Probability Theory II

These notes begin with a brief discussion of independence, and then discuss the three main foundational theorems of probability theory: the weak law of large numbers, the strong law of large numbers, and the central limit theorem. Though we have included a detailed proof of the weak law in Section 2, we omit many of the proofs in Sections 3 and 4.

Independence

Consider an experiment where we flip a coin twice. We begin by flipping once, and the coin comes up heads. How will this outcome affect the second flip?

The answer, of course, is that it doesn't. The second flip is completely independent from the first one. This idea is captured by the following definition:

Definition: Independent Events
Let $(\Omega, \mathcal{E}, P)$ be a probability space. Two events $A, B \subset \Omega$ are **independent** if

$$P(A \cap B) \;=\; P(A)\, P(B).$$

This definition can be phrased in terms of conditional probabilities. If A and B are events and $P(B) \neq 0$, the **probability of A given B** is

$$P(A \text{ given } B) \;=\; \frac{P(A \cap B)}{P(B)}.$$

This represents the probability that A occurs, given the information that B occurs. Using this formula, the definition of independence can be rewritten as

$$P(A \text{ given } B) \;=\; P(A).$$

That is, A and B are independent if the information that B occurs does not affect the probability of A.

The definition of independence can be generalized to more than two events:

Definition: Multiple Independent Events
Events $\{E_n\}$ are **independent** if

$$P(E_{i_1} \cap \cdots \cap E_{i_k}) \;=\; P(E_{i_1}) \cdots P(E_{i_k})$$

for all $i_1 < \cdots < i_k$.

Note that the following statements are different:

1. The events $\{E_n\}$ are independent.
2. E_i and E_j are independent for all $i \neq j$.

That is, independence for multiple events is *not* the same thing as pairwise independence. The following example illustrates this.

EXAMPLE 1 Three Pairwise Independent Events
Consider the following three events for a pair of coin flips:

$E1$: The first coin shows heads.

$E2$: The second coin shows heads.

$E3$: The two coins show the same result.

Each of these events has probability 1/2, and any two of these events are independent. However, all three events together are not independent. In particular,

$$P(E_1 \cap E_2 \cap E_3) \;=\; \frac{1}{4} \;\neq\; P(E_1)\, P(E_2)\, P(E_3). \qquad \blacksquare$$

The notion of independence can also be defined for random variables. Roughly speaking, two random variables are independent if knowledge about the value of the first variable has no effect on the value of the second variable. The following definition formalizes this notion:

Definition: Independent Random Variables
Let $X\colon \Omega \to S$ and $Y\colon \Omega \to T$ be random variables. We say that X and Y are **independent** if

$$P(X \in A \text{ and } Y \in B) \;=\; P(X \in A)\, P(Y \in B)$$

for all measurable subsets $A \subset S$ and $B \subset T$.

More generally, a sequence $\{X_1, X_2, X_3, \ldots\}$ of random variables is independent if

$$P\big(X_i \in A_i \text{ for each } i \in \{1, \ldots, n\}\big) \;=\; \prod_{i=1}^{n} P(X_i \in A_i)$$

for any $n \in \mathbb{N}$ and any finite sequence $A_1, \ldots, A_n$ of measurable sets.

Proposition 1 Functions Preserve Independence

Let $X\colon \Omega \to S$ and $Y\colon \Omega \to T$ be random variables, and let $f\colon S \to S'$ and $g\colon T \to T'$ be measurable functions. If X and Y are independent, then $f(X)$ and $g(Y)$ are independent as well.

PROOF Let $A \subset S'$ and $B \subset T'$ be measurable. Then

$$P\big(f(X) \in A \text{ and } g(Y) \in B\big) \;=\; P\big(X \in f^{-1}(A) \text{ and } Y \in g^{-1}(B)\big)$$

Since X and Y are independent, we can rewrite the quantity on the right to give

$$\begin{aligned} P\big(f(X) \in A \text{ and } g(Y) \in B\big) \;&=\; P\big(X \in f^{-1}(A)\big)\, P\big(Y \in g^{-1}(B)\big) \\ &=\; P\big(f(X) \in A\big)\, P\big(g(Y) \in B\big). \end{aligned}$$

■

It is possible to express the criterion for independence in terms of distributions. If $X\colon \Omega \to S$ and $Y\colon \Omega \to T$ are random variables, the **joint variable** (X, Y) is the Cartesian product $(X, Y)\colon \Omega \to S \times T$. The probability distribution $P_{(X,Y)}$ for (X, Y) is called the **joint distribution**.

Using these definitions, two random variables X and Y are independent if and only if

$$P_{(X,Y)}(A \times B) \;=\; P_X(A)\, P_Y(B)$$

for all measurable subsets $A \subset S$ and $B \subset T$. That is, X and Y are independent if the joint distribution $P_{(X,Y)}$ is the product of the measures P_X and P_Y. We use this criterion to prove the following theorem:

Proposition 2 Expectation of a Product

Let $X, Y\colon \Omega \to \mathbb{R}$ be independent random variables with finite expected values. Then

$$E[XY] \;=\; (EX)(EY).$$

PROOF Observe that

$$\int_{\mathbb{R}}\int_{\mathbb{R}} |xy|\, dP_X(x)\, dP_Y(y) \;=\; \left(\int_{\mathbb{R}} |x|\, dP_X(x)\right)\left(\int_{\mathbb{R}} |y|\, dP_Y(y)\right)$$

$$= E|X|\, E|Y| \;<\; \infty.$$

That is, the function $f(x, y) = xy$ is L^1 with respect to the measure $P_{(X,Y)}$. Therefore, by Fubini's theorem

$$E[XY] \;=\; \int_{\mathbb{R}^2} xy\, dP_{(X,Y)}(x, y) \;=\; \int_{\mathbb{R}}\int_{\mathbb{R}} xy\, dP_X(x)\, dP_Y(y)$$

$$= \left(\int_{\mathbb{R}} x\, dP_X(x)\right)\left(\int_{\mathbb{R}} y\, dP_Y(y)\right) \;=\; (EX)(EY). \qquad \blacksquare$$

This theorem has the following consequence:

Proposition 3 Variance of a Sum

Let $X, Y : \Omega \to \mathbb{R}$ be independent random variables with finite expected values. Then

$$\mathrm{Var}(X + Y) \;=\; \mathrm{Var}(X) + \mathrm{Var}(Y).$$

PROOF Let $X_0 = X - EX$ and $Y_0 = Y - EY$, and note that X_0 and Y_0 are independent. Then

$$\mathrm{Var}(X + Y) \;=\; E\left[(X_0 + Y_0)^2\right] \;=\; E[X_0^2] + 2\, E[X_0 Y_0] + E[Y_0^2].$$

But $E[X_0 Y_0] = (EX_0)(EY_0) = (0)(0) = 0$ by the previous theorem, so

$$\mathrm{Var}(X + Y) \;=\; E[X_0^2] + E[Y_0^2] \;=\; \mathrm{Var}(X) + \mathrm{Var}(Y). \qquad \blacksquare$$

The above formula can be generalized to the sum of any number of independent random variables. Specifically, if $\{X_n\}$ is a sequence of independent random variables, then

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

In particular, if all of the variables X_i have the same variance σ^2, then the sum $X_1 + \cdots + X_n$ has variance $\sigma^2 n$, and therefore has standard deviation $\sigma\sqrt{n}$.

Weak Law of Large Numbers

Suppose we perform the same experiment several times, generating a sequence $\{X_n\}$ of random variables. For example, we might roll a die repeatedly, writing down the result each time. In this case, each iteration of the experiment is called a **trial**, and the resulting random variables $\{X_n\}$ will have the following properties:

1. They will all be independent.
2. They will be **identically distributed**, i.e. all the X_n's will have the same distribution.

In probability textbooks, the phrase "independent and identically distributed" is so commonplace that it is sometimes abbreviated "i.i.d." (We will not follow this practice.)

If $\{X_n\}$ is a sequence of independent, identically distributed random variables, the **sample mean** $\overline{X}_n$ is the average value of the first n results:

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

It is a basic tenet of probability theory that the sample mean $\overline{X}_n$ should approach the mean μ as $n \to \infty$. This principle is known as the law of large numbers:

The Law of Large Numbers

Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite mean μ, and let

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then $\overline{X}_n$ should approach μ as $n \to \infty$.

For example, Figure 1 shows the sample means $\overline{X}_n$ for a sequence of 100,000 die rolls. As you might expect, the samples means for the trials approach 3.5, which is the expected value of a single die roll.

Unfortunately, the law of large numbers stated above is not precise. In particular, the word "approach" is ambiguous—in what sense must the random variables $\overline{X}_n$ approach the mean μ? This must involve some notion of convergence of random variables, but we have not been clear about which notion of convergence we intend. In fact, several different notions of convergence are possible, which leads to several different versions of the law of large numbers.

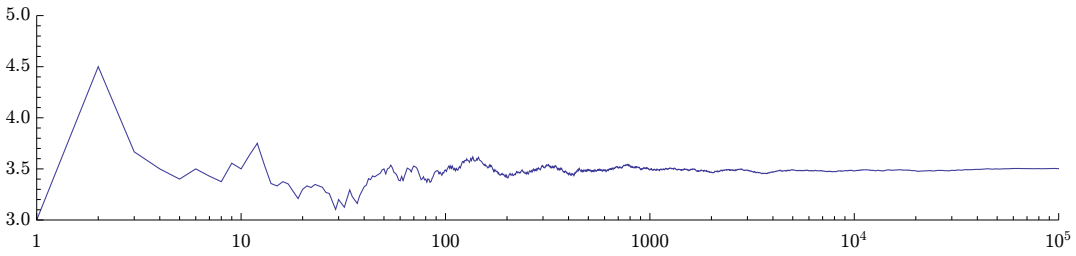


Figure 1: A logarithmic plot showing the sample means for 100,000 die rolls.

In this section, our goal is to prove a version of this law known as the weak law of large numbers. This involves the following notion of convergence:

Definition: Convergence in Probability
Let $\{X_n\}$ be a sequence of random variables, and let X be a random variable. We say that $X_n \to X$ **in probability** if for every $\epsilon > 0$,

$$P\big(|X_n - X| > \epsilon\big) \to 0 \qquad \text{as } n \to \infty.$$

We will spend the remainder of the section proving the following theorem:

Weak Law of Large Numbers

Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite expected value μ. For each n, let

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then $\overline{X}_n \to \mu$ in probability as $n \to \infty$.

To prove this theorem, we must find some bound on $P\big(|\overline{X}_n - \mu| \geq \epsilon\big)$ that goes to zero as $n \to \infty$. We shall use the following two inequalities:

Theorem 4 Markov's Inequality

Let X be a random variable with $E|X| < \infty$, and let $a \in (0, \infty)$. Then

$$P(|X| > a) \;\leq\; \frac{E|X|}{a}.$$

PROOF We may assume that X is nonnegative, so that $|X| = X$. Then

$$\begin{aligned} EX \;&=\; \int_{[0,\infty)} x\, dP_X(x) \;\geq\; \int_{(a,\infty)} x\, dP_X(x) \\ &\geq\; \int_{(a,\infty)} a\, dP_X \;=\; a\, P_X\big((a,\infty)\big) \;=\; a\, P(X > a). \end{aligned}$$ ■

Theorem 5 Chebyshev's Inequality

Let X be a random variable with mean μ and standard deviation σ. Then for any $k \in (0, \infty)$,

$$P\big(|X - \mu| > k\sigma\big) \;\leq\; \frac{1}{k^2}.$$

PROOF Let $Y = (X - \mu)^2$. By Markov's Inequality,

$$P\big(|X - \mu| > k\sigma\big) \;=\; P\big(Y > k^2\sigma^2\big) \;\leq\; \frac{E|Y|}{k^2\sigma^2} \;=\; \frac{\sigma^2}{\sigma^2 k^2} \;=\; \frac{1}{k^2}.$$ ■

Chebyshev's inequality uses the variance of a random variable to bound the probability that it is far away from the mean. We can use this inequality to prove the weak law in the case where the variables have finite variance:

Theorem 6 Weak Law—Finite Variance Version

Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite mean μ and finite variance σ^2, and let

$$\overline{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n}.$$

Then $\overline{X}_n \to \mu$ in probability as $n \to \infty$.

PROOF Observe that $E\overline{X}_n = \mu$ and

$$\mathrm{Var}(\overline{X}_n) \;=\; \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2} \;=\; \frac{\sigma^2}{n},$$

so $\overline{X}_n$ has standard deviation $\sigma/\sqrt{n}$. Therefore, by Chebyshev's Inequality

$$P\big(|\overline{X}_n - \mu| > \epsilon\big) \;=\; P\left(|\overline{X}_n - \mu| > \left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\frac{\sigma}{\sqrt{n}}\right) \;\leq\; \frac{\sigma^2}{\epsilon^2 n}.$$

This approaches 0 as $n \to \infty$, and the theorem follows. ■

Truncation and the General Case

So far, we have succeeded in proving the weak law for random variables that have finite variance. These are sometimes referred to as L^2 variables, since they are precisely the random variables that lie in $L^2(\Omega)$. However, the law holds true for any random variables with finite mean (i.e. for L^1 random variables). To prove this more general case, we must find a way to extend our result to variables with infinite variance.

Given a general variable $X \in L^1(\Omega)$, our plan is to "truncate" X to produce a variable with finite variance:

Definition: Truncation
Let $X\colon \Omega \to \mathbb{R}$ be a random variable, and let $N > 0$. The **truncation** of X at N is the variable $Y\colon \Omega \to [-N, N]$ defined by

$$Y \;=\; \begin{cases} X & \text{if } |X| \leq N \\ 0 & \text{if } |X| > N. \end{cases}$$

Note that any truncation of X is bounded, and therefore has finite variance.

Lemma 7 Truncation Lemma

Let $X\colon \Omega \to \mathbb{R}$ be a random variable with finite expected value, and let $\epsilon > 0$. Then there exists a truncation Y of X so that $E|X - Y| < \epsilon$.

PROOF For each $N \in \mathbb{N}$, let Y_N be the truncation of X at N. It suffices to show that $E|X - Y_N| \to 0$ as $N \to \infty$.

We shall use the dominated convergence theorem, applied to integrals over Ω. Clearly $|X - Y_N| \to 0$ pointwise as $N \to \infty$. Further, all of the functions $|X - Y_N|$

are bounded by $|X|$, and

$$\int_\Omega |X|\, dP \;=\; E|X| \;<\; \infty.$$

Therefore, it follows from the dominated convergence theorem that

$$\int_\Omega |X - Y_N|\, dP \;\to\; 0 \qquad \text{as } N \to \infty.$$

That is, $E|X - Y_N| \to 0$ as $N \to \infty$. ■

Theorem 8 Weak Law of Large Numbers

Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite mean μ, and let

$$\overline{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n}.$$

Then $\overline{X}_n \to \mu$ in probability as $n \to \infty$.

PROOF Let $\epsilon_1 > 0$ and $\epsilon_2 > 0$. We will prove that

$$P\big(\big|\overline{X}_n - \mu\big| > \epsilon_1\big) \;<\; \epsilon_2$$

for sufficiently large values of n.

For convenience of notation, let X be a random variable with same distribution as the X_n's, and let Y be a truncation of X for which

$$E|X - Y| \;<\; \max\Big(\frac{\epsilon_1\epsilon_2}{6},\, \frac{\epsilon_1}{3}\Big).$$

For each n, let Y_n be the corresponding truncation of X_n, and let

$$\overline{Y}_n \;=\; \frac{Y_1 + \cdots + Y_n}{n}.$$

By the triangle inequality, we have:

$$\big|\overline{X}_n - EX\big| \;\le\; \big|\overline{X}_n - \overline{Y}_n\big| \;+\; \big|\overline{Y}_n - EY\big| \;+\; |EY - EX|.$$

We establish a bound for each of these three terms.

1. For the first term, observe that

$$E\big|\overline{X}_n - \overline{Y}_n\big| \;\leq\; \frac{E|X_1 - Y_1| + \cdots + E|X_n - Y_n|}{n} \;=\; E|X - Y| \;<\; \frac{\epsilon_1\epsilon_2}{6}.$$

By Markov's inequality, it follows that

$$P\left(\big|\overline{X}_n - \overline{Y}_n\big| > \frac{\epsilon_1}{3}\right) \;\leq\; \frac{E\big|\overline{X}_n - \overline{Y}_n\big|}{\epsilon_1/3} \;<\; \frac{\epsilon_1\epsilon_2/6}{\epsilon_1/3} \;=\; \frac{\epsilon_2}{2}.$$

2. For the second term, observe that the variables $\{Y_n\}$ are independent, identically distributed, and have finite variance. It follows that $\overline{Y}_n \to EY$ in probability as $n \to \infty$. In particular,

$$P\left(\big|\overline{Y}_n - EY\big| > \frac{\epsilon_1}{3}\right) \;<\; \frac{\epsilon_2}{2}$$

for sufficiently large n.

3. For the third term, observe that

$$|EX - EY| \;=\; |E(X - Y)| \;\leq\; E|X - Y| \;<\; \frac{\epsilon_1}{3}.$$

In particular,

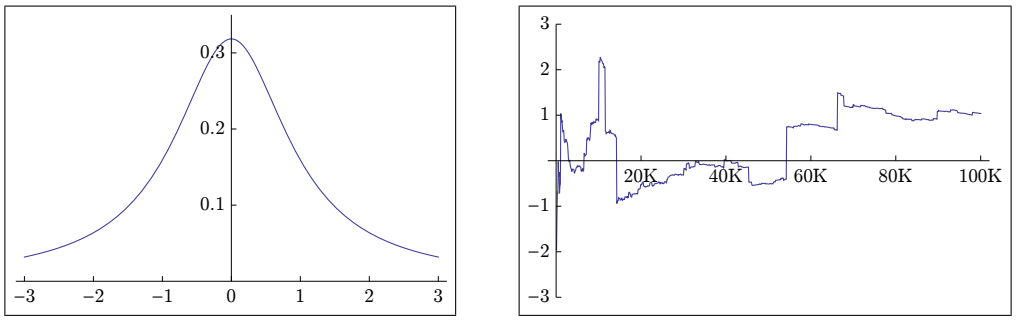
$$P\left(|EX - EY| > \frac{\epsilon_1}{3}\right) \;=\; 0.$$

Combining our results for each of the three terms, we have

$$\begin{aligned}
&P\big(\big|\overline{X}_n - EX\big| > \epsilon_1\big) \\
&\quad\leq\; P\left(\big|\overline{X}_n - \overline{Y}_n\big| > \frac{\epsilon_1}{3} \quad \text{or} \quad \big|\overline{Y}_n - EY\big| > \frac{\epsilon_1}{3} \quad \text{or} \quad |EY - EX| > \frac{\epsilon_1}{3}\right) \\
&\quad\leq\; P\left(\big|\overline{X}_n - \overline{Y}_n\big| > \frac{\epsilon_1}{3}\right) + P\left(\big|\overline{Y}_n - EY\big| > \frac{\epsilon_1}{3}\right) + P\left(|EY - EX| > \frac{\epsilon_1}{3}\right) \\
&\quad<\; \frac{\epsilon_2}{2} + \frac{\epsilon_2}{2} + 0 \;=\; \epsilon_2
\end{aligned}$$

for sufficiently large values of n. ■

Finally, we end this section with a "counterexample" to the weak law of large numbers in the case where the variables X_n do not have an expected value.



(a) (b)

Figure 2: (a) The standard Cauchy distribution. (b) Sample means $\overline{X}_n$ for 100,000 trials using the Cauchy distribution.

EXAMPLE 2 Cauchy Distribution

Let $\{X_n\}$ be an independent sequence of variables with the **standard Cauchy distribution**

$$f_X(x) \;=\; \frac{1}{\pi(1+x^2)}.$$

A plot of this probability density function is shown in Figure 2a. Since the integral

$$\int_{\mathbb{R}} x\, dP_X(x) \;=\; \int_{\mathbb{R}} \frac{x}{\pi(1+x^2)}\, dm(x)$$

does not exist, the expected value for this distribution is undefined.

As you might imagine, the sample means $\overline{X}_n$ for this distribution do not tend to converge. Indeed, all of the sample means $\overline{X}_n$ have precisely the same distribution, which is again the standard Cauchy distribution! Figure 2b shows experimental values of $\overline{X}_n$ for 100,000 trials using this distribution. ■

The Strong Law of Large Numbers

The strong law of large numbers is a version of the law of large numbers that is strictly more powerful than the weak law. It is based on the following notion of convergence:

Definition: Almost Sure Convergence

Let $\{X_n\}$ be a sequence of random variables, and let X be a random variable. We say that $X_n \to X$ **almost surely** if

$$P(X_n \to X) \;=\; 1.$$

That is, $X_n \to X$ almost surely if the functions X_n converge to X pointwise almost everywhere on the sample space. In general, probabilists say that an event occurs **almost surely** if the probability of the event is 1. This is the same as the measure-theoretic notion of "almost everywhere".

The goal of this section is to prove the following theorem:

Strong Law of Large Numbers

Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite expected value μ. For each n, let

$$\overline{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n}.$$

Then $\overline{X}_n \to \mu$ almost surely as $n \to \infty$.

Before we begin to prove this theorem, we should discuss the difference between almost sure convergence and convergence in probability. The following lemma is crucial to understanding this difference:

Theorem 9 Borel-Cantelli Lemma

Let $\{E_n\}$ be a sequence of events on a probability space, and suppose that

$$\sum_{n=1}^{\infty} P(E_n) \;<\; \infty.$$

Then, almost surely, only finitely many of the events E_n occur.

PROOF Let N be a random variable whose value is the number of events E_n that occur. Then

$$N \;=\; \sum_{n=1}^{\infty} \chi_{E_n},$$

where χ_{E_n} is the characteristic function of E_n. By the monotone convergence theorem, it follows that

$$EN \;=\; \sum_{n=1}^{\infty} E[\chi_{E_n}] \;=\; \sum_{n=1}^{\infty} P(E_n) \;<\; \infty.$$

Since EN has finite expected value, it must be the case that $P(N < \infty) = 1$. ■

This lemma gives us a nice test for almost sure convergence:

Theorem 10 Almost Sure Convergence Test

Let $\{X_n\}$ be a sequence of random variables, and let X be a random variable. Suppose that for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P\big(|X_n - X| > \epsilon\big) \;<\; \infty.$$

Then $X_n \to X$ almost surely.

PROOF For each k, let

$$E_k \quad = \quad \text{“}|X_n - X| \geq \frac{1}{k} \text{ for infinitely many } n.\text{”}$$

By the Borel-Cantelli Lemma, $P(E_k) = 0$ for each k. Then $P(\bigcup_{k=1}^{\infty} E_k) = 0$, so $X_n \to X$ almost surely. ■

The following example shows that variables may converge in probability without converging almost surely:

EXAMPLE 3 Convergence in Probability, but not Almost Surely
Let $X_n \colon \Omega \to [1, \infty)$ be a sequence of independent, identically distributed random variables with

$$f_X(x) \;=\; \frac{1}{x^2},$$

and let $Y_n = X_n/n$. Then $Y_n \to 0$ in probability, with

$$P(Y_n > \epsilon) \;=\; P(X_n > n\epsilon) \;=\; \frac{1}{\epsilon n}$$

Since $\sum P(Y_n > \epsilon) = \infty$, these random variables do not satisfy the hypothesis of Theorem 10. Indeed, these random variables do not converge to zero almost surely. In particular,

$$P(Y_n \leq \epsilon \text{ for all } n \geq N) \;=\; \prod_{n=N}^{\infty} \left(1 - \frac{1}{\epsilon n}\right) \;=\; 0$$

for all $\epsilon > 0$ and all $N \in \mathbb{N}$. ■

Proof of the Strong Law

We now turn to the proof of the strong law of large numbers. Before we begin, recall that our proof of the weak law used Chebyshev's inequality to give us the bound

$$P\big(|\overline{X}_n - \mu| > \epsilon\big) \;\leq\; \frac{\sigma^2}{\epsilon^2 n}.$$

Since $\sum 1/n$ diverges, this bound is not useful for proving almost sure convergence. To prove the strong law, we will need a better bound than Chebyshev's inequality can provide.

To obtain a stronger bound, we need a more sensitive measure of variability than variance. The following definition generalizes the notion of variance:

Definition: Moments
Let $X\colon \Omega \to \mathbb{R}$ be a random variable with finite mean μ. If $k \in \{2, 3, 4 \ldots\}$, the ***k*th moment** of X is the quantity

$$E\big[(X - \mu)^k\big].$$

For example, the 2nd moment of X is the same as the variance of X. The moments of X break into two main types:

1. If k is even, then the kth moment is a measure of dispersion, similar to the variance or standard deviation. However, larger values of k give greater weight to values of X that are farther from the mean.

2. If k is odd, then the kth moment counts values less than the mean as negative, and evaluates to zero for distributions that are symmetric about the mean. In this case, the kth moment can be thought of as a measure of the **skewness** (or asymmetry) of a distribution.

Since we are interested in dispersion, we will skip over the third moment and use the fourth moment of a random variable. The following lemma is an analogue of Chebyshev's inequality for the fourth moment:

Lemma 11 Fourth Moment Estimate

Let X be a random variable with finite mean μ and finite fourth moment τ^4. Then for any $k \in (0, \infty)$,

$$P\big(|X - \mu| > k\tau\big) \;\leq\; \frac{1}{k^4}.$$

PROOF Let $Y = (X - \mu)^4$. By Markov's Inequality,

$$P\big(|X - \mu| > k\tau\big) \;=\; P\big(Y > k^4\tau^4\big) \;\leq\; \frac{E|Y|}{k^4\tau^4} \;=\; \frac{\tau^4}{\tau^4 k^4} \;=\; \frac{1}{k^4}.$$ ■

Theorem 12 Strong Law—Finite Fourth Moment Version

Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite mean μ, finite variance σ^2, and finite fourth moment τ^4, and let

$$\overline{X}_n \;=\; \frac{X_1 + \cdots + X_n}{n}.$$

Then $\overline{X}_n \to \mu$ almost surely as $n \to \infty$.

PROOF The first step is to calculate the fourth moment of $\overline{X}_n$. This is tedious but straightforward, and leads to the following result:

$$E\big[(\overline{X}_n - \mu)^4\big] \;=\; \frac{n\tau^4 + 6\binom{n}{2}\sigma^4}{n^4} \;=\; \frac{\tau^4 + \frac{3}{2}(n-1)\sigma^4}{n^3}.$$

In particular,

$$E\big[(\overline{X}_n - \mu)^4\big] \;\leq\; \frac{C}{n^2}$$

where $C = \tau^4 + \frac{3}{2}\sigma^4$. Therefore, by the lemma,

$$P\big(\big|\overline{X}_n - \mu\big| > \epsilon\big) \;=\; P\left(\big|\overline{X}_n - \mu\big| > \sqrt[4]{\frac{\epsilon^4 n^2}{C}} \cdot \sqrt[4]{\frac{C}{n^2}}\right) \;\leq\; \frac{C}{\epsilon^4 n^2}.$$

Since

$$\sum_{n=1}^{\infty} \frac{C^4}{\epsilon^4 n^2} \;<\; \infty,$$

it follows from Proposition 10 that $\overline{X}_n \to \mu$ almost surely as $n \to \infty$. ■

This proves the strong law for random variables with finite fourth moment, i.e. for variables in $L^4(\Omega)$. However, like the weak law, the strong law is true for any random variable with finite expected value. Indeed, it is possible to extend the strong law to arbitrary L^1 variables using a truncation argument, similar to our approach to the weak law. Unfortunately, the details are a bit involved, so we will not pursue the strong law any further.

The Central Limit Theorem

The third major foundational theorem of probability is the central limit theorem. Roughly speaking, this theorem states that the distribution of the sample mean $\overline{X}_n$ tends to converge to a normal distribution as $n \to \infty$.

To state this idea more precisely, we must discuss the idea of convergence of probability measures:

Definition: Weak Convergence
Let $\{P_n\}$ be a sequence of probability measures on $\mathbb{R}$, and let P be a probability measure on $\mathbb{R}$. We say that P_n **converges weakly** to P if

$$\int_{\mathbb{R}} g \, dP_n \;\to\; \int_{\mathbb{R}} g \, dP$$

for every bounded, continuous function $g\colon \mathbb{R} \to \mathbb{R}$.

The following examples should clarify this notion of convergence:

EXAMPLE 4 Discrete Approximations to Lebesgue Measure
For each n, let P_n be the probability measure on $\mathbb{R}$ satisfying

$$P_n\Big(\Big\{\frac{k}{n}\Big\}\Big) \;=\; \frac{1}{n} \qquad \text{for } k \in \{1, 2, \ldots, n\},$$

and let P be Lebesgue measure restricted to the interval $[0, 1]$. Then the measures P_n converge weakly to P. In particular, if $g\colon \mathbb{R} \to \mathbb{R}$ is any bounded, continuous function, then

$$\sum_{k=1}^{n} \frac{1}{n}\, g\Big(\frac{k}{n}\Big) \;\to\; \int_{[0,1]} g \, dm \qquad \text{as } n \to \infty. \qquad \blacksquare$$

EXAMPLE 5 Convergence of Continuous Distributions
In general, a **probability density function** on $\mathbb{R}$ is an L^1 function $f\colon \mathbb{R} \to [0, \infty]$ satisfying $\int_{\mathbb{R}} f = 1$. Every density function f has an associated probability measure P_f defined by

$$P_f(S) \;=\; \int_S f \, dm,$$

where m is Lebesgue measure.

Now let f_n be a sequence of probability density functions, let f be a probability density function, and suppose that $f_n \to f$ in the L^1 norm. In this case, the measures P_{f_n} converge to P_f in probability. In particular, if $g\colon \mathbb{R} \to \mathbb{R}$ is any bounded,

continuous function, then

$$\begin{aligned}\left|\int_R g\,dP_{f_n} - \int_{\mathbb{R}} g\,dP_f\right| &= \left|\int_R f g_n\,dm - \int_R fg\,dm\right| \\ &\leq \int_{\mathbb{R}} |f_n g - fg|\,dm \\ &\leq \|f_n - f\|_1\,\|g\|_\infty\end{aligned}$$

which goes to 0 as $n \to \infty$. ■

EXAMPLE 6 Converging to the Delta Measure

For each n, let P_{f_n} be the continuous probability measure on $\mathbb{R}$ with density function

$$f_n(x) = \begin{cases} n/2 & \text{if } |x| \leq 1/n \\ 0 & \text{if } |x| > 1/n, \end{cases}$$

and let δ be the measure

$$\delta(S) = \begin{cases} 1 & \text{if } 0 \in S \\ 0 & \text{if } 0 \notin S. \end{cases}$$

Then the measures P_{f_n} converge weakly to δ. In particular, if $g\colon \mathbb{R} \to \mathbb{R}$ is any bounded, continuous function, then

$$\frac{n}{2}\int_{[-\frac{1}{n},\frac{1}{n}]} g\,dm \to g(0) \qquad \text{as } n \to \infty.$$ ■

For the following theorem, recall that the **standard normal distribution** is the probability measure on $\mathbb{R}$ defined by the density function

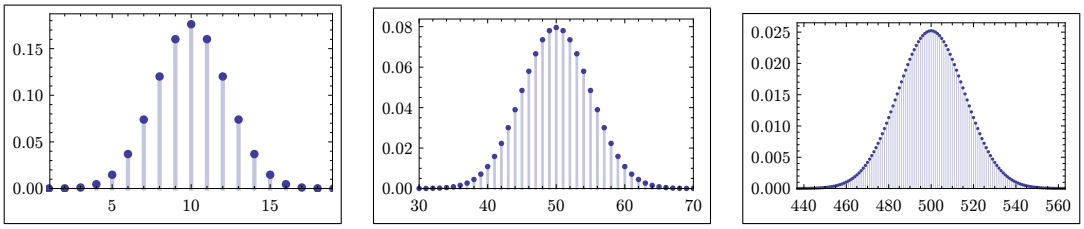
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

Theorem 13 Central Limit Theorem

Let $X_n\colon \Omega \to \mathbb{R}$ be a sequence of independent, identically distributed random variables with finite mean μ and finite variance σ^2. For each n, let

$$Y_n = \frac{X_1 + \cdots X_n - n\mu}{\sigma\sqrt{n}},$$

so Y_n has mean 0 and variance 1. Then P_{Y_n} converges weakly to the standard normal distribution as $n \to \infty$.



(a) (b) (c)

Figure 3: Symmetric binomial distributions corresponding to (a) $n = 20$ (b) $n = 100$ and (c) $n = 1000$.

EXAMPLE 7 Binomial Distributions

Let $C_n : \Omega \to \{0, 1\}$ be a sequence of independent coin flips, and let

$$X_n \;=\; C_1 + \cdots + C_n.$$

Then X_n is a discrete random variable, with probability distribution given by

$$P_{X_n}(\{k\}) \;=\; \frac{1}{2^n}\binom{n}{k} \qquad \text{for } k \in \{0, 1, \ldots, n\}.$$

This probability distribution is known as the **symmetric binomial distribution**, named after the binomial coefficients appearing in the formula. Plots of the distributions of X_{20}, X_{100}, and X_{1000} are shown in Figure 3.

From the figure, it appears that the binomial distributions converge to a normal distribution as $n \to \infty$. Indeed, according to the central limit theorem, the probability distributions for the variables

$$\frac{X_n - n/2}{\sqrt{n}/2}$$

converge weakly to the standard normal distribution as $n \to \infty$. ■

Though we are not in a position to prove the central limit theorem, we can try to convey some of the intuition behind it. In a fundamental way, the central limit theorem involves the distribution of a sum of variables. The following theorem describes the distribution of a sum in the case where one of the variables is continuous:

Proposition 14 Distribution of a Sum

Let $X, Y : \Omega \to \mathbb{R}$ be independent random variables, and let $Z = X + Y$. If X is continuous, then Z is continuous, with

$$f_Z(z) \;=\; \int_{\mathbb{R}} f_X(z - y)\, dP_Y(y).$$

PROOF Let $f\colon \mathbb{R} \to \mathbb{R}$ be the function defined the integral above, and let $S \subset \mathbb{R}$ be measurable. By Fubini's Theorem,

$$\begin{aligned}\int_S f\,dm &= \int_S \int_{\mathbb{R}} f_X(z-y)\,dP_Y(y)\,dm(z) \\ &= \int_{\mathbb{R}} \int_S f_X(z-y)\,dm(z)\,dP_Y(y) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} f_X(z-y)\,\chi_S(z)\,dm(z)\,dP_Y(y).\end{aligned}$$

Substituting $x = z - y$ in the last integral gives

$$\int_S f\,dm = \int_{\mathbb{R}} \int_{\mathbb{R}} f_X(x)\,\chi_S(x+y)\,dm(x)\,dP_Y(y) = \int_{\mathbb{R}} \int_{\mathbb{R}} \chi_S(x+y)\,dP_X(x)\,dP_Y(y).$$

Since X and Y are independent, the product measure $dP_X \times dP_Y$ is the same as the joint distribution $dP_{(X,Y)}$. Therefore, by Fubini's theorem,

$$\int_S f\,dm = \int_{\mathbb{R}^2} \chi_S(x+y)\,dP_{(X,Y)}(x,y) = P(X+Y \in S) = P(Z \in S).$$

Since $S \subset \mathbb{R}$ was an arbitrary measurable set, this proves that Z is continuous and f is a probability density function for Z. ■

In the case where both X and Y are continuous and $Z = X + Y$, the proposition above gives the formula

$$f_Z(z) = \int_{\mathbb{R}} f_X(z-y)\,f_Y(y)\,dm(y) = (f_X * f_Y)(z).$$

That is, f_Z is the convolution f_X and f_Y.

In particular, if $\{X_n\}$ is a sequence of independent, identically distributed, continuous random variables, then the probability density function for the sum $X_1 + \cdots + X_n$ is nth the iterated convolution

$$f_X * f_X * \cdots * f_X$$

where f_X is the probability density function for each X_n. According to the central limit theorem, this iterated convolution tends to converge to a normal distribution as $n \to \infty$.

The following proposition explains why this might be the case:

Proposition 15 Stability of Normal Distributions

The sum of two or more independent, normally distributed random variables is normally distributed.

PROOF Let X and Y be normally distributed random variables, and let $Z = X+Y$. Then

$$f_X(x) \;=\; Ae^{-p(x)} \qquad \text{and} \qquad f_Y(y) \;=\; Be^{-q(y)},$$

where A and B are positive constants, and $p(t)$ and $q(t)$ are quadratic polynomials with positive leading coefficients. Then

$$f_Z(z) \;=\; (f_X * f_Y)(z) \;=\; \int_{\mathbb{R}} f_X(z-y) f_Y(y)\, dm(y) \;=\; \int_{\mathbb{R}} Ae^{-p(z-y)} Be^{-q(y)}\, dm(y).$$

Now, if we complete the square, we can find quadratic polynomials $P(t)$ and $Q(t)$ with positive leading coefficients so that

$$p(z-y) \,+\, q(y) \;=\; P(z) \,+\, Q(z-y).$$

for all $y, z \in \mathbb{R}$. Then

$$\begin{aligned} f_Z(z) \;&=\; \int_{\mathbb{R}} Ae^{-P(z)} Be^{-Q(z-y)}\, dm(y) \;=\; \left(\int_{\mathbb{R}} ABe^{-Q(z-y)}\, dm(y) \right) e^{-P(z)} \\ &=\; \left(\int_{\mathbb{R}} ABe^{-Q(x)}\, dm(x) \right) e^{-P(z)} \;=\; Ce^{-P(z)}. \end{aligned}$$

■

In general, a probability distribution is said to be **stable** if the sum of two independent variables with that distribution again has the same distribution (up to translation and rescaling). For a continuous distribution, this says that

$$(f * f)(x) \;=\; \frac{1}{a}\, f(ax+b)$$

for some constants a and b, where f is the probability density function. According to the above proposition, normal distributions are stable in this sense.

In fact, it can be shown that the normal distribution is the *only* stable distribution with finite mean and variance. That is, the normal distribution is the unique fixed point for the operation of self-convolution. Thus the central limit theorem can be thought of as saying that probability distributions tend to converge to this fixed point under repeated applications of this operation.

Exercises

1. If E and F are independent events, prove that E and F^c are independent.
2. Let E and F be events, and suppose that $P(E) = p$ and $P(F) = q$. What is the maximum possible probability of $P(E \cap F)$? What is the minimum possible probability of $P(E \cap F)$?

3. Let E be an event, let $F_1 \subset F_2 \subset F_3 \subset \cdots$ be an increasing sequence of events, and suppose that E and F_n are independent for each n. Prove that E and $\bigcup_{n=1}^{\infty} F_n$ are independent.

4. Let $X \colon \Omega \to \mathbb{R}$ be a random variable, let $a, b \in \mathbb{R}$, and let $Y = aX + b$.
   a) If X has mean μ and standard deviation σ, what are the mean and standard deviation of Y?
   b) If X is continuous with probability density function f_X, what is the probability density function for Y?

5. Suppose we flip a coin three times. Find four events E_1, E_2, E_3, E_4 for this experiment such that any three are independent, but all four together are not independent.

6. An experiment has 100 possible outcomes, all equally likely. Suppose that $\{E_1, \ldots, E_n\}$ is a collection of independent events for this experiment, each with probability strictly between 0 and 1. What is the maximum possible value for n?

7. Let $\{X_1, \ldots, X_{100}\}$ be a sequence of elements of $[0, 1]$, chosen uniformly at random, and let $Y = X_1 + \cdots + X_{100}$. Prove that
$$P(40 \leq Y \leq 60) \;\geq\; \frac{11}{12}.$$

8. Let $\{X_n\}$ be a sequence of independent, identically distributed continuous random variables with probability density function
$$f_X(x) \;=\; \frac{1}{(1 + |x|)^3},$$
and let $Y_n = X_1 + \cdots + X_n$. Prove that
$$P(-100 \leq Y_{10} \leq 100) \;\geq\; \frac{9}{10}.$$

9. Let $X \colon \Omega \to [0, \infty)$ be a continuous random variable with finite expected value, and suppose that the probability density function $f_X \colon [0, \infty) \to [0, \infty]$ is decreasing. Prove that
$$f_X(x) \;\leq\; \frac{2\,EX}{x^2}$$
for all $x > 0$.

10. Let X and Y be independent random variables with $EX = EY = 0$. Prove that
$$E\big[(X+Y)^4\big] \;=\; E\big[X^4\big] + E\big[Y^4\big] + 6\,\mathrm{Var}(X)\,\mathrm{Var}(Y).$$

11. Let N be the number of heads in 10,000 coin flips.
    a) Find the standard deviation of N.
    b) Use the central limit theorem to estimate $P(4950 \le N \le 5050)$.

12. In general, a **Bernoulli random variable** is any variable $B\colon \Omega \to \{0,1\}$ satisfying
$$P_{\scriptscriptstyle B}(\{0\}) \;=\; 1-p \qquad \text{and} \qquad P_{\scriptscriptstyle B}(\{1\}) = p$$
for some $p \in (0,1)$.

    Let $\{B_n\}$ be a sequence of independent, identically distributed Bernoulli random variables, and let $X_n = B_1 + \cdots + B_n$. Then X_n is said to have a **binomial distribution**

    a) Compute the mean and standard deviation of X_n. Your answers should be formulas involving n and p.
    b) If $k \in \{0, 1, \ldots, n\}$, compute $P(X_n = k)$.