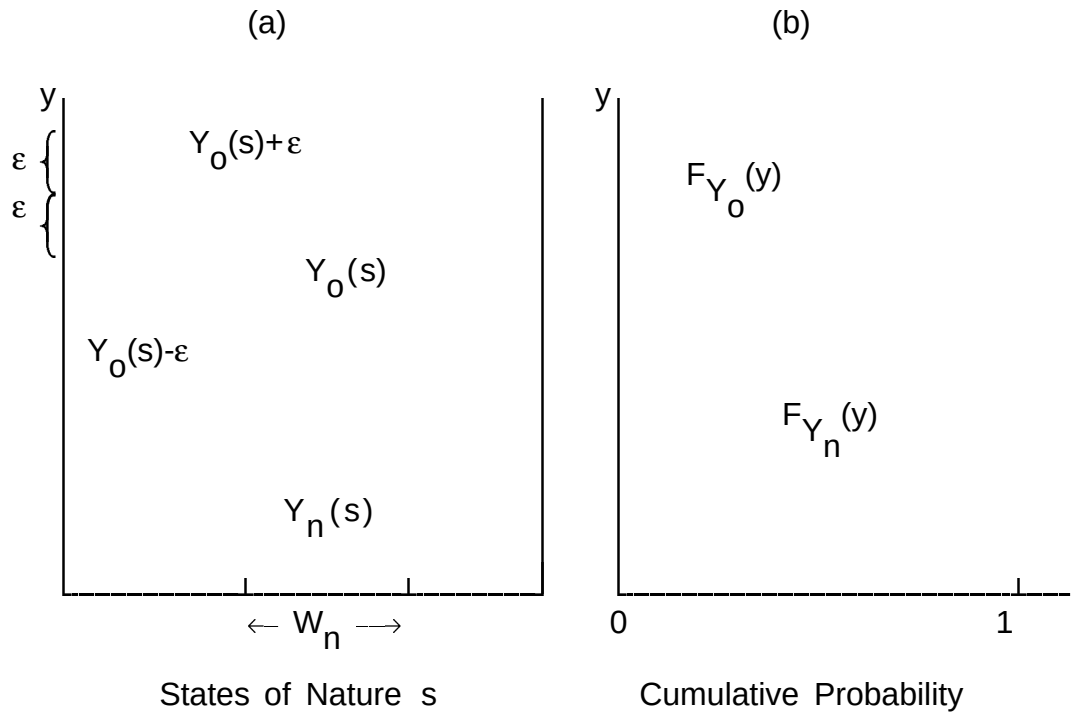
1. Sequences of Random Variables

1.1 Consider a sequence of random variables $Y_1, Y_2 ..., Y_n, ...$. These random variables are all functions of the same state of Nature s, but may depend on different parts of s. The *joint distribution (CDF)* of a finite subsequence $(Y_1,...,Y_n)$, denoted by $F_{Y_1,..,Y_n}(y_1,...,y_n)$, is defined as the probability of a state of nature such that all of the inequalities $Y_1 \leq y_1,...,Y_n \leq y_n$ hold. The random variables in the sequence are *independent* if for every finite sub-sequence $Y_1...,Y_n$, the joint CDF factors:

$$F_{Y_1,...Y_n}(y_1,...,y_n) \equiv F_{Y_1}(y_1)\cdot...\cdot F_{Y_n}(y_n).$$

1.2 There are several possible concepts for the limit of a sequence of random variables, $\lim_{n\to\infty} Y_n = Y_o$. Since the Y_n are functions of states of nature, these limit concepts will correspond to different ways of defining limits of functions. Panel (a) of Figure 1 graphs Y_n and Y_o as functions of the state of nature; W_n is the set of states of Nature for which $Y_o(s)$ and $Y_n(s)$ differ by more than $\varepsilon > 0$. Panel (b) graphs the CDF's of Y_o and Y_n. Note that these CDF's are rotated so the probability is on the horizontal axis. The limit definitions below will be discussed using this figure.

FIGURE 1. Convergence Concepts



1.3 Y_n *converges in probability* to Y_o, denoted $Y_n \xrightarrow{p} Y_o$, if for each $\varepsilon > 0$, $\lim_{n\to\infty} \text{Prob}(|Y_n - Y_o| > \varepsilon) = 0$. In Figure 1, W_n is the set of states of nature for which $|Y_n(w) - Y_o(w)| > \varepsilon$; $Y_n \xrightarrow{p} Y_o$ means $\text{Prob}(W_n) \to 0$.

1.4 Y_n *converges almost surely* to Y_o, denoted $Y_n \xrightarrow{as} Y_o$, if for each $\varepsilon > 0$, $\lim_{n\to\infty} \text{Prob}(\sup_{n'\geq n} |Y_{n'} - Y_o| > \varepsilon) = 0$. For W_n defined in Figure 1, the set of states of nature for which $|Y_{n'}(w) - Y_o(w)| > \varepsilon$ for <u>some</u> $n' \geq n$ is $\bigcup_{n'\geq n} W_{n'}$. Then $Y_n \xrightarrow{as} Y_o$ means $\text{Prob}(\bigcup_{n'\geq n} W_{n'}) \to 0$. An implication of almost sure convergence is $\lim Y_n(s) = Y_o(s)$ for almost all w (i.e., except for a set of states of nature of probability zero); this is

not an implication of $Y_n \xrightarrow{p} Y_o$.

1.5 Y_n *converges in order-*ρ *mean* (also called *convergence in* $\|\cdot\|_\rho$ *norm,* or *convergence in* L_ρ *space*) to Y_o if $\lim_{n\to\infty} \mathbf{E}|Y_n - Y_o|^\rho = 0$. For $\rho = 2$, this is called *convergence in quadratic mean*. In Figure 1(a), the ρ-order mean corresponds to raising $|Y_n(s) - Y_o(s)|$ to power ρ, weighting by the probability measure P(dw), and integrating; the resulting sequence of scalars has limit zero as $n \to \infty$ when this type of convergence holds.

1.6 Y_n *converges in distribution* to Y_o, denoted $Y_n \xrightarrow{d} Y_o$, if the CDF of Y_n converges to the CDF of Y_o at each continuity point of Y_o; i.e., if F_{Y_o} is continuous at y, then $\lim_{n\to\infty} F_{Y_n}(y) = F_{Y_o}(y)$. In Figure 1(b), this means that F_{Y_n} converges to the function F_{Y_o} point by point, except at jumps in F_{Y_o}. If A is an open set, then $Y_n \xrightarrow{d} Y_o$ implies $\lim_{n\to\infty} \inf_{n' \geq n} F_{Y_n}(A) \geq F_{Y_o}(A)$; see Billingsley, Convergence of Probability Measures, Theorem 2.1. Convergence in distribution is also called *weak convergence* in the space of distribution functions.

1.7 The relationships between different types of convergence are summarized in Figure 2. Explanations and examples follow. On first reading, skim Sections 1.8-1.18, and skip the proofs.

FIGURE 2. Relationships Between Stochastic Limits

(Section numbers for details are given in parentheses)

$$Y_n \xrightarrow{as} Y_o \quad \overset{(1.8)}{\models\!\!\Rightarrow} \quad Y_n \xrightarrow{p} Y_o \quad \overset{(1.8)}{\models\!\!\Rightarrow} \quad Y_n \xrightarrow{d} Y_o$$

$$\Updownarrow (1.4) \qquad \Updownarrow (1.3) \qquad \Uparrow (1.10)$$

$$Y_n - Y_o \xrightarrow{as} 0 \quad \overset{(1.8)}{\models\!\!\Rightarrow} \quad Y_n - Y_o \xrightarrow{p} 0 \quad \overset{(1.10)}{\Leftarrow\!\!\models} \quad Y_n - Y_o \xrightarrow{d} 0$$

$$Y_n \xrightarrow{d} c \text{ (a constant)} \quad \overset{(1.10)}{\models\!\!\Rightarrow} \quad Y_n \xrightarrow{p} c$$

$$Y_n \xrightarrow{d} Y_o \quad \overset{(1.11)}{\Leftarrow\!\!\Rightarrow} \quad \mathbf{E}g(Y_n) \to \mathbf{E}g(Y_o) \text{ for all bounded continuous } g$$

$$\mathbf{E}|Y_n - Y_o|^{\rho} \to 0 \text{ (for some } \rho > 0) \quad \overset{(1.12)}{\models\!\!\Rightarrow} \quad Y_n \xrightarrow{p} Y_o$$

$$\mathbf{E}|Y_n - Y_o|^{\rho} \le M \text{ (all n) \& } Y_n \xrightarrow{p} Y_o \quad \overset{(1.13)}{\models\!\!\Rightarrow} \quad \mathbf{E}|Y_n - Y_o|^{\lambda} \to 0 \text{ for } 0 < \lambda < \rho$$

$$Y_n \xrightarrow{p} Y_o \quad \overset{(1.14)}{\models\!\!\Rightarrow} \quad Y_{n_k} \xrightarrow{as} Y_o \text{ for some subsequence } n_k,\ k = 1,2,...$$

$$\sum_n \mathbf{E}|Y_n - Y_o|^{\rho} < +\infty \text{ (for some } \rho > 0) \quad \overset{(1.15)}{\models\!\!\Rightarrow} \quad Y_n \xrightarrow{as} Y_o$$

$$Y_n \xrightarrow{d} Y_o \text{ \& } Z_n - Y_n \xrightarrow{p} 0 \quad \overset{(1.16)}{\models\!\!\Rightarrow} \quad Z_n \xrightarrow{d} Y_o$$

$$Y_n \xrightarrow{p} Y_o \quad \overset{(1.17)}{\models\!\!\Rightarrow} \quad g(Y_n) \xrightarrow{p} g(Y_o) \quad \text{for all continuous } g$$

$$Y_n \xrightarrow{d} Y_o \quad \overset{(1.18)}{\models\!\!\Rightarrow} \quad g(Y_n) \xrightarrow{d} g(Y_o) \quad \text{for all continuous } g$$

1.8 $Y_n \xrightarrow{as} Y_o$ implies $Prob(W_n) \leq Prob(\bigcup_{n' \geq n} W_{n'}) \to 0$, and hence $Y_n \xrightarrow{p} Y_o$. However, one can construct examples where $Prob(W_n) \to 0$ but $\bigcup_{n' \geq n} W_{n'}$ is the set of states of Nature which has probability one. Hence $Y_n \xrightarrow{p} Y_o$ does not imply $Y_n \xrightarrow{as} Y_o$. To illustrate, take the universe of states of nature to be the points on the unit circle with uniform probability, take the W_n to be successive arcs of length π/n, and take Y_n to be 1 on W_n, 0 otherwise. Then $Y_n \xrightarrow{p} 0$ 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 W_n.

1.9 Suppose $Y_n \xrightarrow{p} Y_o$. It is a good exercise in manipulation of probabilities of events to show that $Y_n \xrightarrow{d} Y_o$. Given $\varepsilon > 0$, define W_n as before to be the set of states of Nature where $|Y_n(s) - Y_o(s)| > \varepsilon$. Given y, define A_n, B_o, and C_o to be, respectively, the states of Nature with $Y_n \leq y$, $Y_o \leq y - \varepsilon$, and $Y_o \leq y + \varepsilon$. Then $B_o \subseteq A_n \cup W_n$ (i.e., $Y_o(s) \leq y - \varepsilon$ implies either $Y_n(s) \leq y$ or $|Y_o(s) - Y_n(s)| > \varepsilon$) and $A_n \subseteq C_o \cup W_n$ (i.e., $Y_n(s) \leq y$ implies $Y_o(s) \leq y + \varepsilon$ or $|Y_o(s) - Y_n(s)| > \varepsilon$). Hence, for n large enough so $Prob(W_n) < \varepsilon$,

$$F_{Y_o}(y-\varepsilon) \equiv Prob(B_o) \leq Prob(A_n) + Prob(W_n) < F_{Y_n}(y) + \varepsilon$$

and

$$F_{Y_n}(y) \equiv Prob(A_n) \leq Prob(C_o) + Prob(W_n) < F_{Y_o}(Y+\varepsilon) + \varepsilon,$$

implying $F_{Y_o}(y-\varepsilon) - \varepsilon \leq \lim_{n\to\infty} F_{Y_n}(y) \leq F_{Y_o}(y+\varepsilon) + \varepsilon$. If y is a continuity point of Y_o, then $F_{Y_o}(y-\varepsilon)$ and $F_{Y_o}(y+\varepsilon)$ approach $F_{Y_o}(y)$ as $\varepsilon \to 0$, implying $\lim_{n\to\infty} F_{Y_n}(y) = F_{Y_o}(y)$. This establishes that $Y_n \xrightarrow{d} Y_o$.

Convergence in distribution of Y_n to Y_o does not alone imply that Y_n and Y_o are

close to each other as $n \to \infty$. For example, if Y_n and Y_o are independently identically distributed (i.i.d.) standard normal, then $Y_n \xrightarrow{d} Y_o$ trivially, but clearly not $Y_n \xrightarrow{p} Y_o$ since $Y_n - Y_o$ is normal with variance 2, and $|Y_n - Y_o| > \varepsilon$ with a positive, constant probability.

1.10 Convergence in distribution and convergence in probability to a <u>constant</u> are equivalent. If $Y_n \xrightarrow{p} c$ constant, then $Y_n \xrightarrow{d} c$ as a special case of (1.9) above. Conversely, $Y_n \xrightarrow{d} c$ constant means $F_{Y_n}(y) \to F_c(y)$ at continuity points, where $F_c(y) = 0$ for $y < c$ and $F_c(y) = 1$ for $y \geq c$. Then given $\varepsilon > 0$, $\text{Prob}(|Y_n - c| > \varepsilon) = F_{Y_n}(c-\varepsilon) + 1 - F_{Y_n}(c+\varepsilon) \to 0$, so $Y_n \xrightarrow{p} c$. This result implies particularly that the statements $Y_n - Y_o \xrightarrow{p} 0$ and $Y_n - Y_o \xrightarrow{d} 0$ are equivalent, and $Y_n - Y_o \xrightarrow{d} 0$ is <u>not</u> equivalent to $Y_n \xrightarrow{d} Y_o$.

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-Alexandroff theorem.

1.12 A Chebyshev-like inequality is obtained by noting for a random variable Z that

$$\mathbf{E}|Z|^\rho = \int |z|^\rho f_Z(z)dz \geq \int_{|z| \geq \varepsilon} \varepsilon^\rho f_Z(z)dz = \varepsilon^\rho \text{ Prob}(|Z| > \varepsilon),$$

or $\text{Prob}(|Z| > \varepsilon) \leq \mathbf{E}|Z|^\rho/\varepsilon^\rho$. (When $\rho = 2$, this is the conventional Chebyshev inequality. When $\rho = 1$, one has $\text{Prob}(|Z| > \varepsilon) \leq \mathbf{E}|Z|/\varepsilon$.) Taking $Z = Y_n - Y_o$, one has

$$\lim_{n\to\infty} \text{Prob}(|Y_n - Y_o| > \varepsilon) \le \varepsilon^{-\rho} \lim_{n\to\infty} \mathbf{E}|Y_n - Y_o|^{\rho}.$$

Hence, convergence in ρ-mean (for any $\rho > 0$) implies convergence in probability. An example shows, however, that convergence almost surely or in probability does not necessarily imply convergence in ρ-mean. Suppose Y_n is discrete with $f_{Y_n}(0) = 1 - n^{-2}$ and $f_{Y_n}(e^n) = n^{-2}$. Then $Y_n \xrightarrow{as} 0$ since

$$\text{Prob}(Y_{n'} \neq 0 \text{ for any } n' > n) \le \sum_{n'>n} f_{Y_{n'}}(e^{n'}) \le 1/n,$$

but $\mathbf{E}|Y_n|^{\rho} = e^{\rho n}/n^2 \to +\infty$ for any $\rho > 0$.

1.13 Adding a condition of a uniformly bounded ρ-order mean $\mathbf{E}|Y_n|^{\rho} \le M$ to convergence in probability $Y_n \xrightarrow{p} Y_o$ yields the result that for $0 < \lambda < \rho$, $\mathbf{E}|Y_o|^{\lambda}$ exists and $\mathbf{E}|Y_n|^{\lambda} \to \mathbf{E}|Y_o|^{\lambda}$. 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 2.

To prove this result, we will find useful the following property of moments: $\mathbf{E}|Y|^{\lambda} \le (\mathbf{E}|Y|^{\rho})^{\lambda/\rho}$ for $0 < \lambda < \rho$. This follows from Holder's inequality (Rao, Linear Statistical Models, p. 55, 149), which states

$$\mathbf{E}|UV| \le (\mathbf{E}|U|^{r})^{1/r}(\mathbf{E}|V|^{s})^{1/s} \text{ for } r,s > 0 \text{ and } r^{-1} + s^{-1} = 1,$$

by taking $U = |Y|^{\lambda}$, $V = 1$, and $r = \rho/\lambda$. To show that $\mathbf{E}|Y_o|^{\lambda}$ exists for $0 < \lambda \le \rho$, note that $\mathbf{E}|Y_n|^{\rho} \le M$ implies, by the property of moments above, that

$$M^{\lambda/\rho} \ge \mathbf{E}|Y_n|^{\lambda} = \int |y|^{\lambda} f_{Y_n}(y)dy \ge \int g(y) f_{Y_n}(y)dy$$

where $g(y) = \min(|y|^{\lambda}, k^{\lambda})$. Then, in particular,

$$M \ge \int_{-k}^{k} |y|^{\rho} f_{Y_n}(y)dy \longrightarrow k^{\rho}\text{Prob}(|Y_n| > k).$$

The Helly-Bray theorem implies, since g is continuous, that

$$\int g(y)f_{Y_n}(y)dy \to \int g(y)f_{Y_o}(y)dy = \int_{-k}^{k} |y|^{\lambda} f_{Y_o}(y)dy + k^{\lambda}\text{Prob}(|Y_o| > k).$$

Hence, $M \geq \int_{-k}^{k} |y|^{\rho} f_{Y_o}(y)dy$ for all k, implying that $\mathbf{E}|Y_o|^{\rho} \leq M$ exists, and given $\varepsilon > 0$, there exists k such that $\int_{|y|>k} |y|^{\rho} f_{Y_o}(y)dy < \varepsilon$. Also,

$$|\mathbf{E}|Y_n|^{\lambda} - \mathbf{E}g(Y_n)| \leq \int_{|y|>k} |y|^{\lambda} f_{Y_n}(y)dy \leq k^{\lambda-\rho} \int_{|y|>k} |y|^{\rho} f_{Y_n}(y)dy \leq Mk^{\lambda-\rho}.$$

Then, for k and n sufficiently large, $Mk^{\lambda-\rho} < \varepsilon$ and

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

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 $f_{Z_n}(0) = 1 - \frac{\log(n)}{n}$ and $f_{Z_n}(n) = \frac{\log(n)}{n}$. Then for $\lambda < 1$, $\mathbf{E}|Z_n|^{\lambda} = \log(n)/n^{1-\lambda} \to 0$, but $\mathbf{E}|Z_n|^{1} = \log(n) \to +\infty$.

1.14 If $Y_n \xrightarrow{p} Y_o$, then $\text{Prob}(W_n) \to 0$. Choose a subsequence n_k such that $\text{Prob}(W_{n_k}) \leq 2^{-k}$. Then $\text{Prob}(\bigcup_{k'>k} W_{n_{k'}}) \leq \sum_{k'>k} \text{Prob}(W_{n_{k'}}) \leq \sum_{k'>k} 2^{-k'} = 2^{-k}$, implying $Y_{n_k} \xrightarrow{as} Y_o$.

1.15 By the Chebyshev-like inequality in (1.12), $\text{Prob}(\bigcup_{n'\geq n} W_{n'}) \leq \sum_{n'\geq n} \text{Prob}(W_{n'}) \leq \sum_{n'\geq n} \mathbf{E}|Y_{n'} - Y_o|^{\rho}/\varepsilon^{\rho}$. If this right-hand expression is finite, then it goes to zero as

$n \to \infty$ implying $Y_n \xrightarrow{as} Y_o$. 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.

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 $Z_n \xrightarrow{d} Y_o$ also. In this case, Y_n and Z_n are termed *asymptotically equivalent*. The argument demonstrating this result is similar to that for 1.9. Let y be a continuity point of F_{Y_o} and define the following events:

$$A_n = \{s \mid Z_n(s) < y\},\ B_n = \{s \mid Y_n(s) \le y - \varepsilon\},$$
$$C_n = \{s \mid Y_n(s) \le y + \varepsilon\},\ D_n = \{s \mid\ \mid Y_n(s) - Z_n(s) \mid > \varepsilon\}.$$

Then $A_n \subseteq C_n \cup D_n$ and $B_n \subseteq A_n \cup D_n$, implying

$$F_{Y_n}(y-\varepsilon) - \text{Prob}(D_n) \le F_{Z_n}(y) \le F_{Y_n}(y+\varepsilon) + \text{Prob}(D_n).$$

Given $\delta > 0$, one can choose $\varepsilon > 0$ such that $y-\varepsilon$ and $y+\varepsilon$ are continuity points of F_{Y_o} and $F_{Y_o}(y+\varepsilon) - F_{Y_o}(y-\varepsilon) < \delta/3$. Then one can choose n sufficiently large so that $\text{Prob}(D_n) < \delta/3$, $\mid F_{y_n}(y+\varepsilon) - F_{Y_o}(y+\varepsilon) \mid < \delta/3$ and $\mid F_{Y_n}(y+\varepsilon) - F_{Y_o}(y+\varepsilon) \mid < \delta/3$. Then $\mid F_{Z_n}(y) - F_{Y_o}(y) \mid < \delta$.

1.17 A useful property of convergence in probability is that

$$\boxed{Y_n \xrightarrow{p} Y_o \Longrightarrow g(Y_n) \xrightarrow{p} g(Y_o) \text{ for any continuous function } g.}$$

If A denotes the support of Y_o, then this result holds if g is continuous on a neighborhood of A; i.e., given $\varepsilon > 0$ and $y \in A$, there exists $\delta > 0$ such that $\mid y'-y \mid < \delta$

$\Longrightarrow |g(y') - g(y)| < \varepsilon$. (Note in this definition that y' is not required to be in A.) This results holds for vectors of random variables as well, and specializes to the rules that if $Y_{1n} \xrightarrow{p} Y_{10}$ and $y_{2n} \xrightarrow{p} Y_{20}$, then

(a) $Y_{1n} \cdot Y_{2n} \xrightarrow{p} Y_{10} \cdot Y_{20}$

(b) $Y_{1n} + Y_{2n} \xrightarrow{p} Y_{10} + Y_{20}$

(c) If (Prob($|Y_{20}| < \varepsilon$) = 0 for some $\varepsilon > 0$, then $Y_{1n}/Y_{2n} \xrightarrow{p} Y_{10}/Y_{20}$.

These results obviously continue to hold when Y_{10} and/or Y_{20} are constants.

Demonstrating the result $g(Y_n) \xrightarrow{p} g(Y_o)$ is a good exercise in advanced calculus. Given $\varepsilon > 0$, choose M such that Prob($|Y_o| > M$) $< \varepsilon/2$, and define A_M to be the set of points in A satisfying $|y| \leq M$. Then A_M is closed and bounded. The mathematical properties of continuous functions on closed and bounded sets imply there exists $\delta > 0$ such that for all $y \in A_M$, $|y'-y| < \delta \Longrightarrow |g(y') - g(y)| < \varepsilon$. Choose n_o such that for $n > n_o$, Prob($|Y_n - Y_o| > \delta$) $< \varepsilon/2$. Then Prob($|g(Y_n) - g(Y_o)| > \varepsilon$) $\leq$ Prob($|Y_o| > M$ or $|Y_n - Y_o| > \delta$) $\leq \varepsilon$.

1.18 The preceding result has an analog for convergence in distribution:

$$Y_n \xrightarrow{d} Y_o \Longrightarrow g(Y_n) \xrightarrow{d} g(Y_o) \text{ for any continuous function } g$$

This result holds so long as g is continuous on a neighborhood of the support of Y_o. It also holds for vectors of random variables. To illustrate the result, suppose $Y_n \xrightarrow{d} Y_o$, with Y_o standard normal. Then Y_o^2 is chi-squared, implying that Y_n^2 converges in distribution to a chi-squared random variable.

The result $g(Y_n) \xrightarrow{d} g(Y_o)$ is now proved. Let S_ε be the open set of points a distance less than ε from the support of Y_o; Then g is continuous on $S_{3\varepsilon}$ for some $\varepsilon > 0$. From mathematical analysis, there is a continuous function $\lambda(y)$ that equals 1 for $y \in S_\varepsilon$ and 0 for $y \notin S_{2\varepsilon}$. Then $\tilde{g}(y) = g(y)\lambda(y)$ is continuous for all y. Let F_n

be the CDF of Y_n, F the CDF of Y_o, G_n the CDF of $Z_n = \tilde{g}(Y_n)$, and G the CDF of $Z = \tilde{g}(Y_o) \equiv g(Y_o)$. By 1.11, for any bounded continuous function h, $\mathbf{E}_{Z_n} h(Z_n) \equiv \mathbf{E}_{Y_n} h(\tilde{g}(Y_n)) \longrightarrow \mathbf{E}_{Y_o} h(\tilde{g}(Y_o)) \equiv \mathbf{E}_Z h(Z)$ since $h(\tilde{g}(y))$ is a bounded continuous function and $Y_n \xrightarrow{d} Y_o$. Then, $Z_n \xrightarrow{d} Z$. But for any $\varepsilon' > 0$, $\text{Prob}(|\tilde{g}(Y_n)-g(Y_n)| > \varepsilon') \leq \text{Prob}(Y_n \notin S_\varepsilon) \equiv 1 - F_{Y_n}(S_\varepsilon)$ and, from 1.8, $\limsup F_{Y_n}(S_\varepsilon) \geq F_{Y_o}(S_\varepsilon) = 1$. Then $\tilde{g}(Y_n) - g(Y_n) \xrightarrow{p} 0$, implying by 1.16 that $g(Y_n) \xrightarrow{d} g(Y_o)$.

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 $Y_n \xrightarrow{p} Y_o \Longleftrightarrow Y_n = Y_o + o_p(1)$, and $n^{-\alpha}(Y_n - Y_o) \xrightarrow{p} 0 \Longleftrightarrow Y_n - Y_o = o_p(n^\alpha)$. Thus $o_p(\cdot)$ is a notation for convergence in probability to zero of a suitably normalized sequence of random variables.

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*. An abbreviated list of rules for o_p and O_p is given in Figure 3.

A sequence that is convergent in distribution is stochastically bounded, but the reverse is not necessarily true. We first show convergence in distribution implies stochastic boundedness: If $Y_n \xrightarrow{d} Y_o$, then one can find M and n_o such that $\pm$ M are continuity points of Y_o, $\text{Prob}(|Y_o| \leq M) > 1-\varepsilon/2$, and for $n > n_o$, F_{Y_n} and F_{Y_o} evaluated at M or -M respectively differ by at most $\varepsilon/4$. 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 3. Rules for $O_p(\cdot)$ and $o_p(\cdot)$

Definition: $Y_n = o_p(n^\alpha) \Longleftrightarrow \text{Prob}(|n^{-\alpha}Y_n| > \varepsilon) \to 0$ for each $\varepsilon > 0$.

Definition: $Y_n = O_p(n^\alpha) \Longleftrightarrow$ 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. $Y_n = o_o(n^\alpha)$ & $Z_n = o_p(n^\beta)$ & $\beta \geq \alpha \Longrightarrow Y_n + Z_n = o_p(n^\beta)$
8. $Y_n = O_p(n^\alpha)$ & $Z_n = O_p(n^\beta)$ & $\beta \geq \alpha \Longrightarrow Y_n + Z_n = O_p(n^\beta)$
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. $Y_n = O_p(n^\alpha)$ & $Z_n = o_p(n^\alpha) \Longrightarrow Y_n + Z_n = O_p(n^\alpha)$

As an illustration of 1.19, we prove the very useful rule 6:

$n^{-\alpha}Y_n$ stochastically bounded & $n^{-\beta}Z_n \xrightarrow{p} 0 \Longrightarrow n^{-\alpha-\beta}Y_nZ_n \xrightarrow{p} 0$

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 $\text{Prob}(|n^{-\alpha-\beta}Y_nZ_n| > \varepsilon) \leq \text{Prob}(|n^{-\alpha}Y_n| > M) + \text{Prob}(|n^{-\beta}Z_n| > \varepsilon/M) < \varepsilon$. Demonstration of the remaining rules is left as an exercise.

2. Laws of Large Numbers

2.1 Consider a sequence of random variables $Y_1, Y_2, ..., Y_n, ...$ and a corresponding sequence of averages $X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Laws of large numbers are concerned with the conditions under which the X_n converge to a constant, either in probability (weak laws) or almost surely (strong laws).

2.2 Figure 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 WLLN even for correlated variable provided the correlations of distant random variables approach zero sufficiently rapidly.

To illustrate the basis of laws of large numbers, the proofs of the first three WLLN will be outlined. Consider first Khinchine's theorem. Let $\psi(t)$ be the characteristic function of Y_1. Then X_n has characteristic function $\psi(t/n)^n$. Since $\mathbf{E}Y_1$ exists, ψ has a Taylor's expansion $\psi(t) = 1 + \psi'(\lambda t)t$, where $0 < \lambda < 1$. Then $\psi(t/n)^n = [1 + \frac{t}{n}\psi'(\lambda\frac{t}{n})]^n$. But $\psi'(\lambda t/n) \longrightarrow \psi'(0) = \iota\mu$. There is a mathematical result stating that when a sequence of scalars α_n has a limit, then $[1+\alpha_n/n]^n \longrightarrow \exp(\lim\ \alpha_n)$. Then $\psi(t/n)^n \longrightarrow e^{\iota\mu t}$. But this is the characteristic function of a constant random variable μ, implying $X_n \xrightarrow{d} \mu$, and hence $X_n \xrightarrow{p} \mu$.

Figure 4. Laws of Large Numbers for $X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$

WEAK LAWS (WLLN)

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

2. (Chebyshev) If the Y_n are uncorrelated with $\mathbf{E}\, Y_n = \mu$ and

   $\mathbf{E}(Y_n - \mu)^2 \equiv \sigma_n^2$ satisfying $\lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 = 0$, then $X_n \xrightarrow{p} \mu$

3. If the Y_n have $\mathbf{E}\, Y_n = \mu$, $\mathbf{E}(Y_n-\mu)^2 \equiv \sigma_n^2$, and $|\mathbf{E}(Y_n-\mu)(Y_m-\mu)| \leq$

   $\rho^{|n-m|}\sigma_n\sigma_m$ with $\lim_{n\to\infty} \frac{1}{n^2}\sum_{i=0}^{n} \sigma_i^2 = 0$ and $\sum_{i=0}^{\infty} \rho_i^2 < +\infty$, then $X_n \xrightarrow{p} \mu$

STRONG LAWS (SLLN)

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

2. (Kolmogorov II) If the Y_n are independent, $\mathbf{E}\, Y_n = \mu$, and

   $\mathbf{E}(Y_n-\mu)^2 \equiv \sigma_n^2$ satisfying $\sum_{n=1}^{\infty} \sigma_n^2/n^2 < +\infty$, then $X_n \xrightarrow{as} \mu$

3. (Kolmogorov III) If the Y_n are uncorrelated, $\mathbf{E}\, Y_n = \mu$, and

   $\mathbf{E}(y_n - \mu) = \sigma_n^2$ satisfying $\sum_{n=1}^{\infty} (\log n)^2\sigma_n^2/n^2 < +\infty$, then $X_n \xrightarrow{as} \mu$

4. (Serfling) If the Y_n have $\mathbf{E}\, Y_n = \mu$, $\mathbf{E}(Y_n-\mu)^2 \equiv \sigma_n^2$, and

   $|\mathbf{E}(Y_n-\mu)(Y_m-\mu)| \leq \rho^{|n-m|}\sigma_n\sigma_m$, with $\sum_{i=1}^{\infty} \left(\frac{\log(i)}{i}\right)^2 \sigma_i^2 < +\infty$ and

   $\sum_{i=0}^{\infty} \rho_i^2 < +\infty$, then $X_n \xrightarrow{as} \mu$

Next consider Chebyshev's theorem. One has $\mathbf{E}(X_n-\mu)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_n^2$. If this approaches zero, then Chebyshev's inequality implies $X_n \xrightarrow{p} \mu$.

Finally, consider version 3 of the WLLN. Chebyshev's inequality states that $P(|X_n-\mu| < \varepsilon) < \mathbf{E}(X_n-\mu)^2/\varepsilon^2 \leq \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j\rho_{|i-j|}$. An application of the Cauchy-Schwartz inequality yields

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j\rho_{|i-j|} \leq \left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_j^2\rho_{|i-j|}^2\right)^{1/2}$$

$$\leq \left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\right)^{1/2}\left(\frac{M^2}{n^2}\sum_{j=1}^{n}\sigma_j^2\right)^{1/2} = M\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\right) \longrightarrow 0,$$

where $\sum_{i=1}^{n}\rho_{|i-j|}^2 \leq 2\sum_{k=0}^{\infty}\rho_k^2 \equiv M^2 < +\infty$.

3. Central Limit Theorems

4.1 Consider a sequence of random variables $Y_1,...,Y_n$ with zero means, and the associated sequence of normalized averages $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Y_i$. Central limit theorems (CLT) are concerned with conditions under which the Z_n converge in distribution to a normal random variable Z_o. Figure 5 lists several basic central limit theorems.

Figure 5. Central Limit Theorems for $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$

1. (Lindeberg-Levy) Y_n i.i.d., $\mathbf{E}Y_n = 0$, $\mathbf{E}Y_n^2 = \sigma^2 < +\infty$

   $\Longrightarrow Z_n \xrightarrow{d} Z_o \sim N(0,\sigma^2)$

2. (Lindeberg-Feller) Y_n independent, $\mathbf{E}Y_n = 0$, $\mathbf{E}Y_n^2 = \sigma_n^2$,

   $B_n^2 = \sum_{i=1}^{n} \sigma_i^2 \longrightarrow +\infty$, $\sigma_n^2/B_n^2 \to 0$, and for $\varepsilon > 0$,

   $$\frac{1}{B_n^2} \sum_{i=1}^{n} \int_{|y|>\varepsilon B_n} y^2 f_{Y_i}(y)dy \longrightarrow 0 \Longrightarrow Z_n \xrightarrow{d} Z_o \sim N(0, \lim_{n\to\infty} \frac{B_n^2}{n})$$

3. (Corollary to Lindeberg-Feller) $Z_n \xrightarrow{d} Z_o \sim N(0,\sigma_o^2)$ if the Y_n are independent, $\mathbf{E}Y_n = 0$, $\mathbf{E}Y_n^2 = \sigma_n^2$, $\frac{1}{n}\sum_{i=1}^{n} \sigma_n^2 \longrightarrow \sigma_o^2 < +\infty$, and <u>one</u> of the following conditions holds:

   (i) for some $q > 2$, $\sum_{i=1}^{n} \mathbf{E}|Y_i|^q / \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{q/2} \longrightarrow 0$

   (ii) for some $q > 2$, $\mathbf{E}|Y_n\sigma_n|^q$ is uniformly bounded, all n

   (iii) $\lim_{k\to\infty} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int_{|y|>k\sigma_i} y^2 f_{Y_i}(y)dy \to 0$ and $\sum_{i=1}^{n} \sigma_i^3/n^{3/2} \to 0$

4. (Ibragimov-Linnik) Y_n stationary and strongly mixing with mixing coefficients $\alpha(n)$, $\mathbf{E}Y_n = 0$, $\mathbf{E}Y_n^2 = \sigma^2$, $\mathbf{E}|Y_n|^q < +\infty$ and

   $\sum_{n=1}^{\infty} \alpha(n)^{1-2/q} < +\infty$ for some $q > 2 \Longrightarrow Z_n \xrightarrow{d} Z_o \sim N(0,\sigma^2 + 2\sum_{i=1}^{\infty} \mathbf{E}Y_1Y_{1+i})$

Generally, stronger conditions are needed for CLT than were required for laws of large numbers. The most straightforward results are obtained for *independent and identically distributed* (iid) random variables. When the variables are not independent, one must impose conditions that limit the degree of dependence in order to get a CLT. One such result is the Ibragimov-Linnik CLT in the table. A sequence Y_n is *stationary* if the vector $(Y_1,...,Y_m)$ and the vector $(Y_i,...,Y_{i+m})$ have the same joint distribution, for every m and i. An implication of stationarity is that the Y_i all have the same mean and the same variance (if they exist). Recall the definition of strong mixing from Chap. 3. Let $\mathfrak{I}_t^-$ be the product σ-field of events that use only information up through time t. Let $\mathfrak{I}_{t+n}^+$ denote the product σ-field of events that use only information from time t+n on. The Y_i are *strong mixing* if there exists a scalar $\alpha(n)$ satisfying $\lim_{n\to\infty} \alpha(n) = 0$ such that $|P(\mathbb{A}\cap\mathbb{B}) - P(\mathbb{A})P(\mathbb{B})| \le \alpha(n)$ for all $\mathbb{A} \in \mathfrak{I}_t^-$ and $\mathbb{B} \in \mathfrak{I}_{t+n}^+$; they are *uniform mixing* if $|P(\mathbb{A}|\mathbb{B}) - P(\mathbb{A})| \le \phi(n)$ for all $\mathbb{A} \in \mathfrak{I}_t^-$ and $\mathbb{B} \in \mathfrak{I}_{t+n}^+$ and $\lim_{n\to\infty} \phi(n) = 0$. They are *strict mixing* if $|P(\mathbb{A}\cap\mathbb{B}) - P(\mathbb{A})P(\mathbb{B})| \le \psi(n)P(\mathbb{A})P(\mathbb{B})$ and $\lim_{n\to\infty} \psi(n) = 0$. 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) If the Y_i are uniform mixing with $\mathbf{E}Y_i = 0$ and $\mathbf{E}Y_i^2 = \sigma_i^2 < +\infty$, then $|\mathbf{E}Y_iY_{i+n}| \le 2\sqrt{\phi(n)}\sigma_i\sigma_{i+n}$.

(3) If the Y_i are strong mixing with $\mathbf{E}Y_i = 0$ and $\mathbf{E}|Y_i|^d < +\infty$ for some $d > 2$, then $|\mathbf{E}Y_iY_{i+n}| \le 8\alpha(n)^{1-2/d}\sigma_i\sigma_{i+n}$.

(4) If there exists a sequence ρ_i with $\lim_{i\to\infty} \rho_i = 0$ such that $|\mathbf{E}(U-\mathbf{E}U)(W-\mathbf{E}W)| \le \rho_n\sqrt{(\mathbf{E}(U-\mathbf{E}U)^2)(\mathbf{E}(W-\mathbf{E}W)^2)}$ 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_i are strict mixing.

Only a proof of the Lindeberg-Levy central limit theorem (CLT) will be outlined here. 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 (Chap. 3).

Let $\psi(t)$ be the cf of Y_1. Then Z_n has cf $\psi(t/\sqrt{n})^n$. 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

$$\psi(t/\sqrt{n})^n = [1 + \frac{\psi''(\lambda t/\sqrt{n})t^2}{2n}]^n \to e^{(t^2/2)\lim \psi''(\lambda t/\sqrt{n})} = e^{-\sigma^2 t^2/2} .$$

Thus, the cf of Z_n converges for each t to the cf of $Z_o \sim N(0,\sigma^2)$.

4. Extensions of Limit Theorems

4.1. Limit theorems can be extended in several directions: (1) obtaining results for 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 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 $\lim_{n\to\infty} P(\sup_{\theta} |\frac{1}{n}\sum_{i\le n} Y_i(\theta,\cdot)| > \varepsilon) = 0$. 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.

4.2 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 independent, what are the limiting properties of X_n? The way arrays like a_{in} typically arise is that there are some weighting constants c_i, and either $a_{in} = c_i/\sum_{j=1}^{n} c_j$ or $a_{in} = c_i/\sqrt{\sum_{j=1}^{n} c_j^2}$. If $c_i = 1$ for all i, then $a_{in} = n^{-1}$ or $n^{-1/2}$, respectively, leading to the standard limit theorems.

Assume the Y_i are independently identically distributed with mean zero. *If the* a_{in} *satisfy* $\lim_{n\to\infty} \sum_{j=1}^{n} |a_{jn}| = 0$ *and* $\lim_{n\to\infty} \max_{j\le n} |a_{jn}| = 0$, *then* $X_n \xrightarrow{p} 0$. 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 $|\gamma(t)| \leq \sum_{i=1}^{n} |a_{in} t \cdot \zeta'(\lambda_{in} a_{in} t)| \leq |t| \cdot (\max_{i \leq n} |\zeta'(\lambda_{in} a_{in} t)|) \cdot \sum_{i=1}^{n} |a_{in}|$.

Then $\lim \sum_{i=1}^{n} |a_{in}| < \infty$ and $\lim (\max_{i \leq n} |a_{in}|) = 0$ imply $\gamma(t) \longrightarrow 0$ for each t, and hence X_n converges in distribution, hence in probability, to 0.

Next *assume the* Y_i *are independently identically distributed with mean zero and variance* $\sigma^2 < \infty$, $\lim_{n \to \infty} (\max_{i \leq n} |a_{in}|) = 0$, *and* $\lim_{n \to \infty} \sum_{i=1}^{n} a_{in}^2 = 1$. *Then* X_n *converges in distribution to a normal random variable with mean zero and variance* σ^2.

The proof of this proposition parallels the Lindeberg-Levy CLT proof. The second characteristic function of X_n now has the Taylor's expansion

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

The limit assumptions imply $\gamma(t) + (1/2)\sigma^2 t^2$ is bounded in magnitude by

$$\sum_{i=1}^{n} |\zeta''(\lambda_{in} a_{in} t) + \sigma^2| a_{in}^2 t^2/2 \leq \left(\sum_{i=1}^{n} a_{in}^2 t^2/2 \right) \cdot \max_{i \leq n} (|\zeta''(\lambda_{in} a_{in} t) + \sigma^2| \longrightarrow 0,$$

implying that for each t, $\gamma(t)$ converges to the characteristic function of a normal with mean 0 and variance σ^2.

4.3 Chebyshev bounds give an easy, but crude, bound on the probability in the tail of a density: for $\varepsilon > 0$, $\text{Prob}(Y > \varepsilon) \le \int_{\varepsilon}^{\infty} (|y|/\varepsilon)^{\rho} f_Y(y)dy \le \mathbf{E}_Y |Y|^{\rho}/\varepsilon^{\rho}$. For random variables with well behaved tails, sharper bounds can be found, and used in turn to get sharper limit theorems. First, suppose independent identically distributed random variables Y_i with zero mean and the bound $|Y_i| \le 1$, and $X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Chebyshev's inequality gives $P(X_n > \varepsilon) \le 1/n\varepsilon^2$, since $\text{var}(Y_i) \le 1$. However, a better inequality due to Hoeffding states that $\text{Prob}(X_n > \varepsilon) < e^{-n\varepsilon^2/2}$. This and similar bounds can be found in Pollard, Convergence of Stochastic Processes, and Shorak and Wellner, Empirical Processes. If the Y_i are not necessarily bounded, but have a proper moment generating function, one can get the inequality $P(X_n > \varepsilon) < e^{-\tau\varepsilon\sqrt{n} + \kappa}$, where τ and κ are positive constants determined by the distribution of Y_i.

To illustrate the use of the last inequality, note that

$$P(\sup_{i \ge n} |X_i| > \varepsilon) < \sum_{i=n}^{\infty} 2e^{-\tau\varepsilon\sqrt{i}+\kappa} \le \int_{n-1}^{\infty} 2e^{-\tau\varepsilon\sqrt{z}+\kappa}dz \le 4(1+\tau\varepsilon\sqrt{n-1})e^{-\tau\varepsilon\sqrt{n-1}+\kappa} \longrightarrow 0.$$

This implies $X_n \xrightarrow{as} 0$, a SLLN.

To show the inequality $P(X_n > \varepsilon) < e^{-\tau\varepsilon\sqrt{n}+\kappa}$, first note for any random variable X that $P(X > \varepsilon) \le \inf_{t>0} e^{-\varepsilon t} m_X(t)$, where m_X is the mgf of X. Second, if Y_i has mgf $m_Y(t) < \infty$ for $|t| < 2\tau$, then $m_Y(t) = 1 + m_Y''(\lambda t)t^2/2$ for some $|\lambda| < 1$, for each $|t| < 2\tau$, from the properties of the moment generating function. Define $M = \max_{|t|<\tau} m_Y''(t)$; then $m_Y(t) \le 1 + Mt^2/2$ for $|t| \le \tau$. The mgf of X_n then satisfies $m_X(t) \le (1 + Mt^2/2n^2)^n$ for $|t| \le \tau$. Hence,

$$P(X_n > \varepsilon) \le \inf_{0<t\le\tau} e^{-\varepsilon t}(1 + Mt^2/2n^2)^n.$$

Taking $t = \tau\sqrt{n}$ and using the inequality $(1+M\tau^2/2n)^n \leq e^{M\tau^2/2}$ gives the claimed result with $\kappa = M\tau^2/2$.

4.4 This section states a uniform SLLN for random functions on a subset Θ of a Euclidean space $\mathbb{R}^q$. Let $(S,\mathcal{F},P)$ 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 $(S,\mathcal{F},P)$. 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 few definitions are needed:

A measurable random function $Y(\theta,\cdot)$ is *separable* if there exists a countable dense subset Θ_0 of Θ and a set $A \in \mathcal{F}$ with $P(A) = 1$ such that for each $s \in A$ and $\theta \in \Theta$, there exists a sequence $\theta_i \in \Theta_0$ such that $\lim_{i\to\infty} Y(\theta_i,s) = Y(\theta,s)$.

A measurable random function $Y(\theta,\cdot)$ is *almost surely continuous* at $\theta_0 \in \Theta$ if for each $\varepsilon > 0$, there exist measurable events $A_k(\varepsilon,\theta_0)$ which contain all states of Nature s such that $\sup_{|\theta-\theta_0|\leq 1/k} |Y(\theta,s) - Y(\theta_0 s)| > \varepsilon$ and which converge monotonically as $k \to \infty$ to a set $A_*(\varepsilon,\theta_0)$ that has probability zero.

Finite-dimensional Euclidean spaces contain countable dense subsets (e.g., the points with rational coordinates), and any closed and bounded subset of a space with this property also contains a countable dense subset. Then, the main restriction imposed by separability when Θ is closed and bounded is that Y is regular enough so that its values everywhere are determined by its values on a countable dense subset. This is a much weaker condition than continuity, as we require only that some sequence

yield the limiting value, not all sequences. Furthermore, we can handle a countable number of isolated points of discontinuity simply by including them in Θ_o. Often in the theory of stochastic processes, separability is assumed by construction: a random function $Y'(\theta,s)$ which may not be separable is replaced by an "equivalent" version $Y(\theta,s) = \limsup_{\theta' \in \Theta_o \& \theta' \to \theta} Y'(\theta',s)$ for $\theta \notin \Theta_o$.

The condition of almost sure continuity allows the modulus of continuity to vary with state of Nature, so there is not necessarily a fixed neighborhood of θ_o on which the function does not vary by more than ε, independent of the state of Nature. For example, the function $Y(\theta,s) = \theta^s$ defined for $\theta \in [0,1]$ and s distributed uniformly on the unit interval is continuous at $\theta = 0$ for every s, but $A_k(\varepsilon,0) = (0,\frac{-\log \varepsilon}{\log k})$ has positive probability for all k. The exceptional sets $A_k(\varepsilon,\theta)$ can vary with θ, and there is no requirement that there be a set of states of Nature with probability one, or for that matter with positive probability, where $Y(\theta,s)$ is continuous for all θ. As a result, a function can be pointwise almost surely continuous, and still always have discontinuities. For example, assuming $\theta \in [0,1]$ and s distributed uniformly on the unit interval, and defining $Y(\theta,s) = 1$ if $\theta \geq s$ and $Y(\theta,s) = 0$ otherwise gives a function that always has a discontinuity, but is nevertheless almost surely continuous. For closed and bounded Θ, which always contains a countable dense set, almost sure continuity implies separability. The following result establishes a uniform SLLN for random functions that satisfy almost sure continuity.

Lemma. Assume $Y_i(\theta)$ *are independent identically distributed random functions with a finite mean* $\psi(\theta)$ *for* θ *in a closed bounded set* $\Theta \subseteq \mathbb{R}^q$. *Assume that for each* $\theta \in \Theta$, $Y_i(\cdot)$ *is almost surely continuous at* θ. *Assume there exists a positive*

envelope random variable Z *satisfying* $Z \geq \max_{\theta \in \Theta} |Y_i(\theta)|$ *and* **E** $Z < +\infty$. *Then*, $X_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i(\theta)$ *satisfies* $\sup_{\theta \in \Theta} |X_n(\theta) - \psi(\theta)| \xrightarrow{as} 0$.

Proof: We follow an argument of Tauchen (1985). Let (S,$\mathcal{F}$,P) be the underlying probability space, and write the random function $Y_i(\theta,s)$ to make its dependence on the underlying state of Nature explicit. We have $\psi(\theta) = \int_S Y(\theta,s)P(ds)$. Define $u(\theta_o,s,k) = \sup_{|\theta-\theta_o| \leq 1/k} |Y(\theta,s) - Y(\theta_o,s)|$. Let $\varepsilon > 0$ be given. Let $A_k(\varepsilon/2,\theta_o)$ 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 $A_k(\varepsilon/2,\theta_o)$ is less than $\varepsilon/4 \cdot$(**E** Z). Then,

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

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

Let $B(\theta_o)$ 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 $B(\theta_j)$ with centers at points θ_j for j = 1,...,J. Let $\mu_j = \mathbf{E}u(\theta_j,\cdot,k(\varepsilon/2,\theta_j)) \leq \varepsilon$. For $\theta \in B(\theta_j)$, $|\psi(\theta) - \psi(\theta_j)| \leq \mu_j \leq \varepsilon$. Then

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

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

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

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

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

and

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

With probability at least $1 - \varepsilon/J$, $\sup_{n\geq n_j} \sup_{\theta\in B(\theta_j)} \left|X_n(\theta) - \psi(\theta)\right| \leq 4\varepsilon$. Then, with probability at least $1 - \varepsilon$, $\sup_{n\geq n_o} \sup_{\theta\in\Theta} \left|X_n(\theta) - \psi(\theta)\right| \leq 4\varepsilon$, where $n_o = \max(n_j)$. □

The construction in the proof of the lemma 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 the how rapidly the size of the approximating family can grow with sample size for a given ε, rather than continuity per se; see Pollard (1984).

4.5 Central limit theorems for sums of non-independent random variables are critical for time-series analysis. The result of Ibragimov and Linnik given in Figure 5 is a typical and fairly strong representative of the variety of results that are

available. To illustrate its use, consider a structure that appears frequently in econometric time series analysis. Suppose Z_i are i.i.d. random variables with three finite moments and a zero mean, and that Y_i are weighted averages of current and past Z_i; i.e., $Y_i = \sum_{j=0}^{\infty} \beta_j Z_{j-i}$, a moving average process. If $\sum_{j=0}^{\infty} e^{\lambda j} \cdot \beta_j^2 < +\infty$ for some $\lambda > 0$, so that the moving average weights eventually decay at an exponential rate, then the Y_i are stationary with three finite moments and mean zero, and one strong mixing. This is sufficient for application of the Ibragimov-Linnik CLT.

In general, CLT are much messier to state and prove than the independent case. One powerful tool is martingale theory, which exploits the property that when sums of innovations that have conditional mean zero satisfy some boundedness properties, then these sums behave much like sums of independent innovations. Martingale methods will be presented in a future version of this book.