INFINITE COIN-TOSS AND THE LAWS OF LARGE NUMBERS

The traditional interpretation of the *probability* of an event E is its *asymptotic frequency*: the limit as $n \to \infty$ of the fraction of n repeated, similar, and independent trials in which E occurs. Similarly the "expectation" of a random variable X is taken to be its *asymptotic average*, the limit as $n \to \infty$ of the average of n repeated, similar, and independent replications of X. As statisticians trying to make inference about the underlying probability distribution $f(x|\theta)$ governing observed random variables X_i, this suggests that we should be interested in the probability distribution for large n of quantities like the average of the RV's, $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^n X_i$.

Three of the most celebrated theorems of probability theory concern this sum. For independent random variables X_i, all with the same probability distribution satisfying $\mathsf{E}|X_i|^3 < \infty$, set $\mu = \mathsf{E}X_i$, $\sigma^2 = \mathsf{E}|X_i - \mu|^2$, and $S_n = \sum_{i=1}^n X_i$. The three main results are:

Laws of Large Numbers:

$$\frac{S_n - n\mu}{\sigma n} \longrightarrow 0 \qquad (i.p. \text{ and } a.s.)$$

Central Limit Theorem:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Longrightarrow N(0,1) \qquad (i.d.)$$

Law of the Iterated Logarithm:

$$\limsup_{n\to\infty} \pm \frac{S_n - n\mu}{\sigma\sqrt{2n\log\log n}} = 1.0 \qquad (a.s.)$$

Together these three give a clear picture of how quickly and in what sense $\frac{1}{n}S_n$ tends to μ. We begin with the Law of Large Numbers (LLN), in its "weak" form (asserting convergence *i.p.*) and in its "strong" form (convergence *a.s.*). There are several versions of both theorems. The simplest requires the X_i to be IID and L_2; stronger results allow us to weaken (but not eliminate) the independence requirement, permit non-identical distributions, and consider what happens if the RV's are only L_1 (or worse!) instead of L_2.

The text covers these things well; to complement it I am going to: (1) Prove the simplest version, and with it the Borel-Cantelli theorems; and (2) Show what happens with Cauchy random variables, which don't satisfy the requirements (the LLN fails).

I. Weak version, non-*iid*, L_2: $\mu_i = \mathsf{E}X_i$, $\sigma_{ij} = \mathsf{E}[X_i - \mu_i][X_j - \mu_j]$
   A. $Y_n = (S_n - \Sigma\mu_i)/n$ satisfies $\mathsf{E}Y_n = 0$, $\mathsf{E}Y_n^2 = \frac{1}{n^2}\Sigma_{i\le n}\sigma_{ii} + \frac{2}{n^2}\Sigma_{i<j\le n}\sigma_{ij}$;
      1. If $\sigma_{ii} \le M$ and $\sigma_{ij} \le 0$ *or* $|\sigma_{ij}| < Mr^{|i-j|}$, $r < 1$, Chebychev $\Longrightarrow Y_n \to 0$, *i.p.*
      2. (pairwise) IID L_2 is OK

II. Strong version, non-*iid*, L_2: $\mathsf{E}X_i = 0$, $\mathsf{E}X_i^2 \le M$, $\mathsf{E}X_iX_j \le 0$.
   A. $\mathsf{P}[|S_n| > n\epsilon] < \frac{Mn}{n^2\epsilon^2} = \frac{M}{n\epsilon^2}$
      1. $\mathsf{P}[|S_{n^2}| > n^2\epsilon] < \frac{M}{n^2\epsilon^2}$, $\Sigma_n\mathsf{P}[|S_{n^2}| > n^2\epsilon] < \frac{M\pi^2}{6\epsilon^2}$
      2. Borel-Cantelli: $\mathsf{P}[|S_{n^2}| > n^2\epsilon \text{ } i.o.] = 0$, $\therefore \frac{1}{n^2}S_{n^2} \to 0$ *a.s.*
      3. $D_n = \max_{n^2\le k<(n+1)^2}|S_k - S_{n^2}|$, $\mathsf{E}D_n^2 \le 2n\mathsf{E}|S_{(n+1)^2} - S_{n^2}| \le 4n^2M$
      4. Chebychev: $\mathsf{P}[D_n > n^2\epsilon] < \frac{4n^2M}{n^4\epsilon^2}$, $\therefore D_n \to 0$ *a.s.*
   B. $|S_k/k| \le \frac{|S_{n^2}|+D_n}{n^2} \to 0$ *a.s.*, QED
      1. Bernoulli RV's, normal number theorem, Monte Carlo integration.

III. Weak version, pairwise-*iid*, L_1
   A. Equivalent sequences: $\sum_n \mathsf{P}[X_n \ne Y_n] < \infty$
      1. $\sum_n[X_n - Y_n] < \infty$ *a.s.*
      2. $\sum_{i=1}^n[X_i]$, $a_n\sum_{i=1}^n[X_i]$ converge iff $\sum_{i=1}^n[Y_i]$, $a_n\sum_{i=1}^n[Y_i]$ both converge
      3. $Y_n = X_n 1_{[|X_n|\le n]}$

IV. Counterexamples: Cauchy,
   A. $X_i \sim \frac{dx}{\pi[1+x^2]} \Longrightarrow \mathsf{P}[|S_n|/n \le \epsilon] \equiv \frac{2}{\pi}\tan^{-1}(\epsilon) \not\to 1$, WLLN fails.
   B. $\mathsf{P}[X_i = \pm n] = \frac{c}{n^2}$, $n \ge 1$; $X_i \notin L_1$, and $S_n/n \not\to 0$ *i.p.* or *a.s.*
   C. $\mathsf{P}[X_i = \pm n] = \frac{c}{n^2\log n}$, $n > 1$; $X_i \notin L_1$, but $S_n/n \to 0$ *i.p.* and not *a.s.*
   D. Medians: for ANY RV's $X_n \to X_\infty$ *i.p.*, then $m_n \to m_\infty$ if m_∞ is unique.

Let X_i be *iid* standard Cauchy RV's, with

$$\mathsf{P}[X_1 \le t] = \int_{-\infty}^{t} \frac{dx}{\pi[1+x^2]} = \frac{1}{2} + \frac{1}{\pi}\arctan(t)$$

and characteristic function

$$\mathsf{E}\, e^{i\omega X_1} = \int_{-\infty}^{\infty} e^{i\omega x} \frac{dx}{\pi[1+x^2]} = e^{-|\omega|},$$

so S_n/n has characteristic function

$$\mathsf{E}\, e^{i\omega S_n/n} = \mathsf{E}\, e^{i\frac{\omega}{n}[X_1+\cdots+X_n]} = \left(\mathsf{E}\, e^{i\frac{\omega}{n}X_1}\right)^n = (e^{-|\frac{\omega}{n}|})^n = e^{-|\omega|}$$

and S_n/n also has the standard Cauchy distribution with $\mathsf{P}[S_n/n \le t] = \frac{1}{2} + \frac{1}{\pi}\arctan(t)$; in particular, S_n/n does not converge almost surely, or even in probability.

A LAW OF LARGE NUMBERS FOR CORRELATED SEQUENCES

In many applications we would like a Law of Large Numbers for sequences of random variables that are *not* independent; for example, in Markov Chain Monte Carlo integration, we have a stationary Markov chain $\{X_t\}$ (this means that the distribution of X_t is the same for all t and that the conditional distribution of X_u for $u > t$, given $\{X_s | s \le t\}$, depends only on X_t) and want to estimate the population mean $\mathsf{E}[\phi(X_t)]$ for some function $\phi(\cdot)$ by the sample mean

$$\mathsf{E}[\phi(X_t)] \approx \frac{1}{T}\sum_{t=1}^{T} \phi(X_t).$$

Even though they are identically distributed, the random variables $Y_t \equiv \phi(X_t)$ won't be independent if the X_t aren't independent, so the LLN we already have doesn't quite apply.

A sequence of random variables Y_t is called *stationary* if each Y_t has the same probability distribution and, moreover, each finite set $(Y_{t_1+h}, Y_{t_2+h}, ..., Y_{t_k+h})$ has a joint distribution that doesn't depend on h. The sequence is called "L_2" if each Y_t has a finite variance σ^2 (and hence also a well-defined mean μ); by stationarity it also follows that the *covariance*

$$\gamma_{st} = \mathsf{E}[(Y_s - \mu)(Y_t - \mu)]$$

is finite and depends only on the absolute difference $|t - s|$.

Theorem. If a stationary L_2 sequence has a summable covariance, *i.e.*, satisfies $\sum_{t=-\infty}^{\infty} |\gamma_{st}| \le c < \infty$, then

$$\mathsf{E}[Y_t] = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} Y_t.$$

Proof. Let S_T be the sum of the first T Y_t's and set (as usual) $\bar{Y}_T \equiv S_T/T$. The variance of S_T is

$$\begin{aligned}
\mathsf{E}[(S_T - T\mu)^2] &= \sum_{s=1}^{T}\sum_{t=1}^{T} \mathsf{E}[(X_s - \mu)(X_t - \mu)] \\
&\le \sum_{s=1}^{T}\sum_{t=-\infty}^{\infty} |\gamma_{st}| \\
&\le T\, c,
\end{aligned}$$

so $\bar{Y}_T$ had variance $\mathsf{V}[\bar{Y}_T] \le c/T$ and by Chebychev's inequality

$$\begin{aligned} \mathsf{P}[|\bar{Y}_T - \mu| > \epsilon] &\le \frac{\mathsf{E}[(\bar{Y}_T - \mu)^2]}{\epsilon^2} \\ &= \frac{\mathsf{E}[(S_T - T\mu)^2]}{T^2\epsilon^2} \\ &\le \frac{T\,c}{T^2\epsilon^2} \\ &= \frac{c}{T\epsilon^2} \to 0 \quad \text{as } T \to \infty. \end{aligned}$$

A strong LLN follows with a bit more work, just as for *iid* random variables.

Examples

1. **IID:** If X_t are independent and identically distributed, and if $Y_t = \phi(X_t)$ has finite variance σ^2, then Y_t has a well-defined finite mean μ and $\bar{Y}_T \to \mu$.
   Here $\gamma_{st} = \begin{cases} \sigma^2 & \text{if } s = t \\ 0 & \text{if } s \ne t \end{cases}$, so $c = \sigma^2 < \infty$.
2. **AR$_1$:** If Z_t are *iid* $N(0,1)$ for $-\infty < t < \infty$, $\mu \in \mathbb{R}$, $\sigma > 0$, $-1 < \rho < 1$, and
   $$\begin{aligned} X_t &\equiv \mu + \sigma \sum_{s=0}^{\infty} \rho^s Z_{t-s} \\ &= \rho X_{t-1} + \alpha + \sigma Z_t, \end{aligned} \qquad (*)$$
   where $\alpha = (1-\rho)\mu$, then the X_t are identically distributed (all with the $N(\mu, \frac{\sigma^2}{1-\rho^2})$ distribution) but not independent (since $\gamma_{st} = \frac{\sigma^2}{1-\rho^2}\rho^{|s-t|} \ne 0$); this is called an "autoregressive process" (because of equation (*), expressing X_t as a regression of previous X_s's) of order one (because only one earlier X_s appears in (*)), and is about the simplest non-*iid* sequence occuring in applications. Since the covariance is summable,
   $$\sum_{t=-\infty}^{\infty} |\gamma_{st}| = \frac{\sigma^2}{1-\rho^2}\frac{1+|\rho|}{1-|\rho|} = \frac{\sigma^2}{(1-|\rho|)^2} < \infty,$$
   we again have $\bar{X}_T \to \mu$ as $T \to \infty$.
2. **Geometric Ergodicity:** If for some $0 < \rho < 1$ and $c > 0$ we have $\gamma_{st} \le c\rho^{|s-t|}$ for a Markov chain Y_t the chain is called *Geometrically Ergodic* (because $c\rho^t$ is a geometric sequence), and the same argument as for AR$_1$ shows that $\bar{Y}_t$ converges; Meyn & Tweedie (1993), Tierney (1994), and others have given conditions for MCMC chains to be Geometric Ergodic, and hence for the almost-sure convergence of sample averages to population means.
3. **General Ergodicity:** Consider the three sequences of random variables on $(\Omega, \mathcal{F}, \mathsf{P})$ with $\Omega = (0,1]$ and $\mathcal{F} = \mathcal{B}(\Omega)$, each with $X_0(\omega) = \omega$:
   1. $X_{n+1} \equiv 2\,X_n \pmod 1$;
   2. $X_{n+1} \equiv X_n + \alpha \pmod 1$ (Does it matter if α is rational?);
   3. $X_{n+1} \equiv 4X_n(1-X_n)$.

   For each, find a probability measure P (equivalently find a distribution for X_0) such that the X_n are all identically distributed; the sequence is called *ergodic* if each $E \in \mathcal{F}$ left invariant by the transformation T that takes X_n to X_{n+1}, $E = T^{-1}(E)$, is trivial in the sense that $\mathsf{P}[E] = 0$ or $\mathsf{P}[E] = 1$. The *Ergodic Theorem* asserts that $\bar{X}_n$ converges almost-surely to a T-invariant limit X_∞ as $n \to \infty$; since only constants are T-invariant for ergodic sequences, it follows that $\bar{X}_n \to \mu = \mathsf{E}X_n$. The conditions here are weaker than those for the usual LLN; in all three cases above, for example, each X_n is completely determined by X_0 so there is complete dependence!

Stable Limit Laws

Let $S_n = X_1 + ... + X_n$ be the partial sum of *iid* random variables. IF the random variables are all square integrable, THEN the Central Limit Theorem applies and necessarily $\frac{S_n}{n\sigma^2} - \mu \Longrightarrow \mathsf{No}(0,1)$. But what if each X_n is *not* square integrable? We have already seen CLT fail for Cauchy variables X_j. Denote by $F(x) = \mathsf{P}[X_n \le x]$ the common CDF of the $\{X_n\}$.

Theorem (Stable Limit Law).

There exist constants $A_n > 0$ and $B_n \in \mathbb{R}$ and a distribution μ for which the

$$\frac{S_n}{A_n} - B_n \Longrightarrow \mu$$

if and only if there are constants $0 < \alpha \le 2$, $M^- \ge 0$, and $M^+ \ge 0$, with $M^- + M^+ > 0$, such that the following limits hold for every $\xi > 0$ as $x \to +\infty$:

1. $\dfrac{F(-x)}{1-F(x)} = \dfrac{\mathsf{P}[X \le -x]}{\mathsf{P}[X > x]} \to \dfrac{M^-}{M^+}$;
2. $M^+ > 0 \Rightarrow \dfrac{1-F(x\xi)}{1-F(x)} \to \xi^{-\alpha}$ $\qquad M^- > 0 \Rightarrow \dfrac{F(-x\xi)}{F(-x)} \to \xi^{-\alpha}$.

In this case the limit is the **Stable Distribution** with index α, with characteristic function

$$\mathsf{E}[e^{i\omega Y}] = e^{i\delta\omega - \gamma|\omega|^\alpha[1 - i\beta\tan\frac{\pi\alpha}{2}\,\mathrm{sgn}(\omega)]},$$

where $\beta = \frac{M^+}{M^- + M^+}$ and $\gamma = (M^- + M^+)$. The sequence A_n must be essentially $A_n \propto n^{1/\alpha}$ (more precisely, the sequence $C_n = n^{-1/\alpha} A_n$ is *slowly changing* in the sense that

$$1 = \lim_{n\to\infty} \frac{C_{cn}}{C_n}$$

for every $c > 0$); thus partial sums converge to stable distributions at rate $n^{-1/\alpha}$, more slowly (*much* more slowly, if α is close to one) than in the L^2 (Gaussian) case of the central limit theorem.

Note that the **Cauchy** distribution is the special case of $(\alpha, \beta, \gamma, \delta) = (1, 0, 1, 0)$ and the **Normal** distribution is the special case of $(\alpha, \beta, \gamma, \delta) = (2, 0, \sigma^2/2, \mu)$. Although each Stable distribution has an absolutely continuous distribution with continuous probability density function $f(y)$, these two cases and the "inverse gamma distribution" with $\alpha = 1/2$ and $\beta = \pm 1$ are the only ones where the p.d.f. can be given in closed form. Moments are easy enough to compute; for $\alpha < 2$ the Stable distribution only has finite moments of order $p < \alpha$ and, in particular, *none* of them has a finite variance. The Cauchy has finite moments of order $p < 1$ but does not have a well-defined mean.

Condition 2. says that each tail must be fall off like a power (sometimes called *Pareto tails*), and the powers must be identical; Condition 1. gives the tail ratio. In a common special case, $M^- = 0$; for example, random variables X_n with the Pareto distribution (often used to model income) given by $P[X_n > t] = (k/t)^\alpha$ for $t \ge k$ will have a stable limit for their partial sums if $\alpha < 2$, and (by CLT) a normal limit if $\alpha \ge 2$. You can find out more details reading Chapter 9 of Breiman's book.