7 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 $X$. For 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 $E|X_i|^3 < \infty$, set $\mu = E X_i$, $\sigma^2 = E|X_i - \mu|^2$, and $S_n = \sum_{i=1}^{n} X_i$. The three main results are:

Laws of Large Numbers:
$$\frac{S_n - n\mu}{\sigma n} \rightarrow 0 \quad \text{(i.p. and a.s.)}$$

Central Limit Theorem:
$$\frac{S_n - n\mu}{\sigma \sqrt{n}} \Rightarrow \mathrm{No}(0,1) \quad \text{(in dist.)}$$

Law of the Iterated Logarithm:
$$\limsup_{n \to \infty} \pm \frac{S_n - n\mu}{\sigma \sqrt{2n \log \log n}} = 1.0 \quad \text{(a.s.)}$$

Together these three give a clear picture of how quickly and in what sense $\frac{1}{n} S_n$ tends to $\mu$. 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 we relax the $L_2$ requirement and allow the RV’s to be only $L_2$ or $L_1$ (or worse!).

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).

7.1 Proofs of the Weak and Strong Laws

Here are two simple versions (one Weak, one Strong) of the Law of Large Numbers; first we prove an elementary but very useful result:

**Proposition 1 (Markov’s Inequality)** Let $\phi(x) \geq 0$ be non-decreasing on $\mathbb{R}_+$. For any random variable $X \geq 0$ and constant $a \in \mathbb{R}_+$,

$$P[X \geq a] \leq P[\phi(X) \geq \phi(a)] \leq E[\phi(X)]/\phi(a)$$

To see this, set $Y = \phi(a)1_A$ for the event $A = \{\phi(X) \geq \phi(a)\}$ and note $Y \leq \phi(X)$ so $EY \leq E\phi(X)$.

**Theorem 1 (L\textsubscript{2} WLLN)** Let $\{X_n\}$ be independent random variables with the same mean $\mu = E[X_n]$ and uniformly bounded variance $E(X_n - \mu)^2 \leq B$ for some fixed bound $B < \infty$. Set $S_n = \sum_{j \leq n} X_j$ and $\bar{X}_n \equiv S_n/n = \frac{1}{n} \sum_{j \leq n} X_j$. Then:

$$\forall \epsilon > 0 \quad P[|\bar{X}_n - \mu| > \epsilon] \to 0. \quad (1)$$

**Proof.**

$$E(S_n - n\mu)^2 = \sum_{i=1}^{n} E(X_i - \mu)^2 \leq nB$$

so for $\epsilon > 0$

$$P[|\bar{X}_n - \mu| > \epsilon] = P[(S_n - n\mu)^2 > (n\epsilon)^2] \leq E[(S_n - n\mu)^2]/n^2\epsilon^2 \leq B/n\epsilon^2 \to 0 \quad \text{as } n \to \infty.$$

This Law of Large Numbers is called *weak* because its conclusion is only that $\bar{X}_n$ converges to zero *in probability* (Eqn (1)); the *strong* Law of Large Numbers asserts convergence of a stronger sort, called *almost sure* convergence (Eqn (2) below). If $P[|\bar{X}_n - \mu| > \epsilon]$ were *summable* then by B-C
we could conclude almost-sure convergence; unfortunately we have only the bound \( P[|\bar{X}_n - \mu| > \epsilon] < c/n \) which tends to zero but isn’t summable. It is summable along the subsequence \( n^2 \), however; our approach to proving a strong LLN is to show that \( |S_k - S_{n^2}| \) isn’t too big for any \( n^2 < k < (n+1)^2 \).

**Theorem 2 (L2 SLLN)** Under the same conditions,

\[
P[\bar{X}_n \to \mu] = 1.
\]

**Proof.** Without loss of generality take \( \mu = 0 \) (otherwise subtract \( \mu \) from each \( X_n \)), and fix \( \epsilon > 0 \). Set \( S_n \equiv \sum_{j \leq n} X_j \). Then

\[
P[|S_n| > n\epsilon] \leq E|S_n|^2/(n\epsilon)^2 \leq nB/n^2\epsilon^2 = B/n\epsilon^2
\]

\[
P[|S_{n^2}| > n^2\epsilon \text{ i.o.}] = 0 \text{ by B-C } \Rightarrow S_{n^2}/n^2 \to 0 \text{ a.s.}
\]

Set

\[
D_n \equiv \max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|
\]

\[
ED_n^2 = E \left[ \max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|^2 \right]
\]

\[
\leq E \sum_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|^2 = \sum_{n^2 \leq k < (n+1)^2} E|S_k - S_{n^2}|^2
\]

\[
\leq \sum_{n^2 \leq k < (n+1)^2} (k - n^2)B \leq 4n^2B, \text{ so}
\]

\[
P[D_n > n^2\epsilon] \leq 4n^2B/n^4\epsilon^2 \Rightarrow D_n/n^2 \to 0 \text{ a.s.}
\]

\[
\left| \frac{S_k}{k} \right| \leq \frac{|S_{n^2}| + D_n}{n^2} \to 0 \text{ a.s. as } k \to \infty, \text{ where } n = \lfloor \sqrt{k} \rfloor.
\]

Each of these LLNs required only that \( \text{Cov}(X_n, X_m) \leq 0 \), not pairwise (let alone full) independence; we’ll see below in Section (7.2) that even positive correlations are okay if the fall off fast enough (e.g., if \( |\text{Cov}(X_n, X_m)| \leq ar^n \) for some \( a > 0, 0 < r < 1 \), with a similar proof. The uniform \( L_2 \) bound isn’t necessary either. There are a variety of LLNs with either or both of the \( L_2 \) bound and independence weakened in some way, but they can’t be dispensed with altogether—consider iid Cauchy random variables, for example, to show \( L_2 \) isn’t entirely surperfluous, or \( X_n \equiv X_1 \) with any nontrivial distribution to show the need for at least a modicum of independence.

Here’s a summary of LLN facts:
I. Weak version, non-iid, \(L_2\): \(\mu_i = \mathbb{E}X_i, \sigma_{ij} = \mathbb{E}[X_i - \mu_i][X_j - \mu_j]\)

A. \(Y_n = (S_n - \Sigma \mu_i)/n\) satisfies \(\mathbb{E}Y_n = 0, \mathbb{E}Y_n^2 = \frac{1}{n^2}\Sigma_{i \leq n} \sigma_{ii} + \frac{2}{n^2}\Sigma_{i < j \leq n} \sigma_{ij}\):

1. If \(\sigma_{ii} \leq M\) and \(\sigma_{ij} \leq 0\) or \(|\sigma_{ij}| < Mr^{i-j}, r < 1\), Chebychev \(\implies Y_n \rightarrow 0, i.p.\)
2. (pairwise) IID \(L_2\) is OK

II. Strong version, non-iid, \(L_2\): \(\mathbb{E}X_i = 0, \mathbb{E}X_i^2 \leq M, \mathbb{E}X_i X_j \leq 0\).

A. \(\mathbb{P}[|S_n| > ne] < \frac{Mn}{n^2\epsilon^2} = \frac{M}{n\epsilon^2}\)

1. \(\mathbb{P}[|S_n|^2 > n^2\epsilon] < \frac{Mn^2}{n^2\epsilon^2}, \Sigma_n \mathbb{P}[|S_n|^2 > n^2\epsilon] < \frac{Mn^2}{6\epsilon^2} < \infty\)
2. Borel-Cantelli: \(\mathbb{P}[|S_n|^2 > n^2\epsilon \ i.o.] = 0\), \(\therefore \frac{1}{n^2} S_n^2 \rightarrow 0 \ a.s.\)
3. \(D_n \equiv \max_{n^2 \leq k < (n+1)^2} |S_k - S_n^2|\), so \(\mathbb{E}D_n^2 \leq 2n\mathbb{E}|S_{(n+1)^2} - S_n^2| \leq 4n^2 M\)
4. Chebychev: \(\mathbb{P}[D_n > n^2\epsilon] < \frac{4n^2 M}{n^2\epsilon^2}, \therefore D_n/n^2 \rightarrow 0 \ a.s.\)

B. \(|S_k/k| \leq \frac{|S_n|^2 + D_n}{n^2} \rightarrow 0 \ a.s.\) as \(k \rightarrow \infty\), QED

1. Bernoulli RV’s, normal number theorem, Monte Carlo integration.

III. Weak version, pairwise-iid, \(L_1\)

A. Equivalent sequences: \(\sum_n \mathbb{P}[X_n \neq Y_n] < \infty\)

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 both \(\sum_{i=1}^n [Y_i], a_n \sum_{i=1}^n [Y_i]\) converge
3. \(Y_n = X_n\mathbb{1}_{|[X_n]| \leq n}\)

IV. Counterexamples: Cauchy,

A. \(X_i \sim \frac{dx}{\pi [1+x^2]} \implies \mathbb{P}[|S_n|/n \leq \epsilon] \equiv \frac{2}{\pi} \tan^{-1}(\epsilon) \neq 1, \text{ WLLN fails.}\)
B. \(\mathbb{P}[X_i = \pm n] = \frac{1}{n^2}, n \geq 1; X_i \notin L_1, \text{ and } S_n/n \neq 0 \ i.p. \text{ or } a.s.\)
C. \(\mathbb{P}[X_i = \pm n] = \frac{c}{n^2 \log n}, n > 1; X_i \notin L_1, \text{ but } S_n/n \rightarrow 0 \ i.p. \text{ and not } a.s.\)

D. Medians: for ANY RV’s \(X_n \rightarrow X_\infty \ i.p., \text{ then } m_n \rightarrow m_\infty \text{ if } m_\infty \text{ is unique.}\)
Let $X_i$ be iid standard Cauchy RV's, with

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

and characteristic function

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

so $S_n/n$ has characteristic function

\[ \mathbb{E} e^{i\omega S_n/n} = \left( \mathbb{E} e^{i\omega X_1} \right)^n = (e^{-|\omega|})^n = e^{-|\omega|} \]

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

### 7.2 An LLN 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 $\{X_s|s \leq t\}$, depends only on $X_t$) and want to estimate the population mean $\mathbb{E} [\phi(X_t)]$ for some function $\phi(\cdot)$ by the sample mean

\[ \mathbb{E} [\phi(X_t)] \approx \frac{1}{T} \sum_{t=0}^{T-1} \phi(X_t). \]

Even though they are identically distributed, the random variables $Y_t = \phi(X_t)$ 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 $(Y_{t_1+h}, Y_{t_2+h}, \ldots, Y_{t_k+h})$ has a joint distribution that doesn’t depend on $h$. The sequence is called “$L_2$” if each $Y_t$ has a finite variance $\sigma^2$ (and hence also a well-defined mean $\mu$); by stationarity it also follows that the covariance

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

is finite and depends only on the absolute difference $|t - s|$.
Theorem 3 If a stationary $L_2$ sequence has a summable covariance, i.e., satisfies $\sum_{t=\infty}^{t=-\infty} |\gamma_{st}| \leq c < \infty$, then

$$E[Y_t] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} Y_t.$$ 

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

$$E[(S_T - T\mu)^2] = \sum_{s=0}^{T-1} \sum_{t=0}^{T-1} E[(X_s - \mu)(X_t - \mu)]$$

$$\leq \sum_{s=0}^{T-1} \sum_{t=-\infty}^{\infty} |\gamma_{st}|$$

$$\leq T \cdot c,$$

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

$$P[|\bar{Y}_T - \mu| > \epsilon] \leq \frac{E[(\bar{Y}_T - \mu)^2]}{\epsilon^2} = \frac{E[(S_T - T\mu)^2]}{T^2\epsilon^2} \leq \frac{T \cdot c}{T^2\epsilon^2} = \frac{c}{T\epsilon^2} \to 0 \quad \text{as } T \to \infty.$$ 

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

7.2.1 Examples

1. IID: If $X_t$ are independent and identically distributed, and if $Y_t = \phi(X_t)$ has finite variance $\sigma^2$, then $Y_t$ has a well-defined finite mean $\mu$ and $\bar{Y}_T \to \mu$.

Here $\gamma_{st} = \begin{cases} \sigma^2 & \text{if } s = t, \\ 0 & \text{if } s \neq t, \end{cases}$ so $c = \sigma^2 < \infty$. 

Page 6 July 12, 2011
2. AR₁: If \( Z_t \) are iid \( \mathcal{N}(0, 1) \) for \( -\infty < t < \infty, \mu \in \mathbb{R}, \sigma > 0, -1 < \rho < 1 \), and

\[
X_t \equiv \mu + \sigma \sum_{s=0}^{\infty} \rho^s Z_{t-s} = \rho X_{t-1} + \alpha + \sigma Z_t, \tag{*}
\]

where \( \alpha = (1-\rho)\mu \), then the \( X_t \) are identically distributed (all with the \( \mathcal{N}(\mu, \sigma^2) \) distribution) but not independent (since \( \gamma_{st} = \sigma^2 (1-\rho^{|s-t|}) \neq 0 \)); 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 occurring 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 \( T \to \infty \).

3. Geometric Ergodicity: If for some \( 0 < \rho < 1 \) and \( c > 0 \) we have \( \gamma_{st} \leq \rho^{s-t} \) 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 AR₁ 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.

4. General Ergodicity: Consider the three sequences of random variables on \((\Omega, \mathcal{F}, P)\) with \( \Omega = (0, 1] \) and \( \mathcal{F} = \mathcal{B}(\Omega) \), each with \( X_0(\omega) = \omega \):

(a) \( X_{n+1} \equiv 2 X_n \) (mod 1);
(b) \( X_{n+1} \equiv X_n + \alpha \) (mod 1) (Does it matter if \( \alpha \) is rational?);
(c) \( X_{n+1} \equiv 4X_n(1-X_n) \).

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 “almost trivial” in the sense that \( P[E] = 0 \) or \( P[E] = 1 \). Birkhoff’s Ergodic Theorem asserts that \( \bar{X}_n \) converges almost-surely to a \( T \)-invariant limit \( X_\infty \) as \( n \to \infty \); since
only constants are $T$-invariant for ergodic sequences, it follows that $\bar{X}_n \to \mu = E X_n$. 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!

For any $L_1$ distribution $\mu(dx)$ on $(\mathbb{R}, \mathcal{B})$, we can construct iid random variables $X_n$ on the product probability space $(\Omega = \mathbb{R}^\infty, \mathcal{F} = \mathcal{B}^\infty, \mathbb{P} = \otimes \mu_n)$ and a measure-preserving transformation $T : \Omega \to \Omega$ by

$$T(\omega_1, \omega_2, \omega_3, \ldots) = (\omega_2, \omega_3, \omega_4, \ldots),$$

the left-shift. The $\sigma$-algebra $\mathcal{T} = \{A : A = T^{-1}(A)\}$ of $T$-invariant sets is just the tail $\sigma$-algebra for the independent random variables $\{X_n : X_n(\omega) = \omega_n\}$, so by Kolmogorov’s zero-one law $\mathcal{T}$ is almost-trivial and so $T$ is ergodic. It follows from Birkhoff’s Ergodic Theorem that

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$$

converges almost-surely to a $T$-invariant and hence almost-surely constant random variable whose value must be $\mu$, proving a strong LLN for iid random variables that assumes only $L_1$:

**Theorem 4 (L$_1$ iid SLLN)** Let $\{X_n\}$ be iid $L_1$ random variables with mean $\mu = E[X_n]$. Set $S_n = \sum_{j \leq n} X_j$ and $\bar{X}_n \equiv S_n/n = \frac{1}{n} \sum_{j \leq n} X_j$. Then:

$$\mathbb{P}[\bar{X}_n \to \mu] = 1.$$