7. Weak Law of Large Numbers

We are now in a position to establish various laws of large numbers, which give conditions for the arithmetic average of repeated observations to converge in certain senses. Among other things, these laws justify and formalize our intuitive notions of probability as representing some kind of measure of long-term relative frequency.

Convergence in L^p and Probability.

The weak law of large numbers is concerned with convergence in probability where

Definition. A sequence of random variables $X_1, X_2, ...$ is said to *converge to X in probability* if for every $\varepsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$. In this case, we write $X_n \to_p X$.

In analysis we would call this convergence in measure.

Note that if $X_n \to_p X$, then $\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$ for all $\varepsilon > 0$, while $X_n \to X$ a.s. implies that $P(\lim_{n\to\infty} |X_n - X| < \varepsilon) = 1$ for all $\varepsilon > 0$. The following proposition and example show the importance of the placement of the limit in the two definitions..

Proposition 7.1. *If $X_n \to X$ a.s., then $X_n \to_p X$.*

Proof. Let $\varepsilon > 0$ be given and define

$$A_n = \bigcup_{m \geq n} \{ |X_m - X| > \varepsilon \},$$
$$A = \bigcap_{n=1}^{\infty} A_n,$$
$$E = \left\{\omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right\}.$$

Since $A_1 \supseteq A_2 \supseteq ...$, continuity from above implies that $P(A) = \lim_{n\to\infty} P(A_n)$.

Now if $\omega \in A$, then for every $n \in \mathbb{N}$, there is an $m \geq n$ with $|X_m(\omega) - X(\omega)| > \varepsilon$, so $\lim_{n\to\infty} X_n(\omega) \neq X(\omega)$, and thus $A \subseteq E$.

Because we also have the inclusion $\{|X_n - X| > \varepsilon\} \subseteq A_n$, monotonicity implies that

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} P(A_n) = P(A) \leq P(E) = 0$$

where the final equality is the definition of almost sure convergence. □

Example 7.1 (Scanning Interval)**.** On the interval $[0, 1)$ with Lebesgue measure, define

$$X_1 = 1_{[0,1)}, X_2 = 1_{[0,\frac{1}{2})}, X_3 = 1_{[\frac{1}{2},1)}, ..., X_{2^n+k} = 1_{[\frac{k}{2^n},\frac{k+1}{2^n})}, ...$$

It is straight forward that $X_n \to_p 0$ (for any $\varepsilon \in (0, 1)$, $m \geq 2^n$ implies $P(|X_m - 0| > \varepsilon) \leq \frac{1}{2^n}$), but $\lim_{n\to\infty} X_n(\omega)$ does not exist for any ω (there are infinitely many values of n with $X_n(\omega) = 1$ and infinitely many values with $X_n(\omega) = 0$), thus $X_n \nrightarrow 0$ a.s.

The preceding shows that convergence in probability is weaker than almost sure convergence. In fact, this is the source of "weak" in the weak law of large numbers.

Our first set of weak laws make use of L^2 convergence where

Definition. For $p \in (0, \infty]$, a sequence of random variables $X_1, X_2, ...$ is said to *converge to* X *in* L^p if $\lim_{n\to\infty} \|X_n - X\|_p = 0$. (For $p \in (0, \infty)$, this is equivalent to $E[|X_n - X|^p] \to 0$.)

Our first observation about L^p convergence is

Proposition 7.2. *For any* $1 \le r < s \le \infty$, *if* $X_n \to X$ *in* L^s, *then* $X_n \to X$ *in* L^r.

Proof. If $X_n \to X$ in L^s, then Corollary 5.2 implies $\|X_n - X\|_r \le \|X_n - X\|_s \to 0$. □

To see how L^p convergence compares with our other notions of convergence, note that

Proposition 7.3. *If* $X_n \to X$ *in* L^p *for* $p > 0$, *then* $X_n \to_p X$.

Proof. For any $\varepsilon > 0$, Chebychev's inequality gives

$$P(|X_n - X| > \varepsilon) = P(|X_n - X|^p > \varepsilon^p) \le \varepsilon^{-p} E[|X_n - X|^p] \to 0.$$ □

Example 7.2. On the interval $[0,1]$ with Lebesgue measure, define a sequence of random variables by $X_n = n^{\frac{1}{p}} 1_{(0,n^{-1}]}$. Then $X_n \to 0$ a.s. (and thus in probability) since for all $\omega \in (0,1]$, $X_n(\omega) = 0$ whenever $n > \omega^{-1}$. However, $E[|X_n - 0|^p] = 1$ for all n, so $X_n \nrightarrow 0$ in L^p.

Proposition 7.3 and Example 7.2 show that L^p convergence is stronger than convergence in probability.

Example 7.2 also shows that almost sure convergence need not imply convergence in L^p
(unless one makes additional assumptions such as boundedness or uniform integrability).

Conversely, Example 7.1 shows that L^p convergence does not imply almost sure convergence.

It is perhaps worth noting that a.s. convergence and convergence in probability are preserved by continuous functions. (The latter claim can be shown directly from the ε-δ definition of continuity, but we will give an easier proof in Theorem 8.2.) However, L^p convergence need not be. For example, on $[0,1]$ with Lebesgue measure, $X_n = n^{\frac{1}{2}} 1_{(0,n^{-p})}$ converges to 0 in L^p, $p > 0$, but if $f(x) = x^2$, $\|f(X_n) - f(0)\|_p = 1$ for all n.

Now recall that random variables X and Y with finite second moments are said to be uncorrelated if $E[XY] = E[X]E[Y]$.

If we denote $E[X] = \mu_X$, $E[Y] = \mu_Y$, then the *covariance* of X and Y is defined as

$$\begin{aligned}\text{Cov}(X,Y) :&= E[(X - \mu_X)(Y - \mu_Y)] = E[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y] \\ &= E[XY] - 2\mu_X\mu_Y + \mu_X\mu_Y = E[XY] - E[X]E[Y],\end{aligned}$$

so uncorrelated is equivalent to zero covariance and finite second moments.

We say that a family of random variables $\{X_i\}_{i\in I}$ is uncorrelated if $\text{Cov}(X_i, X_j) = 0$ for all $i \ne j$.

Before stating our first weak law, we record the following simple observation about sums of uncorrelated random variables.

Lemma 7.1. *If $X_1, X_2, ..., X_n$ are uncorrelated, then*

$$\mathrm{Var}(X_1 + ... + X_n) = \mathrm{Var}(X_1) + ... + \mathrm{Var}(X_n).$$

Proof. Let $\mu_i = E[X_i]$ and $S_n = \sum_{i=1}^n X_i$. Then $E[S_n] = \sum_{i=1}^n \mu_i$ by linearity, so

$$\begin{aligned}\mathrm{Var}(S_n) = E\left[(S_n - E[S_n])^2\right] &= E\left[\left(\sum_{i=1}^n (X_i - \mu_i)\right)^2\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu_i)^2 + \sum_{i \neq j} (X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_{i=1}^n E\left[(X_i - \mu_i)^2\right] + \sum_{i \neq j} E\left[(X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^n \mathrm{Var}(X_i).\end{aligned}$$

□

We also observe that the for any $a, b \in \mathbb{R}$,

$$\mathrm{Var}(aX + b) = E\left[((aX + b) - (a\mu_X + b))^2\right] = a^2 E\left[(X - \mu_X)^2\right] = a^2 \mathrm{Var}(X).$$

With these results in hand, the L^2 weak law follows easily.

Theorem 7.1. *Let $X_1, X_2, ...$ be uncorrelated random variables with common mean $E[X_i] = \mu$ and uniformly bounded variance $\mathrm{Var}(X_i) \le C < \infty$. Writing $S_n = X_1 + ... + X_n$, we have that $\frac{1}{n} S_n \to \mu$ in L^2 and in probability.*

Proof. Since $E\left[\frac{1}{n} S_n\right] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$, we see that

$$E\left[\left(\frac{1}{n} S_n - \mu\right)^2\right] = \mathrm{Var}\left(\frac{1}{n} S_n\right) = \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}(X_i) \le \frac{nC}{n^2} \to 0$$

as $n \to \infty$, hence $\frac{1}{n} S_n \to \mu$ in L^2. By Proposition 7.3, $\frac{1}{n} S_n \to_p \mu$ as well. □

Specializing to the case where the $X_i's$ are *independent and identically distributed* (or *i.i.d.*), we have the oft-quoted weak law

Corollary 7.1. *If $X_1, X_2, ...$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, then $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ converges in probability to μ.*

The statistical interpretation of Corollary 7.1 is that under mild conditions, if the sample size is sufficiently large, then the sample mean will be close to the population mean with high probability.

Examples.

Our first applications of these ideas involve statements that appear to be unrelated to probability.

Example 7.3. Let $f : [0,1] \to \mathbb{R}$ be a continuous function and let

$$f_n = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f\left(\frac{k}{n}\right)$$

be the *Bernstein polynomial of degree n associated with f*. Then $\lim_{n\to\infty} \sup_{x\in[0,1]} |f_n(x) - f(x)| = 0$.

Proof.

Given any $p \in [0,1]$, let $X_1, X_2, ...$ be i.i.d. with $P(X_1 = 1) = p$ and $P(X_1 = 0) = 1-p$.

One easily calculates $E[X_1] = p$ and $\mathrm{Var}(X_1) = p(1-p) \le \frac{1}{4}$.

Letting $S_n = \sum_{i=1}^n X_i$, we have that $P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$, so $E\left[f\left(\frac{1}{n}S_n\right)\right] = f_n(p)$.

Also, Corollary 7.1 shows that $\overline{X}_n = \frac{1}{n}S_n$ converges to p in probability.

To establish the desired result, we have to appeal to the proof of our weak law.

First, for any $\alpha > 0$, Chebychev's inequality and the fact that $E\left[\overline{X}_n\right] = p$, $\mathrm{Var}\left(\overline{X}_n\right) = \frac{p(1-p)}{n} < \frac{1}{4n}$ gives

$$P\left(\left|\overline{X}_n - p\right| \ge \alpha\right) \le \frac{\mathrm{Var}(\overline{X}_n)}{\alpha^2} \le \frac{1}{4n\alpha^2}.$$

Now since f is continuous on the compact set $[0,1]$ it is uniformly continuous and uniformly bounded. Let $M = \sup_{x\in[0,1]} |f(x)|$, and for a given $\varepsilon > 0$, let $\delta > 0$ be such that $|x-y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$ for all $x, y \in [0,1]$. Since the absolute value function is convex, Jensen's inequality yields

$$\left|E\left[f\left(\overline{X}_n\right) - f(p)\right]\right| \le E\left|f\left(\overline{X}_n\right) - f(p)\right| \le \varepsilon P\left(\left|\overline{X}_n - p\right| < \delta\right) + 2MP\left(\left|\overline{X}_n - p\right| \ge \delta\right) \le \varepsilon + \frac{M}{2n\delta^2}.$$

As this does not depend on p, the result follows upon letting $n \to \infty$. □

Our next amusing result can be interpreted as saying that a high-dimensional cube is almost a sphere.

Example 7.4. Let $X_1, X_2, ...$ be independent and uniformly distributed on $[-1,1]$. Then $X_1^2, X_2^2, ...$ are also independent with $E[X_i^2] = \int_{-1}^1 \frac{x^2}{2} dx = \frac{1}{3}$ and $\mathrm{Var}(X_i^2) \le E[X_i^4] \le 1$, so Corollary 7.1 shows that $\frac{1}{n}\sum_{i=1}^n X_i^2$ converges to $\frac{1}{3}$ in probability.

Now given $\varepsilon \in (0,1)$, write $A_{n,\varepsilon} = \left\{x \in \mathbb{R}^n : (1-\varepsilon)\sqrt{\frac{n}{3}} \le \|x\| \le (1+\varepsilon)\sqrt{\frac{n}{3}}\right\}$ where $\|x\| = (x_1^2 + ... + x_n^2)^{\frac{1}{2}}$ is the usual Euclidean distance, and let m denote Lebesgue measure. We have

$$\begin{aligned}
\frac{m\left(A_{n,\varepsilon} \cap [-1,1]^n\right)}{2^n} &= P\left((X_1, ..., X_n) \in A_{n,\varepsilon}\right) = P\left((1-\varepsilon)\sqrt{\frac{n}{3}} \le \sqrt{\sum_{i=1}^n X_i^2} \le (1+\varepsilon)\sqrt{\frac{n}{3}}\right) \\
&= P\left(\frac{1}{3}(1 - 2\varepsilon + \varepsilon^2) \le \frac{1}{n}\sum_{i=1}^n X_i^2 \le \frac{1}{3}(1 + 2\varepsilon + \varepsilon^2)\right) \\
&\ge P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{3}\right| \le \frac{2\varepsilon - \varepsilon^2}{3}\right),
\end{aligned}$$

so that $\frac{m(A_{n,\varepsilon}\cap[-1,1]^n)}{2^n} \to 1$ as $n \to \infty$. In words, most of the volume of the cube $[-1,1]^n$ comes from $A_{n,\varepsilon}$, which is almost the boundary of the ball centered at the origin with radius $\sqrt{\frac{n}{3}}$.

The next set of examples concern the limiting behavior of row sums of *triangular arrays*, for which we appeal to the following easy generalization of Theorem 7.1.

Theorem 7.2. *Given a triangular array of integrable random variables, $\{X_{n,k}\}_{n\in\mathbb{N},1\leq k\leq n}$, let $S_n = \sum_{k=1}^n X_{n,k}$ denote the nth row sum, and write $\mu_n = E[S_n]$, $\sigma_n^2 = \mathrm{Var}(S_n)$. If the sequence $\{b_n\}_{n=1}^\infty$ satisfies $\lim_{n\to\infty} \frac{\sigma_n^2}{b_n^2} = 0$, then*

$$\frac{S_n - \mu_n}{b_n} \to_p 0.$$

Proof. By assumption, $E\left[\left(\frac{S_n-\mu_n}{b_n}\right)^2\right] = \frac{\mathrm{Var}(S_n)}{b_n^2} \to 0$ as $n \to \infty$, so the result follows since L^2 convergence implies convergence in probability. □

Example 7.5 (Coupon Collector's Problem)**.** Suppose that there are n distinct types of coupons and each time one obtains a coupon it is, independent of prior selections, equally likely to be any one of the types. We are interested in the number of draws needed to obtain a complete set. To this end, let $T_{n,k}$ denote the number of draws needed to collect k distinct types for $k = 1, ..., n$ and note that $T_{n,1} = 1$. Set $X_{n,1} = 1$ and $X_{n,k} = T_{n,k} - T_{n,k-1}$ for $k = 2, ..., n$ so that $X_{n,k}$ is the number of trials needed to obtain a type different from the first $k - 1$. The number of draws needed to obtain a complete set is given by

$$T_n := T_{n,n} = 1 + \sum_{k=2}^n (T_{n,k} - T_{n,k-1}) = 1 + \sum_{k=2}^n X_{n,k}.$$

By construction, $X_{n,2}, ..., X_{n,n}$ are independent with $P(X_{n,k} = m) = \left(\frac{n-k+1}{n}\right)\left(\frac{k-1}{n}\right)^{m-1}$ for $m \in \mathbb{N}$.

Now a random variable X with $P(X = m) = p(1 - p)^{m-1}$ is said to be *geometric with success probability p*. A little calculus gives

$$\begin{aligned} E[X] &= \sum_{m=1}^\infty mp(1-p)^{m-1} = p\sum_{m=1}^\infty -\frac{d}{dp}(1-p)^m \\ &= -p\frac{d}{dp}\sum_{m=1}^\infty (1-p)^m = -p\frac{d}{dp}\frac{1-p}{p} = \frac{1}{p} \end{aligned}$$

and

$$\begin{aligned} E[X^2] &= \sum_{m=1}^\infty m^2 p(1-p)^{m-1} = \sum_{m=1}^\infty [m(m-1)+m]p(1-p)^{m-1} \\ &= p(1-p)\sum_{m=1}^\infty m(m-1)(1-p)^{m-2} + \sum_{m=1}^\infty mp(1-p)^{m-1} \\ &= p(1-p)\sum_{m=2}^\infty \frac{d^2}{dp^2}(1-p)^m + E[X] = p(1-p)\frac{d^2}{dp^2}\frac{(1-p)^2}{p} + \frac{1}{p} \\ &= \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2}, \end{aligned}$$

hence

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{1-p}{p^2} \leq \frac{1}{p^2}.$$

It follows that

$$E[T_n] = 1 + \sum_{k=2}^{n} E[X_{n,k}] = 1 + \sum_{k=2}^{n} \frac{n}{n-k+1} = 1 + n\sum_{j=1}^{n-1}\frac{1}{j} = n\sum_{j=1}^{n}\frac{1}{j}$$

and

$$\operatorname{Var}(T_n) = \sum_{k=2}^{n}\operatorname{Var}(X_{n,k}) \le \sum_{k=2}^{n}\left(\frac{n}{n-k+1}\right)^2 = n^2\sum_{j=1}^{n-1}\frac{1}{j^2} \le n^2\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{\pi^2 n^2}{6}.$$

Taking $b_n = n\log(n)$ we have $\frac{\operatorname{Var}(T_n)}{b_n^2} \le \frac{\pi^2}{6\log(n)^2} \to 0$, so Theorem 7.2 implies $\frac{T_n - n\sum_{k=1}^{n} k^{-1}}{n\log(n)} \to_p 0$.
Using the inequality

$$\log(n) \le \sum_{k=1}^{n}\frac{1}{k} \le \log(n) + 1$$

(which can be seen by bounding $\log(n) = \int_1^n \frac{dx}{x}$ with the upper Riemann sum $\sum_{k=1}^{n-1}\frac{1}{k} \le \sum_{k=1}^{n}\frac{1}{k}$ and the lower Riemann sum $\sum_{k=2}^{n}\frac{1}{k} = \sum_{k=1}^{n}\frac{1}{k} - 1$), we conclude that $\dfrac{T_n}{n\log(n)} \to_p 1$.

Example 7.6 (Occupancy Problem)**.** Suppose that we drop r_n balls at random into n bins where $\dfrac{r_n}{n} \to c$.
Letting $X_{n,k} = 1\{\text{bin } k \text{ is empty}\}$, the number of empty bins is $X_n = \sum_{k=1}^{n} X_{n,k}$.
It is clear that

$$E[X_n] = \sum_{k=1}^{n} E[X_{n,k}] = \sum_{k=1}^{n} P(\text{bin } k \text{ is empty}) = n\left(\frac{n-1}{n}\right)^{r_n}$$

and

$$\begin{aligned} E[X_n^2] &= E\left[\sum_{k=1}^{n} X_{n,k}^2 + 2\sum_{i<j} X_{n,i}X_{n,j}\right] = \sum_{k=1}^{n} E[X_{n,k}] + 2\sum_{i<j} E[X_{n,i}X_{n,j}] \\ &= \sum_{k=1}^{n} P(\text{bin } k \text{ is empty}) + 2\sum_{i<j} P(\text{bins } i \text{ and } j \text{ are empty}) \\ &= n\left(\frac{n-1}{n}\right)^{r_n} + 2\binom{n}{2}\left(\frac{n-2}{n}\right)^{r_n} = n\left(1-\frac{1}{n}\right)^{r_n} + n(n-1)\left(1-\frac{2}{n}\right)^{r_n}, \end{aligned}$$

so

$$\operatorname{Var}(X_n) = E[X_n^2] - E[X_n]^2 = n\left(1-\frac{1}{n}\right)^{r_n} + n(n-1)\left(1-\frac{2}{n}\right)^{r_n} - n^2\left(1-\frac{1}{n}\right)^{2r_n}.$$

Now L'Hospital's rule gives $\lim_{n\to\infty}\dfrac{\log\left(\frac{n-1}{n}\right)}{n^{-1}} = \lim_{n\to\infty}\dfrac{n^{-2}}{-n^{-2}}\cdot\dfrac{n}{n-1} = -1$, so, since $\dfrac{r_n}{n} \to c$, we have that $\log\left[\left(\dfrac{n-1}{n}\right)^{r_n}\right] = \dfrac{r_n}{n}\cdot\dfrac{\log\left(\frac{n-1}{n}\right)}{n^{-1}} \to -c$ and thus $\left(\dfrac{n-1}{n}\right)^{r_n} \to e^{-c}$ as $n \to \infty$.
Similarly, $\left(1-\dfrac{2}{n}\right)^{r_n}, \left(1-\dfrac{1}{n}\right)^{2r_n} \to e^{-2c}$.
Consequently, $\dfrac{E[X_n]}{n} = \left(\dfrac{n-1}{n}\right)^{r_n} \to e^{-c}$ and

$$\frac{\operatorname{Var}(X_n)}{n^2} = \frac{\left(1-\frac{1}{n}\right)^{r_n}}{n} + \frac{n(n-1)}{n}\left(1-\frac{2}{n}\right)^{r_n} - \left(1-\frac{1}{n}\right)^{2r_n} \to 0 + 1\cdot e^{-2c} - e^{-2c} = 0$$

as $n \to \infty$, so taking $b_n = n$ in Theorem 7.2 shows that the proportion of empty bins, $\dfrac{X_n}{n}$, converges to e^{-c} in probability.

Weak Law of Large Numbers.

We begin by providing a simple analysis proof of the weak law in its classical form. The general trick is to use truncation in order to consider cases where we have control over the size and the probability, respectively.

Theorem 7.3. *Suppose that $X_1, X_2, \ldots$ are i.i.d. with $E|X_1| < \infty$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu = E[X_1]$. Then $\frac{1}{n}S_n \to \mu$ in probability.*

Proof.

In what follows, the arithmetic average of the first n terms of a sequence of random variables $Y_1, Y_2, \ldots$ will be denoted by $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.

We first note that, by replacing X_i with $X_i' = X_i - \mu$ if necessary, we may suppose without loss of generality that $E[X_i] = 0$.

Thus we need to show that for given $\varepsilon, \delta > 0$, there is an $N \in \mathbb{N}$ such that $P\left(\left|\overline{X}_n\right| > \varepsilon\right) < \delta$ whenever $n \geq N$.

To this end, we pick $C < \infty$ large enough that $E\left[|X_1| 1\{|X_1| > C\}\right] < \eta$ for some η to be determined.

(This is possible since $|X_1| 1\{|X_1| \leq n\} \leq |X_1|$ and $E|X_1| < \infty$, so $\lim_{n\to\infty} E\left[|X_1| 1\{|X_1| \leq n\}\right] = E|X_1|$ by the dominated convergence theorem, hence $E\left[|X_1| 1\{|X_1| > n\}\right] = E|X_1| - E\left[|X_1| 1\{|X_1| \leq n\}\right] \to 0$.)

Now define

$$W_i = X_i 1\{|X_i| \leq C\} - E\left[X_i 1\{|X_i| \leq C\}\right]$$
$$Z_i = X_i 1\{|X_i| > C\} - E\left[X_i 1\{|X_i| > C\}\right].$$

By assumption, we have that

$$E|Z_i| \leq 2E\left[|X_1| 1\{|X_i| > C\}\right] < 2\eta,$$

and thus, for every $n \in \mathbb{N}$,

$$E\left|\overline{Z}_n\right| = E\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| \leq \frac{1}{n}\sum_{i=1}^n E|Z_i| \leq 2\eta.$$

Also, the $W_i's$ are i.i.d. with mean zero and satisfy $|W_i| \leq 2C$ by construction, so

$$E\left[\overline{W}_n^2\right] = \frac{1}{n^2}\left(\sum_{i=1}^n E[W_i^2] + \sum_{i\neq j} E[W_i W_j]\right) = \frac{E[W_1^2]}{n} \leq \frac{4C^2}{n},$$

and thus

$$E\left[\left|\overline{W}_n\right|\right]^2 \leq E\left[\overline{W}_n^2\right] \leq \frac{4C^2}{n}$$

by Jensen's inequality.

Consequently, if $n \geq N := \left\lceil \frac{4C^2}{\eta^2} \right\rceil$, then $E\left|\overline{W}_n\right| \leq \eta$.

Finally, Chebychev's inequality and the fact that

$$\left|\overline{X}_n\right| = \left|\overline{W}_n + \overline{Z}_n\right| \leq \left|\overline{W}_n\right| + \left|\overline{Z}_n\right|$$

imply that for $n \geq N$,

$$P\left(\left|\overline{X}_n\right| > \varepsilon\right) \leq P\left(\left|\overline{W}_n\right| + \left|\overline{Z}_n\right| > \varepsilon\right) \leq \frac{E\left|\overline{W}_n\right| + E\left|\overline{Z}_n\right|}{\varepsilon} < \frac{3\eta}{\varepsilon}.$$

Taking $\eta = \frac{\varepsilon\delta}{3}$ completes the proof. □

We now turn to a weak law for triangular arrays which can be useful even in situations involving infinite means.

Theorem 7.4. *For each $n \in \mathbb{N}$, let $X_{n,1}, ..., X_{n,n}$ be independent. Let $\{b_n\}_{n=1}^{\infty}$ be a sequence of positive numbers with $\lim_{n\to\infty} b_n = \infty$ and let $\widetilde{X}_{n,k} = X_{n,k} 1\{|X_{n,k}| \le b_n\}$. Suppose that as $n \to \infty$*

(1) $\sum_{k=1}^{n} P(|X_{n,k}| > b_n) \to 0$
(2) $b_n^{-2} \sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}^2\right] \to 0$.

If we let $S_n = \sum_{k=1}^{n} X_{n,k}$ and $a_n = \sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}\right]$, then $\dfrac{S_n - a_n}{b_n} \to_p 0$.

Proof. Let $\widetilde{S}_n = \sum_{k=1}^{n} \widetilde{X}_{n,k}$. By partitioning the event $\left\{\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right\}$ according to whether or not $S_n = \widetilde{S}_n$, we see that

$$P\left(\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right) \le P(S_n \ne \widetilde{S}_n) + P\left(\left|\frac{\widetilde{S}_n - a_n}{b_n}\right| > \varepsilon\right).$$

To estimate the first term, we observe that

$$P(S_n \ne \widetilde{S}_n) \le P\left(\bigcup_{k=1}^{n} \{X_{n,k} \ne \widetilde{X}_{n,k}\}\right) \le \sum_{k=1}^{n} P\left(X_{n,k} \ne \widetilde{X}_{n,k}\right) = \sum_{k=1}^{n} P(|X_{n,k}| > b_n) \to 0$$

where the first inequality is due to the fact that $S_n \ne \widetilde{S}_n$ implies that there is some $k \in [n]$ with $X_{n,k} \ne \widetilde{X}_{n,k}$, and the second inequality is countable subadditivity.

For the second term, we use Chebychev's inequality, $E\left[\widetilde{S}_n\right] = a_n$, the independence of the $\widetilde{X}'_{n,k}s$, and our second assumption to obtain

$$P\left(\left|\frac{\widetilde{S}_n - a_n}{b_n}\right| > \varepsilon\right) \le \varepsilon^{-2} E\left[\left(\frac{\widetilde{S}_n - a_n}{b_n}\right)^2\right] = \varepsilon^{-2} b_n^{-2} \mathrm{Var}(\widetilde{S}_n)$$

$$= \varepsilon^{-2} b_n^{-2} \sum_{k=1}^{n} \mathrm{Var}\left[\widetilde{X}_{n,k}^2\right] \le \varepsilon^{-2}\left(b_n^{-2} \sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}^2\right]\right) \to 0. \qquad \square$$

Theorem 7.4 was so easy to prove because we assumed exactly what we needed. Essentially, these are the correct hypotheses for the weak law, but they are a little clunky so we usually talk about special cases that take a nicer form.

In order to prove our weak law for sequences of i.i.d. random variables, we need the following simple lemma.

Lemma 7.2 (Layer cake representation)**.** *If $Y \ge 0$ and $p > 0$, then*

$$E[Y^p] = \int_0^{\infty} p y^{p-1} P(Y > y) dy.$$

Proof. Tonelli's theorem gives

$$\int_0^{\infty} p y^{p-1} P(Y > y) dy = \int_0^{\infty} p y^{p-1} \left(\int_{\Omega} 1\{Y > y\} dP\right) dy$$

$$= \int_{\Omega} \left(\int_0^{\infty} p y^{p-1} 1\{y < Y\} dy\right) dP$$

$$= \int_{\Omega} \left(\int_0^{Y} p y^{p-1} dy\right) dP = \int_{\Omega} Y^p dP = E[Y^p]. \qquad \square$$

We now have all the necessary ingredients for

Theorem 7.5 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with*

$$xP(|X_1| > x) \to 0 \text{ as } x \to \infty.$$

Let $S_n = X_1 + \ldots + X_n$ and $\mu_n = E\left[X_1 1\{|X_1| \le n\}\right]$. Then $\frac{1}{n}S_n - \mu_n \to 0$ in probability.

Proof. We will apply Theorem 7.4 with $X_{n,k} = X_k$ and $b_n = n$ (hence $a_n = n\mu_n$).

The first assumption is satisfied since

$$\sum_{k=1}^{n} P\left(|X_{n,k}| > n\right) = nP(|X_1| > n) \to 0.$$

For the second assumption, we have $\widetilde{X}_{n,k} = X_k 1\{|X_k| \le n\}$, so we must show that

$$\frac{1}{n}E\left[\widetilde{X}_{n,1}^2\right] = \frac{1}{n^2}\sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}^2\right] \to 0.$$

Lemma 7.2 shows that

$$E\left[\widetilde{X}_{n,1}^2\right] = \int_0^\infty 2yP\left(\left|\widetilde{X}_{n,1}\right| > y\right)dy \le \int_0^n 2yP\left(|X_1| > y\right)dy$$

since $P\left(\left|\widetilde{X}_{n,1}\right| > y\right) = 0$ for $y > n$ and $P\left(\left|\widetilde{X}_{n,1}\right| > y\right) = P\left(|X_1| > y\right) - P\left(|X_1| > n\right)$ for $y \le n$, so we will be done once we prove that

$$\frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right)dy \to 0.$$

To see that this is the case, note that since $2yP\left(|X_1| > y\right) \to 0$ as $y \to \infty$, for any $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that $2yP\left(|X_1| > y\right) < \varepsilon$ whenever $y \ge N$. Because $2yP\left(|X_1| > y\right) < 2N$ for $y < N$, we see that for all $n > N$,

$$\begin{aligned}\frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right)dy &= \frac{1}{n}\int_0^N 2yP\left(|X_1| > y\right)dy + \frac{1}{n}\int_N^n 2yP\left(|X_1| > y\right)dy \\ &\le \frac{1}{n}\int_0^N 2Ndy + \frac{1}{n}\int_N^n \varepsilon dy = \frac{2N^2}{n} + \frac{n-N}{n}\varepsilon,\end{aligned}$$

hence

$$\limsup_{n\to\infty}\frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right)dy \le \limsup_{n\to\infty}\frac{2N^2}{n} + \frac{n-N}{n}\varepsilon = \varepsilon,$$

and the result follows since ε was arbitrary. □

Remark. Theorem 7.5 implies Theorem 7.3 since if $E|X_1| < \infty$, then the dominated convergence theorem gives

$$\mu_n = E[X_1 1\{|X_1| \le n\}] \to E[X_1] = \mu \text{ as } n \to \infty,$$

$$xP\left(|X_1| > x\right) \le E\left[|X_1| 1\{|X_1| > x\}\right] \to 0 \text{ as } x \to \infty.$$

On the other hand, the improvement is not vast since $xP(|X_1| > x) \to 0$ implies that there is $M \in \mathbb{N}$ so that $xP(|X_1| > x) \leq 1$ for $x \geq M$, and thus for any $\varepsilon > 0$, Lemma 7.2 with $p = 1 - \varepsilon$ yields

$$E\left[|X_1|^{1-\varepsilon}\right] = \int_0^\infty (1-\varepsilon)y^{-\varepsilon}P(|X_1| > y)\,dy = (1-\varepsilon)\int_0^\infty y^{-(1+\varepsilon)} \cdot yP(|X_1| > y)\,dy$$
$$\leq (1-\varepsilon)\int_0^M y^{-(1+\varepsilon)}M\,dy + (1-\varepsilon)\int_M^\infty y^{-(1+\varepsilon)}dy < \infty.$$

Example 7.7 (The St. Petersburg Paradox)**.** Suppose that I offered to pay you 2^j dollars if it takes j flips of a fair coin for the first head to appear. That is, your winnings are given by the random variable X with $P(X = 2^j) = 2^{-j}$ for $j \in \mathbb{N}$. How much would you pay to play the game n times? The paradox is that $E[X] = \sum_{j=1}^\infty 2^j \cdot 2^{-j} = \infty$, but most sensible people would not pay anywhere near \$40 a game.

Using Theorem 7.4, we will show that a fair price for playing n times is $\$\log_2(n)$ per play, so that one would need to play about a trillion rounds to reasonably expect to break even at \$40 a play.

Proof. To cast this problem in terms of Theorem 7.4, we will take $X_1, X_2, \ldots$ to be independent random variables which are equal in distribution to X and set $X_{n,k} = X_k$. Then $S_n = \sum_{k=1}^n X_k$ denotes your total winnings after n games. We need to choose b_n so that

$$nP(X > b_n) = \sum_{k=1}^n P(X_{n,k} > b_n) \to 0,$$
$$\frac{n}{b_n^2}E\left[X^2 1\{X \leq b_n\}\right] = b_n^{-2}\sum_{k=1}^n E\left[(X_{n,k}1\{|X_{n,k}| \leq b_n\})^2\right] \to 0.$$

To this end, let $m(n) = \log_2(n) + K(n)$ where $K(n)$ is such that $m(n) \in \mathbb{N}$ and $K(n) \to \infty$ as $n \to \infty$. If we set $b_n = 2^{m(n)} = n2^{K(n)}$, we have

$$nP(X > b_n) \leq nP(X \geq b_n) = n\sum_{i=m(n)}^\infty 2^{-i} = n2^{-m(n)+1} = 2^{-K(n)+1} \to 0$$

and

$$E\left[X^2 1\{X \leq b_n\}\right] = \sum_{i=1}^{m(n)} 2^{2i} \cdot 2^{-i} = 2^{m(n)+1} - 2 \leq 2b_n,$$

so that

$$\frac{n}{b_n^2}E\left[X^2 1\{|X| \leq b_n\}\right] \leq \frac{2n}{b_n} = 2^{-K(n)+1} \to 0.$$

Since

$$a_n = \sum_{k=1}^n E\left[X_{n,k}1\{|X_{n,k}| \leq b_n\}\right] = nE\left[X1\{X \leq b_n\}\right] = n\sum_{i=1}^{m(n)} 2^i \cdot 2^{-i} = nm(n),$$

Theorem 7.4 gives

$$\frac{S_n - n\log_2(n) - nK(n)}{n2^{K(n)}} \to_p 0.$$

If we take $K(n) \leq \log_2(\log_2(n))$, then the conclusion holds with $n\log_2(n)$ in the denominator, so we get

$$\frac{S_n}{n\log_2(n)} \to_p 1.$$

□