BOL. SOC. BRAS. MAT. 6(1975), 25 – 38

Some Applications of the Law of Large Numbers

Jerome A. Goldstein

1. Introduction

This expository paper is aimed largely at analysts who know little or no probability theory. By presenting some surprising, nontrivial applications of an elementary probability limit theorem (a variant of the weak law of large numbers), we hope to persuade these analysts that it is worthwhile to study probability theory (if for no other reason) to get a new perspective from which to review other parts of analysis. The basic limit theorem presented below is at the level of easy advanced calculus, whereas the consequences of the limit theorem are usually thought to be more difficult.

Much of what we present is contained in Chapter 7 of Feller's book [2]. Some of the results presented here are new, but we are more interested in publicizing Feller's clever techniques than in extending his results.

The plan of the paper is a follows. Section 2 contains Feller's elementary limit theorem. This section is written in nonprobabilistic language to indicate that one need know *no* prabability theory to understand the theorem and its easy proof. (Section 2 will look awkward to a probabilist.) Section 3 contains the result of Section 2 translated into probabilistic terminology. Section 4 contains applications, namely: the Weierstrass approximation theorem, a difference quotient version of Taylor's theorem, inversion of Laplace transforms, and the moment problem. We pinpoint exactly where probabilistic thinking comes into play in the applications. Section 5 deals with rates of convergence. Section 6 contains a multidimensional generalization and an application. Finally, Section 7 contains E. Borel's elegant application of the strong law of large numbers to show that almost all real numbers are normal.

2. Feller's limit theorem

A *distribution function* is a nondecreasing function $F: \mathbb{R} =]-\infty, \infty[\longrightarrow [0, 1]$ satisfying $\lim\limits_{x\to-\infty} F(x) = 0$, $\lim\limits_{x\to\infty} F(x) = 1$. (One could also re-

Recebido em 23 de Maio de 1975.

quire that F is continuous from the right, but this is irrelevant for our purposes.) We write $F \in \mathscr{D}\mathscr{F}$ to indicate that F is a distribution function. We shall be concerned with integrals of the form $\int_b^a g(x)\,dF(x)$ where $-\infty \le a < b \le \infty$, $F \in \mathscr{D}\mathscr{F}$, and g is continuous. These Riemann-Stieltjes integrals are limits of sums of the form $\sum_{i=1}^{n} g(y_i)\,(F(x_i) - F(x_{i-1}))$. We won't bother to explain this notation as we assume the reader is familiar with such integrals. Actually, there are only two cases which interest us: (i) F has a piecewise continuous derivative f, so that

$$\int_a^b g(x)\,dF(x) = \int_a^b g(x)\,f(x)\,dx;$$

(ii) F is constant in an interval J except for jumps of size a_j at x_j, so that

$$\int_J g(x)\,dF(x) = \Sigma_j g(x_j)a_j\,;$$

the latter expression is either a finite sum or infinite series. Note that $\int_{-\infty}^{\infty} dF(x) = 1$ for $F \in \mathscr{D}\mathscr{F}$. We shall write $\int$ for $\int_{-\infty}^{\infty}$.

Lemma 1 (Chebyshev's inequality). *If $F \in \mathscr{D}\mathscr{F}$, $\int x^2 dF(x) < \infty$, and $\varepsilon > 0$, then for any real number μ,*

$$\int_{|x-\mu| \le \varepsilon} dF(x) \le \varepsilon^{-2} \int (x-\mu)^2\,dF(x).$$

Proof. $F \in \mathscr{D}\mathscr{F}$ and $\int x^2\,dF(x) < \infty$ implies $\int (x-\mu)^2\,dF(x) < \infty$. We have

$$\begin{aligned}\int (x-\mu)^2\,dF(x) &= \int_{|x-\mu|<\varepsilon} (x-\mu)^2\,dF(x) + \int_{|x-\mu|\le\varepsilon} (x-\mu)^2\,dF(x)\\ &\ge \int_{|x-\mu|\ge\varepsilon} (x-\mu)^2\,dF(x) \ge \varepsilon^2 \int_{|x-\mu|\ge\varepsilon} dF(x).\end{aligned}$$

Theorem 1 (Feller). *Let θ be a real parameter varying in a compact (i.e. closed and bounded) interval J. Let $\{F_n(\cdot\,;\theta)\colon \theta \in J\} \subset \mathscr{D}\mathscr{F}$ satisfy for $n = 1, 2, \ldots,$*

(i) $\int x\,dF_n(x;\theta) = \theta$ *for all* n, θ,

(ii) $\int (x-\theta)^2\,dF_n(x;\theta) \equiv \sigma_n^2(\theta) \longrightarrow 0$ *as* $n \longrightarrow \infty$,

uniformly for $\theta \in J$.

Then for every bounded continuous function f on $\mathbb{R}$,

$$\int f(x)\,dF_n(x;\theta) \longrightarrow f(\theta) \qquad \textit{as} \qquad n \longrightarrow \infty,$$

uniformly for $\theta \in J$.

Proof. Let $\varepsilon > 0$ be given. Choose $M \varepsilon\,]0, \infty[$ such that $|f(x)| \le M$ for all $x \in R$. For $\theta \in J$,

$$\Big|\int f(x)\,dF_n(x;\theta) - f(\theta)\Big| = \Big|\int (f(x) - f(\theta))\,dF_n(x;\theta)\Big|$$

$$\le \int_{|x-\theta|<\delta} |f(x) - f(\theta)|\,dF_n(x;\theta) + \int_{|x-\theta|\ge\delta} |f(x) - f(\theta)|\,dF_n(x;\theta) \equiv I_1 + I_2 \tag{1}$$

for each $\delta > 0$. Since f is uniformly continuous on compact intervals we choose $\delta > 0$ such that $|f(x) - f(\theta)| < \varepsilon/2$ for $\theta \in J$ and $|x - \theta| < \delta$. Fix this δ. Now choose $N = N_\varepsilon$ such that

$$n \ge N \quad \text{implies} \quad \sigma_n^2(\theta) \le \delta^2 \varepsilon/4M. \tag{2}$$

Then

$$I_1 < \tfrac{\varepsilon}{2} \int_{|x-\theta|<\delta} dF_n(x;\theta) \le \varepsilon/2 \tag{3}$$

and

$$\begin{aligned} I_2 &\le 2M \int_{|x-\theta|\ge\delta} dF_n(x;\theta)\\ &\le 2M\sigma_n^2(\theta)/\delta^2 \quad \text{by Chebyshev's inequality}\\ &< \varepsilon/2 \quad \text{if} \quad n \ge N. \qquad (4)\end{aligned}$$

Combining (1), (3) and (4) completes the proof.

3. The weak law of large Numbers

Let (Ω, Σ, P) be a probability space (i.e. a measure space such that $P(\Omega) = 1$). Let X be a *random variable* (i.e. a real-valued Σ- measurable function on Ω). The *distribution function* F_X of X is defined by: $F_X(x) = P\{\omega \in \Omega\colon X(\omega) \le x\}$ $(= P\{X \le x\}$ for short). Clearly $F_X \in \mathscr{D}\mathscr{F}$. If $g\colon \mathbb{R} \longrightarrow \mathbb{R}$ is continuous (or more generally, Borel measurable), then the *expectation* $E(g(X))$ of the random variable $g(X)$ is defined to be the Lebesgue integral $\int_\Omega g(X)\,dP$. In practice, this is always computed with the aid of the *law of the unconscious statistician*:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)\,dF_X(x). \tag{5}$$

(More precisely the left side exists if and only if the right side exists, in which case equality holds.) In undergraduate courses (5) is often used as the definition of $E(g(X))$. Two choices of g are especially popular: $E(X)$ is called the *mean* of X (take $g(x) \equiv x$), and $Var(X) = E((X - E(X))^2)$ is called the *variance* of X (take $g(x) = (x-\mu)^2$ where $\mu = E(X)$). One final bit of notation: $\chi_A(\omega)$ denotes the random variable which is 1 if $\omega \in A$ and 0 if $\omega \in \Omega \setminus A$, where $A \in \Sigma$.

Lemma 1' (Chebyshev's inequality). *Let X be a random variable having a finite variance. Then for any $\varepsilon > 0$,*

$$P\{|X - E(X)| \ge \varepsilon\} \le \varepsilon^{-2}\,Var(X).$$

Proof.

$$\begin{aligned} Var(X) &= E\{(X - E(X))^2 \chi_{\{|X-E(X)| \geq \varepsilon\}}\} + E\{(X - E(X))^2 \chi_{\{|X-E(X)| < \varepsilon\}}\} \\ &\geq E\{\varepsilon^2 \chi_{\{|X-E(X)| \geq \varepsilon\}}\} + 0 = \varepsilon^2 P\{|X - E(X)| \geq \varepsilon\}. \end{aligned}$$

The above is evidently a repetition of the statement and proof of Lemma 1.

The random variables $X_1, X_2, \ldots$ are *independent* if for each n and each $A_i \in \Sigma$,

$$P\left\{\bigcap_{i=1}^{n} (X_i \in A_i)\right\} = \prod_{i=1}^{n} P\{X_i \in A_i\}.$$

They are *identically distributed* if they have a common distribution function, i.e. if $F(x) = P\{X_i \leq x\}$ does not depend on i.

Theorem 0. (Weak Law of Large Numbers). *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables having a finite mean $\mu = E(X_i)$ and variance. Then for each $\varepsilon > 0$,*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right\} \longrightarrow 0$$

as $n \longrightarrow \infty$.

Thus the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability (or weakly) to the mean μ as $n \longrightarrow \infty$.

Proof. Let $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $E(Y_n) = \mu$, and using the independence of $X_1, \ldots, X_n$ it is easy to show that $Var(Y_n) = n^{-1}\, Var(X_1)$. Chebyshev's inequality yields

$$P\{|Y - \mu| \geq \varepsilon\} \leq \varepsilon^{-2}\, Var(Y_n) = \varepsilon^{-2}\, Var(X_1)/n \longrightarrow 0$$

as $n \longrightarrow \infty$.

The following result is a special case of Feller's Theorem (Theorem 1).

Theorem 1'. *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with mean $E(X_i) = \theta$, which we consider as a real variable varying over a compact interval J. Suppose also that $K = \sup_{\theta \in J} Var(X_i) < \infty$. Then for every bounded continuous function f on $\mathbb{R}$,*

$$E\left\{f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\right\} \longrightarrow f(\theta)$$

as $n \longrightarrow \infty$, uniformly for $\theta \in J$.

Let $F_n(\cdot\,;\theta)$ be the distribution function of $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\int x\,dF_n(x;\theta) = \theta$ for all θ, n; and $\int (x - \theta)^2\, dF_n(x;\theta) = Var(Y_n) = n^{-1}\, Var(X_j) \leq K/n \longrightarrow 0$ as $n \longrightarrow \infty$, uniformly for $\theta \in J$. Since

$$E\left\{f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\right\} = \int f(x)\, dF_n(x;\theta),$$

we see that indeed Theorem 1' is a special case of Theorem 1. There is no need to repeat the proof.

Note that Theorem 0 is a special case of Theorem 1' (take J $= \{\theta\}$, $f(x) \equiv x$). But they are not really that different; if one stares at them long enough one becomes convinced that they are almost the same! This is why we call Theorem 1 a variant of the weak law of large numbers.

4. Applications

Specific choices of $F_n(\cdot\,;\theta)$ in Theorem 1 lead to interesting applications. Question: How does one choose $F_n(\cdot\,;\theta)$? Answer: Choose the basic distribution functions arising in probability theory. This is the point where probability theory enters in a crucial way.

In the examples (except Examples 3) that follow, $F_n(\cdot\,;\theta)$ will be the distribution function of $\frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_1, X_2, \ldots$ are independent, identically distributed random variables with mean θ (cf. Theorem 1').

Example 1. Let $X_i = 1$ with probability θ, $X_i = 0$ with probability $1 - \theta$, $\theta \in J = [0, 1]$. Then $\sum_{i=1}^{n} X_i$ has the binomial distribution with parameters n and θ. Thus $F_n(\cdot\,;\theta)$ is a step function whose jumps occur at k/n $(k = 0, 1, \ldots, n)$, the magnitude of the jump at k/n being $\binom{n}{k}\theta^k(1 - \theta)^{n-k}$ where $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ is the binomial coefficient. A simple calculation shows that

$$\int x\,dF_n(x;\theta) = \theta, \quad \int (x - \theta)^2\, dF_n(x;\theta) = \frac{\theta(1 - \theta)}{n} \leq \frac{1}{4n} \longrightarrow 0$$

as $n \longrightarrow \infty$ uniformly in $\theta \in [0, 1]$. Consequently for any continuous function f on $[0, 1]$,

$$f(\theta) \longleftarrow \int f(x)\, dF_n(x;\theta) = \sum_{k=0}^{n} f\left(\tfrac{k}{n}\right)\binom{n}{k}\theta^k(1 - \theta)^{n-k} \tag{6}$$

as $n \longrightarrow \infty$, the convergence being uniform in $\theta \in [0, 1]$. Thus any continuous function on $[0, 1]$ can be uniformly approximated by polynomials. This is the Weierstrass approximation theorem. Moreover, (6) contains an explicit recipe for computing the polynomials which approximate f; the polynomials on the right hand side of (6) are called *Bernstein polynomials*. The idea of this proof goes back to S. Bernstein.

Example 2. Let X_i have a Poisson distribution with parameter θ. Then $\sum_{i=1}^{n} X_i$ has a Poisson distribution with parameter $n\theta$. That is, $F_n(\cdot\,; \theta)$ is a step function having a jump of size $e^{-\theta n}(\theta n)^k/k!$ at k/n, $k = 0, 1, 2, \ldots$. Let J be the interval $[0, M_1]$, where $M_1 > 0$ is fixed but arbitrary. One easily shows that $\int x dF_n(x; \theta) = \theta$, $\int (x - \theta)^2\, dF_n(x; \theta) = \dfrac{\theta}{n} \leq \dfrac{M_1}{n} \longrightarrow 0$ as $n \longrightarrow \infty$, uniformly for $\theta \in J$. Hence for every bounded continuous function f on $[0, \infty[$,

$$f(\theta) \leftarrow \int f(x)\, dF_n(x; \theta) = \sum_{k=0}^{\infty} e^{-\lambda n} f\left(\frac{k}{n}\right) \frac{(\lambda n)^k}{k!}$$

$$= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} f(kh) \left(\frac{\lambda}{h}\right)^{j+h} \frac{1}{k!} \quad \text{where} \quad h = \frac{1}{n}$$

$$(7) \qquad = \sum_{m=0}^{\infty} \left(\frac{\lambda}{h}\right)^m \frac{1}{m!} \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} f(kh),$$

the last equality being obtained by setting $m = j + k$. Define the difference quotient operator Δ_h by

$$(\Delta_h f)(x) = \frac{f(x + h) - f(x)}{h}.$$

$\Delta_h{}^n$ is the *nth* power of Δ_h; thus $\Delta_h{}^n f = \Delta_h(\Delta_h{}^{n-1} f)$. A simple induction argument yields

$$\Delta_h{}^N f(x) = \frac{1}{h^N} \sum_{k=0}^{N} (-1)^{N-h} \binom{N}{h} f(x + kh).$$

Plugging this into (7) yields

$$f(\theta) \longleftarrow \sum_{m=0}^{\infty} \frac{\theta^m}{m!} \Delta_h{}^m f(0)$$

as $h \longrightarrow 0$, uniformly for $\theta \in J$. Changing variables we conclude that if f is any bounded continuous function on $\mathbb{R}$ and if $x \in \mathbb{R}$,

$$\sum_{m=0}^{\infty} \frac{y^m}{m!} \Delta_h{}^m f(x) \longrightarrow f(x + y)$$

as $h \longrightarrow 0$, uniformly for y in any bounded interval. This is a generalized Taylor series expansion of f with difference quotients replacing derivatives; it was first obtained by E. Hille. Note that f need not be differentiable here.

Example 3. Let g_n be the density function of a Beta distribution with parameters $n\nu_1$, $n\nu > 0$. That is,

$$g_n(x) = \frac{\Gamma(n\nu_1 + n\nu)}{\Gamma(n\nu_1)\,\Gamma(n\nu)} x^{n\nu_1 - 1} (1 - x)^{n\nu - 1} \quad \text{if} \quad 0 < x < 1$$

and $g(x) = 0$ otherwise; here Γ denotes the gamma function. Then calculus shows that

$$\int x g_n(x)\, dx = \frac{\nu_1}{\nu_1 + \nu}, \quad \int \left(x - \frac{\nu_1}{\nu_1 + \nu}\right)^2 g_n(x)\, dx = \frac{\nu_1}{n(\nu_1 + \nu)(\nu_1 + \nu + \frac{1}{n})}.$$

Let ν be a fixed positive rational, let $\theta = \dfrac{\nu_1}{\nu_1 + \nu}$, and let $J = [0, 1]$. If $F_n(x; \theta) = \int_0^x g_n(t)\, dt$, then

$$\int x dF_n(x; \theta) = \theta, \; \int (x - \theta)^2\, dF_n(x; \theta) \leq \frac{\theta\nu}{n(\nu_1 + \nu)^2} \leq \frac{\theta}{n\nu} \longrightarrow 0$$

as $n \longrightarrow \infty$, uniformly for $\theta \in J$. (Recall that ν is a fixed positive rational.) Consequently if f is a continuous function on $[0, 1]$, then as $n \longrightarrow \infty$,

$$(8) \quad f(\theta) \leftarrow \int_0^1 f(x) dF_n(x; \theta) = \frac{\Gamma\left(\dfrac{n\nu}{1 - \theta}\right)}{\Gamma\left(\dfrac{n\nu\theta}{1 - \theta}\right)\Gamma(n\nu)} \int_0^1 f(x) x^{\frac{n\nu\theta}{1 - \theta}} (1 - x)^{n\nu - 1}\, dx.$$

The convergence is uniform for $\theta \in J$. When $n\nu$ is an integer we can expand $(1 - x)^{n\nu - 1}$ as polynomial in x. Let θ be a rational number in $[0, 1]$ (so that $\dfrac{\theta}{1-\theta}$ is rational). Now let $n \longrightarrow \infty$ through a subsequence S of integers so that $n\nu - 1$ and $\dfrac{n\nu\theta}{1-\theta}$ are positive integers for each $n \in S$. Then (8) implies that we can recover $f(\theta)$ from the *moments* $\int_0^1 f(x)\, x^m\, dx$, $m = n_0$, $n_0 + 1, \ldots$, where n_0 is an arbitrary integer. Since θ is an arbitrary rational in $[0, 1]$, it follows that $f(x)$, $0 \leq x \leq 1$ can be recovered from $\int_0^1 f(x)\, x^m\, dx$, $m = n_0, n_0 + 1, \ldots$, for any continuous function f on $[0, 1]$ and we have exhibited a specific algorithm for doing so. For a different approach to a slightly different moment problem based on Theorem 1, see Feller [2, pp. 224-227].

Example 4. Other distribution lead to other results. Discrete distribution lead to various theorems concerning approximation by rational functions. The normal distribution (X_i has normal distribution with mean θ and variance σ^2,

$$F_n(x; \theta) = (n/2\pi\sigma^2)^{1/2} \int_{-\infty}^{x} e^{-n(t-\theta)^2/2\sigma^2}\, dt, \qquad \theta \in \mathbb{R},$$

σ a fixed positive number) leads immediately to the result.

$$\left(\frac{n}{2\pi\sigma^2}\right)^{1/2} \int f(x)\, e^{-n(x-\theta)^2/2\sigma^2}\, dx \longrightarrow f(\theta)$$

for each bounded continuous function f on $\mathbb{R}$, uniformly for θ in bounded intervals. This result is used in the study of initial value problem for the one dimensional heat equation.

Example 5. The final result we mention is the use of the gamma distribution. (Here we let X_i have gamma distribution with parameters α, β and use Theorem 1′.) Let

$$f_n(x) = \frac{n\alpha}{\Gamma(n\beta)} (n\alpha x)^{n\beta - 1}\, e^{-n\alpha x}$$

for $x > 0$ and $f_n(x) = 0$ for $x \leq 0$. Here $\alpha > 0$, $\beta > 0$. We let $\theta = \alpha/\beta$ and regard β as being fixed and $\theta \in J = [0, M_1]$, M_1 a fixed but arbitrary positive integer. If

$$F_n(x; \theta) = \int_0^x f_n(t)\, dt,$$

then $\int x dF_n(x; \theta) = \theta$, $\int (x - \theta)^2\, dF_n(x; \theta) = \theta/n\beta$. Hence for any bounded continuous function f on $[0, \infty[$,

$$\text{(9)} \qquad f(\theta) \longleftarrow \int f(x)\, dF_n(x; \theta) = \frac{\theta^{n\beta}\, (n\beta)^{n\beta}}{\Gamma(n\beta)} \int_0^\infty f(x)\, x^{n\beta - 1}\, e^{-nn\beta\theta x}\, dx$$

as $n \longrightarrow \infty$, uniformly for θ in bounded intervals of $[0, \infty[$. Let

$$\varphi(\lambda) = \int_0^\infty e^{-\lambda x}\, f(x)\, dx \qquad (\lambda > 0)$$

be the Laplace transform of f. Then φ is infinitely differentiable and

$$(d^n/d\lambda^n)\, \varphi(\lambda) = \int_0^\infty e^{-\lambda x}\, x^n\, f(x)\, dx, \qquad \lambda > 0.$$

Thus if we know the Laplace transform φ of f and if β is an integer, say $\beta = 1$, the right hand side of (9) is a constant times $(d^{n-1}/d\lambda^{n-1})\, \varphi(\lambda)$ evaluated at $\lambda = n\theta$. Thus (9) is an inversion formula for Laplace transforms.

5. Rates of Convergence

Suppose that we can approximate a continuous function f on an interval J uniformly by "nice" functions f_n, so that given $\varepsilon > 0$ there exists an N_ε such that

$$\sup_{\theta \in J} |f(\theta) - f_n(\theta)| \leq \varepsilon$$

for all $n \geq N_\varepsilon$. We know that N_ε exists; the problem we consider now is to find N_ε explicitly as a function of ε. This problem of the rate of convergence of approximating functions is important in many contexts, for example, in numerical computations. The next theorem shows how fast convergence takes place in Theorem 1 when the function being approximated is Lipschitzian.

Theorem 2. *Let $\{F_n(\cdot\,; \theta);\ \theta \in J\}$ be as in Theorem 1, and let the hypotheses of Theorem 1 hold. Let f be a bounded function on $\mathbb{R}$ satisfying a uniform Lipschitz condition, i.e. suppose there are constants M, L such that*

$$|f(x)| \leq M, \qquad |f(x) - f(y)| \leq L\,|x - y|$$

for all x, $y \in \mathbb{R}$. Then, given $\varepsilon > 0$,

$$|f(\theta) - \int f(x)\, dF_n(x; \theta)| \leq \varepsilon$$

for all $\theta \in J$ whenever $n \geq N_\varepsilon$, where N_ε is chosen so that

$$n \geq N_\varepsilon \quad \textit{implies} \quad \sigma_n^2(\theta) \leq \varepsilon^3/16L^2M.$$

Proof. As in the proof of Theorem 1, for any $\delta > 0$,

$$|f(\theta) - \int f(x)\, dF_n(x; \theta)| \leq I_1 + I_2\,;$$

$$I_1 = \int_{|x-\theta| < \delta} |f(x) - f(\theta)|\, dF_n(x; \theta) \leq L\delta \int_{|x-\theta| < \delta} dF_n(x; \theta) \leq L\delta;$$

$$I_2 = \int_{|x-\theta| \geq \delta} |f(x) - f(\theta)|\, dF_n(x; \theta) \leq 2M\sigma_n^2(\theta)/\delta^2.$$

Choose $\delta = \varepsilon/2L$. Then $I_1 \leq \varepsilon/2$, and $I_2 \leq \varepsilon/2$ if $2M\sigma_n^2(\theta)/\delta^2 = 8L^2M\sigma_n^2(\theta)/\varepsilon^2 \leq$ $\leq \varepsilon/2$, i.e. if $\sigma_n^2(\theta) \leq \varepsilon^3/16L^2M$. The theorem is proved.

In Example 1,

$$\sigma_n^2(\theta) = \frac{\theta(1 - \theta)}{n} \leq \frac{1}{4n} \leq \frac{\varepsilon^3}{16L^2M}$$

for $n \geq N_\varepsilon$ if $N_\varepsilon = \{4L^2M/\varepsilon^3\}$ where $\{x\}$ denotes the least integer $\geq x$. In particular, the error after n terms satisfies

$$\sup_{\theta \in [0,1]} |f(\theta) - B_n(\theta; f)| = 0(1/n^{1/3})$$

for every continuously differentiable function f on $[0,1]$, where $B_n(\theta; f)$ is the Bernstein polynomial defined by (6).

Similarly in Example 2 we get $N_\varepsilon = \{16L^2MM_1/\varepsilon^3\}$, so again the error after n terms is $0(n^{-1/3})$. The error estimate $0(n^{-1/3})$ is also valid for the other examples.

6. The multidimensional case

Let $\mathbb{R}^m$ denote m-dimensional Euclidean space. A *random vector* $X = (X_1, \ldots, X_m)$ is an m-tuple of random variables on a probability space (Ω, Σ, P). The *distribution function* of X is the function

$$F_X(x) = P\{X_1 \leq x_1, \ldots, X_m \leq x_m\}$$

defined for $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$. $F\colon \mathbb{R}^m \longrightarrow [0,1]$; $F(x_1, \ldots, x_m) \longrightarrow 0$ as $x_i \longrightarrow -\infty$ for $i = 1, \ldots, m$ and the other variables fixed; $F(x_1, \ldots, x_m) \longrightarrow 1$ as $x_1 \longrightarrow \infty, \ldots$, and $x_m \longrightarrow \infty$; and F satisfies a monotonicity property which we state for $m = 2$: if $x \leq x'$, $y \leq y'$,

$$F(x', y') - F(x, y') - F(x', y) + F(x, y) \geq 0.$$

This says that $P\{(X_1, X_2) \in \,]x, x'] \times \,]y, y']\} \geq 0$. The monotonicity condition for general m is similar but messier. These conditions enable us to distribution functions on $\mathbb{R}^m$ without reference to random variables, but for the sake of simplicity we won't.

The next result is an m-dimensional generalization of Theorem 1.

Theorem 3. *Let J be a compact set in $\mathbb{R}^m$ and let $\theta = (\theta_1, \ldots, \theta_m)$ be a parameter varying in J. Let $X^n(\theta) = (X_1^n(\theta), \ldots, X_m^n(\theta))$ be a random vector having distribution function $F_n(\cdot; \theta) = F_{X^{n}(\theta)}(\cdot)$ on $\mathbb{R}^m$. Assume*

(i) *$E(X_i^n(\theta)) = \theta_i$ for all n, i, θ,*

(ii) *$Var(X_i^n(\theta)) \equiv \sigma_{n,i}^2(\theta) \longrightarrow 0$ as $n \longrightarrow \infty$, uniformly for $\theta \in J$. Then for any bounded continuous function f on $\mathbb{R}^m$,*

$$\int_{\mathbb{R}^m} f(x)\, dF_n(x; \theta) \longrightarrow f(\theta)$$

as $n \longrightarrow \infty$, uniformly for $\theta \in J$.

Proof. Define $|x| = \left(\sum_{i=1}^{m} x_i^2\right)^{1/2}$ for $x \in \mathbb{R}^m$. Let the hypotheses of the theorem hold and choose M such that $|f(x)| \leq M$ for all $x \in \mathbb{R}^m$. Let $\varepsilon > 0$ be given.

$$|f(\theta) - \int_{\mathbb{R}^m} f(x)\, dF_n(x; \theta)| \leq I_1 + I_2$$

where

$$I_1 = \int_{|x-\theta| < \delta} |f(^\theta) - f(x)|\, dF_n(x; \theta),$$
$$I_2 = \int_{|x-\theta| \geq \delta} |f(\theta) - f(x)|\, dF_n(x; \theta).$$

By uniform continuity, choose (and fix) $\delta > 0$ such that $\theta \in J$, $|x - \theta| < \delta$ implies $|f(x) - f(\theta)| \leq \varepsilon/2$. Then $I_1 \leq \varepsilon/2$ follows trivially, Next,

$$I_2 \leq 2M \int_{|x-\theta| \geq \delta} dF_n(x; \theta) = 2MP\{|X^n(\theta) - \theta| \geq \delta\}$$
$$= 2M\, P\left\{\sum_{i=1}^{m} |X_i^n(\theta) - \theta_i|^2 \geq \delta^2\right\} \leq 2M \sum_{i=1}^{m} P\{|X_i^n(\theta) - \theta_i|^2 \geq \delta^2/m\}$$
$$\leq 2Mm\delta^{-2} \sum_{i=1}^{m} \sigma_{n,i}^2(\theta) \quad \text{by Chebyshev's inequality}$$
$$\leq \varepsilon/2$$

if $n < N_\varepsilon$, where N_ε is chosen (by (ii)) so that $\sum_{i=1}^{m} \sigma_{n,i}^2(\theta) \leq \varepsilon\delta^2/4Mm$ for all $\theta \in J$ and $n \geq N_\varepsilon$. Then $I_1 + I_2 \leq \varepsilon$, and we are done.

This proof contains a rate of convergence estimate for f satisfying a uniform Lipschitz condition. Thus Theorem 2 generalizes to the m-dimensional case too.

Example 6. Let $X^n(\theta) = (X_1^n(\theta), \ldots, X_m^n(\theta))$ have multinomial distribution with parameters n and $\tilde{\theta}$ where $\tilde{\theta} = (\theta, \theta_{m+1})$, $\theta \in J = \{\theta \in \mathbb{R}^m\colon \theta_i \geq 0, \sum_{i=1}^{m} \theta_i \leq 1\}$, $\sum_{i=1}^{m+1} \tilde{\theta}_i = \sum_{i=1}^{m+1} \theta_i = 1$. Thus if $x = (x_1, \ldots, x_m)$ is an m-tuple of nonnegative integers with $\sum_{i=1}^{m} x_i \leq n$, then $X^n(\theta) = x$ with probability

$$\frac{n!}{x_1! \ldots x_m! \left(n - \sum_{i=1}^{m} x_i\right)!} \theta_1{}^{x_1} \ldots \theta_m x^m \left(1 - \sum_{i=1}^{m} \theta_i\right)$$

Let $F_n(\cdot; \theta)$ be the distribution function of $n^{-1}X^n(\theta)$. Then $F_n(\cdot; \theta)$ satisfies the hypotheses of Theorem 3 with

$$\sigma_{n,i}^2(\theta) = \frac{\theta_i(1 - \theta_i)}{n} \leq \frac{1}{4n}.$$

Hence for any continuous function f on the simplex J, $f(\theta) \longleftarrow \int_{\mathbb{R}^m} f(x)\,dF_n(x;\theta) =$

$$= \Sigma f\left(\frac{x_1}{n},\ldots,\frac{x_m}{n}\right)\frac{n!\,\theta_1{}^{x_1}\ldots\theta_m{}^{x_m}\left(1-\sum_{i=1}^{m}\theta_i\right)}{x_1!\ldots x_m!\left(n-\sum_{i=1}^{m}x_i\right)!}$$

as $n \longrightarrow \infty$, uniformly for $\theta \in J$, where the summation is over the set of $x = (x_1, \ldots, x_m)$ with x_i a nonnegative integer and $\sum_{i=1}^{m} x_i \leq n$. This is the multivariate Wierstrass approximation theorem with multivariate Bernstein polynomials. Once again, if $f \in C^1(J)$, then the error estimate is

$$|f(\theta) - \int_{\mathbb{R}^m} f(x)\,dF_n(x;\theta)| = 0(n^{-1/3}),$$

uniformly for $\theta \in J$.

By Examining the Dirichlet distribution (a multidimensional generalization of the Beta distribution), once can solve a higher dimensional moment problem. We omit the calculations, which are similar to those in Example 3.

7. The strong law and normal numbers

The strong law of large numbers is the weak law with weak convergence (i.e. convergence in probability) replace by strong convergence (i.e. convergence almost everywhere).

Theorem 4. (Kolmogorov's Strong law of large Numbers) *Let X_1, $X_2, \ldots$ be independent identically distributed random variables having a finite mean $\mu = E(X_i)$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu$$

as $n \longrightarrow \infty$ almost everywhere, i.e.

$$P\left\{\omega \in \Omega : \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \longrightarrow \mu \quad as \quad n \longrightarrow \infty\right\} = 1.$$

For a proof see [2, p. 238 ff]. For a proof of the special case of Theorem 4 we will use, see [1, p. 190 ff].

Let d be an integer, $d \leq 2$. Then any number $x \in [0, 1[$ can be written uniquely in the form

$$x = \frac{x_1}{d} + \frac{x_2}{d^2} + \ldots + \frac{x_n}{d^n} + \ldots$$

where $x_i \in \{0, 1, \ldots, d-1\}$ and the sequence $\{x_1, x_2, \ldots\}$ does not end in a string of $(d-1)'s$, i.e. given n there is an $m > n$ such that $x_m \neq d-1$. Then one writes

$$x = 0\,.\,x_1x_2x_3\ldots$$

and calls this the base d expansion of x. We shall say that x is *normal to the base d* if

$$card\,\{x_i : 1 \leq i \leq n,\ x_i = c\}/n \longrightarrow 1/d$$

as $n \longrightarrow \infty$ for $c = 0, 1, \ldots, d-1$. Here $card(A)$ is the number of members of A. In other words, x is normal to the base d if each of the digits 0, 1, ..., $d-1$ occurs with equal asymptotic frequency. x is called *normal* if it is normal to the base d for every $d \geq 2$. Clearly no rational number is normal. (It base d expansion ends in a string of 0's if $x = c/d$.) It is not obvious that any number is normal.

Theorem 5 (E. Borel) *Almost all numbers are normal. That is the set of nonnormal numbers forms a Lebesgue null set.*

Proof. (See [1, pp. 195-197].) Let $d \geq 2$. Let $c \in \{0, 1, \ldots, d-1\}$. Define $X_i : [0, 1] \longrightarrow \mathbb{R}$ by $X_i(x) = 1$ or 0 according as $x_i = c$ or $x_i \neq c$, x_i being the ith entry in the base d expansion of x. Here $[0, 1]$ is regarded as a probability space equipped with Lebesgue measure. It is straightforward to check that $X_1, X_2, \ldots$ are independent, identically distributed random variables with mean $1/d$, and that

$$\frac{1}{n}\sum_{i=1}^{n} X_i(x) = n^{-1}\,card\,\{x_i : 1 \leq i \leq n,\ x_i = c\}.$$

The strong law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow 1/d \qquad a.e.,$$

i.e. almost every number is normal to the base d. Let N_d be the set of numbers which are not normal to the base d. Then the set of nonnormal numbers is $N = \bigcup_{i=1}^{\infty} N_i$, which is a Lebesgue null set since each N_i is.

An example of a normal number is the number whose base 10 (i.e. decimal) expansion is

$$0.1234567891011121314151617181920212223\ldots,$$

but this is much harder to prove.

References

[1] W. Feller, *An Introduction to Probability Theory, and Its Applictions*, Vol. I, 2nd ed., Wiley, New York, 1957.

[2] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed., Wiley, New York, 1971.

Departamento de Matemática
Universidade de Brasília
70.000 Brasília – D. F.
Brasil

and

Department of Mathematics
Tulane University
New Orleans, Louisiania 70118
U. S. A.