THE LAW OF LARGE NUMBERS

BY

D. D. KOSAMBI, *Tata Institute of Fundamental Research, Bombay.*

1. Introduction. Let $X_1, \ldots, X_n$ be a sequence of random (stochastic) variables with expectations $E(X_r) = m_r$. Then, LLN * states that under suitable restrictions upon the X_r, the difference $(X_1 + \cdots + X_n)/n - (m_1 + \cdots + m_n)/n$ may be made arbitrarily small in absolute value with probability arbitrarily close to unity by taking n sufficiently large (U, chap. 10). The proof is based upon Tshebysheff's Lemma (U, p. 182): *If X is a positive random variable, then*

$$P\{X \leqslant t^2 E(X)\} > 1 - \frac{1}{t^2}.$$

Thus, if $B_n = E(X_1 - m_1 + \cdots + X_n - m_n)^2$ exists, the lemma applied to the positive random variable $\sum_1^n (X_r - m_r)^2$ gives at once

$$\left| \frac{X_1 + \cdots + X_n}{n} - \frac{m_1 + \cdots + m_n}{n} \right| \leqslant \varepsilon \text{ with } P_n > 1 - \frac{B_n}{n^2\varepsilon^2}.$$

So, for convergence in probability, it suffices that $B_n/n^2 \to 0$. On the other hand, if for $r = 1, 2, \ldots, n$, $\max |X_r - m_r| \leqslant C_n < \infty$, and LLN holds, so that $(1 - P_n) \to 0$, we may use the easily derivable inequality $B_n < n^2 C_n^2 (1 - P_n) + n^2\varepsilon^2 P_n$ to prove that $B_n/n^2 \to 0$ follows as a necessary condition from LLN under the restriction $C_n^2 (1 - P_n) \to 0$; hence, in particular when the variables X_r are uniformly bounded, $C_n \leqslant C$ (U, pp. 185–6). This may be and has been extended in various directions, beginning with the Bienayme-Tshebysheff inequality (Cr. pp. 21; 38–39). What interests us here is the analysis of the structure of the proof, using only text book methods as far as possible.

In what follows I consider LLN *only* in the particularly useful special case when the variable X_r are all independent.

2. The law of large numbers. To cover general types of distributions, we assume at the outset that each X_r has a distribution function $F_r(x)$ which is positive, non-decreasing, with $F_r(-\infty) = 0$, $F_r(+\infty) = 1$ for all r and $P\{X_r \leqslant x\} = F_r(x)$ for all values

---

* Abbreviations used are: LLN for the law of large numbers; cf. for characteristic function; U for J.V. Uspensky, "Introduction to Mathematical Probability" (New York, 1937); Cr. for Harald Cramér, "Random Variables and Probability Distributions" (Cambridge Tract no. 36, Cambridge, 1937).

of the real variable x. The integral being taken in the sense of Lebesgue-Stieltjes, we may assume the existence of $E(X_r) = \int x dF_r(x)$ over the entire real line $-\infty \leqslant x \leqslant \infty$. Otherwise LLN has to be given a special meaning for the occasion. We make the further assumption that $\int |x| dF_r(x)$ also exists for each value of the index; this is "reasonable" in that when the distribution is given only discrete values for the observable variable, we have an infinite series replacing the integral and in view of the fact that a preferential order is hardly reconcilable with the intuitive idea of randomness, the series in question would have to converge absolutely, i. e. independently of order, to the same value.

These fundamental assumptions are then also fulfilled for the independent stochastic variable $X_r - E(X_r)$ with which alone LLN is concerned, so that there is no further loss of generality in assuming $E(X_r) = 0$ for all r. We are therefore dealing with a sequence of random independent variables $X_1, \ldots, X_n$ such that $\int x dF_r(x) = 0$ for all r, and $\int |x| dF_r(x)$ exists. LLN holds for these if and only if $(X_1 + \cdots + X_n)/n$ converges in probability to zero. For each n and N, we define non-negative functions of the two variables n, N (of which the first is only a positive integer, the second a continuous variable) as follows:

$$(1) \qquad \int_{|x|>N} |x| dF_n(x) = h(n, N) \to 0 \text{ as } N \to \infty \text{ for each } n,$$

$$h_0 = \max h(r, N), \quad r = 1, 2, \ldots, n; \quad H = \sum_1^n h(r, N).$$

$$\int_{|x|\leqslant N} |x| dF_n(x) = c(n, N); \quad c_0 = \max c(r, N), \; r = 1, 2, \ldots, n; \quad C = \sum_1^n c(r, N).$$

Following a standard device (U, p. 192) the variables X are each split up into two additive stochastic components as follows:

$$(2) \quad X_i = u_i + v_i; \text{ if } |X_i| \leqslant N, \; u_i = X_i, \; v_i = 0; \text{ otherwise } u_i = 0, \; v_i = X_i.$$

We then define $b_i = E(u_i) = -E(v_i)$ and it follows that

$$\begin{aligned} |X_1 + \cdots + X_n| &\leqslant |u_1 + \cdots + u_n| + |v_1 + \cdots + v_n| \\ &\leqslant |u_1 - b_1 + \cdots + u_n - b_n| + |b_1 + \cdots + b_n| \\ &\quad + |v_1 + \cdots + v_n|. \end{aligned}$$

By definition it also follows that

$$(3\text{—a}) \quad |b_n| = \left|\int_{|x|>N} x dF_n(x)\right| \leqslant h(n, N); \quad |b_1 + \cdots + b_n| \leqslant H(n, N).$$

$$\text{b)} \quad P\{v_n \neq 0\} = P\{|X_n| > N\} = \int_{|x|>N} dF_n(x)$$

$$= \frac{N}{N}\int_{|x|>N} dF_n(x) \leqslant \frac{1}{N}\int_{|x|>N} |x|\, dF_n(x),$$

whence $P\{v_n \neq 0\} \leqslant \frac{h(n, N)}{N}$ and $\Sigma P\{v_r \neq 0\} \leqslant H(n, N)/N$.

From (2) and the above, we have

$$(4)\quad P\left\{\frac{|X_1 + \cdots + X_n|}{n} \leqslant \varepsilon\right\}$$

$$> P\left\{\frac{|u_1 - b_1 + \cdots + u_n - b_n|}{n} \leqslant \varepsilon - \frac{H(n,N)}{n}\right\} - \frac{H(n, N)}{N}.$$

The argument is that $|X_1 + \cdots + X_n| \leqslant A$ may hold in two mutually exclusive ways: when all the $v_r = 0$, or when at least one $v \neq 0$. The probability for the latter event is allowed for by subtracting the term H/N, which it can never exceed.

The ε in (4) may be chosen arbitrarily small, so that H/n must tend to zero with increasing n. For LLN to hold, H/N must also tend to zero with increasing N, for some method of having n, N both $\to \infty$. A third condition has to be satisfied, however. The stochastic variables $u_i - b_i$ are independent with zero expectation each, and bounded so that their second moment exists; the second moment of their sum is the sum of their second moments. It is easily proved that, for any random variable, $E(X^2) \geqslant E[X - E(X)]^2$, so that $E[\Sigma (u_i - b_i)^2] = B_n \leqslant \Sigma E(u_i^2) \leqslant N\Sigma E(|u_i|)$, whence $B_n \leqslant NC$.

Applying LLN in its classical form to $u_i - b_i$, we obtain

$$(5)\quad P\left\{\frac{|u_i - b_i + \cdots + u_n - b_n|}{n} \leqslant \varepsilon - \frac{H}{n}\right\} > 1 - \frac{NC}{n^2(\varepsilon - H/n)^2}.$$

That is,

$$(6)\quad P\left\{\frac{|X_1 + \cdots + X_n|}{n} \leqslant \varepsilon\right\} > 1 - \frac{NC}{(n\varepsilon - H)^2} - \frac{H}{N}.$$

So, for *LLN* to hold, the third condition is that $NC/(n\varepsilon - H)^2 \to 0$. In this limit, the quantity $\varepsilon - H/n$ may be ignored, which gives our main result: LLN *holds*, $(X_1 + \cdots + X_n)/n$ *converging in probability to zero, when* $E(X_r) = 0$ *and* $E(|X_r|)$ *exists for all* r, *if for some manner of approach of* n *and* N *to infinity the conditions*

$$(7)\quad \frac{1}{n}\sum_1^n \int_{|x|>N} |x|\, dF_r(x) \to 0 \quad ; \quad \frac{1}{N}\sum_1^n \int_{|x|>N} |x|\, dF_r(x) \to 0,$$

$$\frac{N}{n^2}\sum_1^n \int_{|x|\leqslant N} |x|\, dF_r(x) \to 0$$

are all satisfied.

In particular, as a corollary, LLN *holds if* $h(n, N) < G(N) \to 0$. In this case we need not attempt refinement by asking the condition to hold for all large n in view of the fact that $h(n, N) \to 0$ for each n as $N \to \infty$; so that if the condition holds for $n \geqslant k$, the larger of the functions $h_0(k, N)$ and $G(N)$ will do for all n. The proof of this corollary is simple: the condition leads at once to $H(n, N) < nG(N)$, $H/n < G(N) \to 0$; $H/N < n\sqrt{G} \cdot \sqrt{G}/N \to 0$ also, if we take $n = N/\sqrt{G} \to \infty$.

One implication of this corollary is that *the absolute expectations* $E(|X_r|)$ *are bounded.* For,

$$E(|X_n|) \leqslant N + h_0 < N + G(N) < N + \varepsilon$$

for some N with ε as small as desired, and for all n. The corollary includes as special cases

1) Markoff's: LLN *holds if* $E(|X_r|^{1+\delta})$ *exist and are bounded for all* r *and some* $\delta > 0$. In this case, again, no loss of generality is caused by taking $E(X_r) = 0$. Therefore,

$$\int |x|^{1+\delta}\, dF_r(x) < A \text{ leads to } h < A/N^{\delta} = G(N).$$

2) Khintchine's LLN *holds if all the* X_r *have the same distribution and* $E(|X|)$ *exists.* In this case, for all n,

$$h = \int_{|x|>N} |x|\, dF(x) = G(N) \to 0,$$

taking $E(X) = 0$ as before. Our result is only a slight and almost obvious refinement of existing deductions. The basic process (due apparently to Markoff is the division of the variables into two portions, of which one is bounded and the other contains values of negligible probability. The question still remains: what is the function of the *second* moment for the bounded part in the deduction?

3. The law of large numbers and the central limit theorem. For a simple random variate X, bounded with $|X| \leqslant M$ and zero expectation, the probability of X lying outside a given interval centered at the origin can be increased by increase of the dispersion. Under the conditions of the problem, there is an optimum "most scattered" distribution, independent of the particular and variable limits outside which X may be asked to lie from stage to stage; that is, $X = \pm M$ with probability ½ each will give the greatest possible scattering once for all. Then $P\{|X| > S\} = 0$ if $S \geqslant M$ and $= 1$ if $S < M$. If the value of S be preassigned and not greater than M, we have some choice in limiting the random variable, but the distribution given here is the optimum in that it gives the greatest probability, independently of S.

For two independent variables of the same type, the sum $X_1 + X_2$ can be dispersed in a similar manner. Here, $P\{|X_1 + X_2| > S\} = 1, \frac{1}{2}$, or 0 according as $S < M$, $M \leqslant S < 2M$, or $S \geqslant 2M$. The first can be obtained in particular by taking one of the variables $\pm M$ with probability $\frac{1}{2}$ each and the other 0 with probability 1. In the last case no permissible choice of distributions can possibly give any other probability than zero. In the middle case, however, there does exist an optimum, i.e. X_1, $X_2 = \pm M$, $p = \frac{1}{2}$ each, so that the value of the sum would be $\pm 2M$ with $p = \frac{1}{4}$ each and 0 with $p = \frac{1}{2}$. This, incidentally, points out the essential reason for the validity of our LLN, in that the values of the independent stochastic variables with zero expectations cancel out. But it is of basic importance for the optimum to exist, in such addition, that the interval $2S$ be sufficiently large. For the purpose of the preceding section, it suffices to note that S must lie between $(n-1)M$ and nM to enforce the optimum scattering for the sum of n variates. In the variables $u_i - b_i$, S corresponds to $n\varepsilon - H$, and M to N, or rather to $N + h_0$. So, N must be of the same order as $\varepsilon - H/n$, which will not do in view of the fact that ε is arbitrarily small while N has to be taken arbitrarily large.

If we take the distribution with optimum scattering, $X_i = X = \pm M$, with $p = \frac{1}{2}$ each, then each X has the c.f.

$$\tfrac{1}{2}(e^{-iMt} + e^{+iMt}) = \cos Mt = 1 - 2\sin^2 \tfrac{1}{2} Mt.$$

The c.f. of the sum of n is therefore $(1 - 2\sin^2 \frac{1}{2} Mt)^n$, which, for the average of n tends to $(1 - M^2t^2/2n^2)^n \to \exp(-M^2t^2/2n)$. So the general distribution, bounded by the most scattered case, is thus bounded by a normal distribution with zero mean and variance M^2/n, in substance. If M and N are of the same order, we must have $(N + h_0)^2/n \to 0$ simultaneously with $n\varepsilon - H \sim n(N + h_0)$; but in LLN it is essential to have $N \to \infty$ unless we restrict ourselves to almost trivial cases. These conditions obviously contradict each other.

The function of the second moment, or any moment higher than the first, is to bridge this gap between the requirements of LLN and the approach of this section, by consideration of the most scattered variables.

Suppose that to our preceding assumptions $E(X) = 0$, $|X| \leqslant M$, we add the condition $E(X^2) \leqslant kM$. In the former case, $E(X^2) = M^2$ for the most scattered distribution, so that for $k < M$ (*which we assume hereafter*), we have less concentrated scattering. The extreme limits $\pm M$ can no longer be attained as before with $p = \frac{1}{2}$. The greatest concentrated scattering is, in fact, now given by $\pm \sqrt{kM}$, $p = \frac{1}{2}$ each.

Nevertheless, $P\{|X| \geqslant S\}$ need not vanish if we push S beyond $\sqrt{kM}$ so that the actual optimum in this case is much less clear-cut. For $M > S \geqslant \sqrt{kM}$, the best choice is clearly to take $X = \pm S$ with probability $kM/2S^2$ for each value, and $X=0$ with probability $1-kM/S^2$. If the probability is reduced at either of the two extremes, it would be necessary to add an extra value on the same side of zero, or to cut down the probability for the other extreme, in order to preserve the mean value $E(X)=0$. The optimum for this three-valued distribution is more clearly dependent upon the choice of interval, and the probabilities for the greatest scattering also depend upon S. The c. f. for a single such scattered variable is $1-(2kM/S^2)\sin^2 \frac{1}{2} St$. For the sum of n, we raise this to the nth power, for the average of n, we have only to replace t by t/n in raising to nth power. Similarly, for the sum of n, the second moment is $B_n \leqslant nkM$; for the average, we replace each X by X/n and B_n by $B_n/n^2 \leqslant kM/n$. So, in getting the distribution of the average of n values we should not only raise the c. f. to the nth power but also replace t by t/n and S by $S/\sqrt{n}$. Making these substitutions, we get the most scattered distribution of the average as having the c. f.

$$(8) \quad \left(1-\frac{2kMn}{S^2}\sin^2\frac{St}{2n\sqrt{n}}\right)^n \to \left(1-\frac{kMt^2}{2n^2}\right)^n \to e^{-kMt^2/2n}.$$

In spite of its fundamental rôle, S has cancelled out in the process of deduction, to give a normal distribution for the bounding scattered variate. The mean is zero as before, the variance reduced to kM/n. This must tend to zero if LLN is to hold. For the attack of section 2 we should have to take $M=N+h_0$, $k=c_0$, $S \sim \sqrt{c_0 nM} = n(\varepsilon - h_0)$. The condition would then take on the form $h_0 \to 0$, $c_0(N+h_0)/n \to 0$, which it is possible to fulfil, as for example under the assumptions of the preceding section.

That is, the second moment or any moment higher than the first helps only by restricting the admissible variation, the maximum possible scattering. The analysis of this section shows in addition to this a general connection between LLN and the central limit theorem. In fact, we have the limiting values of the bounding distributions as normal distributions.