# 2 Independence

## 2.1 Borel's Strong Law of Large Numbers

**Standing Convention:** $q = 1 - p$

**Definition 2.1.** A sequence $\{X_n\}_{n\geq 1}$ of random variables defined on a common probability space $(\Omega, \mathscr{F}, P)$ taking values in $\{0,1\}$ is said to be independent, identically distributed Bernoulli$-p$ if for every finite sequence $e_i$ of 0s and 1s,

$$P\{X_i = e_i \text{ for each } i \leq m\} = p^{\sum_{i=1}^m e_i} q^{m - \sum_{i=1}^m e_i}.$$

**Theorem 2.2.** *(Borel) Assume that $X_1, X_2, \dots$ are independent, identically distributed Bernoulli$-p$ random variables defined on a probability space $(\Omega, \mathscr{F}, P)$, and for each $n \geq 1$ let $S_n = \sum_{i=1}^n X_i$. Then*

$$P\{\lim_{n\to\infty} S_n / n = p\} = 1. \tag{2.1}$$

*Proof.* Convergence of a sequence to the limit $p$ means that for every $\varepsilon > 0$ only finitely many terms of the sequence lie outside the interval $(p-\varepsilon, p+\varepsilon)$, and since the rational numbers are dense in $\mathbb{R}$, only *rational* values of $\varepsilon$ need be considered. Thus, we must show that with probability one, for every rational $\varepsilon > 0$ only finitely many terms of the sequence $S_n/n$ are not between $p-\varepsilon$ and $p+\varepsilon$. Since the rational numbers are countable, it is enough to prove that for each fixed rational $\varepsilon > 0$,

$$P\{|S_n - p| \geq \varepsilon \;\; i.o\} = 0.$$

Consequently, by the Borel-Cantelli lemma, it suffices to prove that for each $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\{S_n \geq np + n\varepsilon\} < \infty \quad \text{and} \tag{2.2}$$

$$\sum_{n=1}^{\infty} P\{S_n \leq np - n\varepsilon\} < \infty. \tag{2.3}$$

The problem is now reduced to finding upper bounds on the tail probabilities for the distribution of $S_n$, a common problem in probability theory. The first tool that might come to mind, the *Chebyshev inequality*, will not work here, because it gives upper bounds on the order of $c/n$ (for some $c > 0$ depending on $\varepsilon$ and the variance $pq$), and unfortunately the sequence $1/n$ is not summable. Nevertheless, usable bounds can be gotten by a strategy not unlike that underlying the Chebyshev inequality: in particular, if $f : \mathbb{R} \to (0, \infty)$ is any *non-decreasing*, positive function, then for any $\alpha > 0$

$$P\{S_n \geq n\alpha\} \leq \sum_{k=[n\alpha]}^{n} \frac{f(k)}{f([n\alpha])} P\{S_n = k\},$$

where $[\cdot]$ is the greatest integer function. The trick will be to find the right function $f$. Let's try $f(x) = e^{\theta x}$ for some parameter $\theta > 0$; this leads us to

$$\begin{aligned}
P\{S_n \geq np + n\varepsilon\} &\leq \exp\{-[np\theta + n\varepsilon\theta]\} \sum_{k=[np+n\varepsilon]}^{n} e^{\theta k} P\{S_n = k\} \\
&\leq \exp\{-[np\theta + n\varepsilon\theta]\} \sum_{k=0}^{n} e^{\theta k} P\{S_n = k\} \\
&= \exp\{-[np\theta + n\varepsilon\theta]\} \sum_{k=0}^{n} \binom{n}{k} (pe^{\theta})^k q^{n-k} \\
&= \exp\{-[np\theta + n\varepsilon\theta]\} \left(pe^{\theta} + q\right)^n \qquad (2.4)
\end{aligned}$$

This chain of inequalities holds for any value of $\theta > 0$. Consequently, if we can find a particular value of $\theta > 0$ for which the bound is summable in $n$, then inequality (2.2) will be proved. Now the bound is (essentially) of the form (something)$^n$ (the greatest integer in the first exponential can be dropped at a cost of at most $e^1$); hence, we should look for a value of $\theta$ that will make (something) less than 1. But

$$(\text{something}) = e^{-p\theta + \varepsilon\theta}(pe^{\theta} + q);$$

this has the value 1 when $\theta = 0$, and the derivative with respect to $\theta$ at $\theta = 0$ is (exercise: do the calculus!)

$$-(p + \varepsilon) + p = -\varepsilon < 0.$$

Thus, for small values of $\theta > 0$ the value of (something) will be *less than* the value at $\theta = 0$, which is one. This proves that for small values of $\theta > 0$ the upper bounds in the inequalities (2.4) will be exponentially small in $n$, and so (2.2) holds. A similar argument (using $f(x) = e^{-\theta x}$) proves (2.3). □

## 2.2 Independence and Kolmogorov's 0–1 Law

**Standing Assumption:** $(\Omega, \mathscr{F}, P)$ is a probability space. An *event* is an element of the $\sigma-$algebra $\mathscr{F}$, and a *random variable* is a measurable transformation $X : \Omega \to \mathbb{R}$.

**Definition 2.3.** Events $A_1, A_2, \ldots, A_m$ are said to be *independent* if for every sub-collection $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}). \qquad (2.5)$$

An infinite collection of events $\{A_\theta\}_{\theta \in \Theta}$ is independent if every finite sub-collection is independent.

**Definition 2.4.** For each $i = 1,2,3,\ldots$ let $\mathscr{C}_i$ be a collection of events. The collections $\mathscr{C}_1, \mathscr{C}_2, \ldots$ are said to be *independent* if if for every choice of events $C_i \in \mathscr{C}_i$ the events $C_1, C_2, \ldots$ are independent.

In many instances the collections $\mathscr{C}_i$ of interest will be $\sigma-$algebras. In this case, checking that equation (2.5) holds for all possible choices $A_i = C_i \in \mathscr{C}_i$ might be difficult; fortunately, there are some useful shortcuts. The following criterion is especially useful: it states that one need only check that (2.5) holds for events $A_i$ in $\pi-$systems that generate the $\sigma-$algebras.

**Proposition 2.5.** *Suppose that for each $i = 1,2,\ldots$ the collection $\mathscr{A}_i$ is a $\pi-$system of events. If the collections $\mathscr{A}_i$ are independent, then the collections $\sigma(\mathscr{A}_i)$ are independent.*

*Proof.* $\pi - \lambda$ lemma. □

**Example 2.6.** Let $X_1, Y_1, X_2, Y_2, \ldots$ be independent, identically distributed Bernoulli$-p$ random variables defined on $(\Omega, \mathscr{F}, P)$. Let $A$ be the event that $\sum_{i=1}^n (1-2X_{i)=0}$ for infinitely many $n$, and let $B$ be the event that $\sum_{i=1}^n (1-2Y_{i)=0}$ for infinitely many $n$. Proposition 2.5 implies that $A$ and $B$ are independent, because $A$ is in the $\sigma-$algebra generated by the cylinder sets for the sequence $X_1, X_2, \ldots$, and $B$ is in $\sigma-$algebra generated by the cylinder sets for the sequence $Y_1, Y_2, \ldots$.

Example 2.6 helps to explain the usefulness of extending the notion of independence to $\sigma-$algebras. Often (usually?) a $\sigma-$algebra arises in connection with a collection of random variables in the same way that the two $\sigma-$algebras in Example 2.6 are associated with the sequences $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$. The next definition formalizes this.

**Definition 2.7.** If $\{X_\theta\}_{\theta\in\Theta}$ is a family of random variables then $\sigma(\{X_\theta\}_{\theta\in\Theta})$ is the smallest $\sigma-$algebra $\mathscr{G}$ such that all of the random variables $X_\theta$ are measurable with respect to $\mathscr{G}$. Equivalently,

$$\sigma(\{X_\theta\}_{\theta\in\Theta}) := \sigma(\{X_\theta^{-1}(B)\}_{\theta\in\Theta, B\in\mathscr{B}})$$

where $\mathscr{B}$ is the family of Borel subsets of $\mathbb{R}$.

**Definition 2.8.** Random variables $X_1, X_2, \ldots$ are said to be *independent* if the $\sigma-$algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent.

Borel's strong law of large numbers is the first of many theorems we will see in which some event involving a sequence of independent random variables turns out to have probability 1. You might wonder why it's always 1, and not (say) $\pi/6$ or $\sqrt{2}/2$ or $\cdots$. The next theorem, the *Kolmogorov 0–1 Law*, explains why.

**Definition 2.9.** Let $\mathscr{F}_1, \mathscr{F}_2, \ldots$ be $\sigma-$algebras of events (i.e., each $\mathscr{F}_i \subset \mathscr{F}$). The associated *tail field*[1] is defined to be the $\sigma-$algebra

$$\mathscr{T} := \bigcap_{m=1}^{\infty} \sigma\left(\bigcup_{n=m}^{\infty} \mathscr{F}_n\right)$$

[1] It should really be called the *tail $\sigma-$algebra*, but everyone is now used to calling it the *tail field*.

**Theorem 2.10.** *If the $\sigma-$algebras $\mathscr{F}_1, \mathscr{F}_2, \ldots$ are independent then every event of the associated tail field has probability either 0 or 1.*

*Proof.* The strategy is to show that every event $A \in \mathscr{T}$ is independent of itself, so that $P(A) = P(A)P(A)$. For this we will show that $A$ is independent of every event in $\mathscr{H} := \sigma(\cup_{n=1}^{\infty} \mathscr{F}_n)$; since $A \in \mathscr{H}$ it will then follow that $A$ is independent of $A$. By Proposition 2.5, it suffices to show that for each $m \geq 1$ the event $A$ is independent of $\sigma(\cup_{n=1}^{m} \mathscr{F}_n)$. But another application of Proposition 2.5 shows that for every $r > m$ the $\sigma-$algebras

$$\sigma(\bigcup_{n=1}^{m} \mathscr{F}_n) \quad \text{and} \quad \sigma(\bigcup_{n=m+1}^{r} \mathscr{F}_n)$$

are independent, and so a third application of Proposition 2.5 shows that the $\sigma-$algebras

$$\sigma(\bigcup_{n=1}^{m} \mathscr{F}_n) \quad \text{and} \quad \sigma(\bigcup_{n=m+1}^{\infty} \mathscr{F}_n)$$

are independent. Since $A$ is an element of this last $\sigma-$algebra, the result follows. □

## 2.3 SLLN for Bounded Random Variables

**Definition 2.11.** A real random variable $X$ defined on $(\Omega, \mathscr{F}, P)$ is called *simple* if there is a *finite* set $F \subset \mathbb{R}$ such that $P\{X \in F\} = 1$. If $X$ is simple and $a_1, a_2, \ldots, a_m$ are real numbers such that $P(\cup_{i \leq m}\{X = a_i\})$ then the *expectation* of $X$ is defined to be

$$EX = \sum_{i=1}^{m} a_i P\{X = a_i\}. \tag{2.6}$$

The definition (2.6) is the only one that makes sense if we want expectation to be *linear* and to agree with $P$ on indicators of events, i.e.,

(i) $E(aX) = a(EX)$ for all scalars $a \in \mathbb{R}$;
(ii) $E(X + Y) = (EX) + (EY)$ for any two random variables $X, Y$; and
(iii) $E\mathbf{1}_F = P(F)$ for any indicator $\mathbf{1}_F$ where $F \in \mathscr{F}$.

**Theorem 2.12.** *If $X_1, X_2, \ldots$ are independent, identically distributed* simple *random variables then with probability one,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = EX_1$$

*Proof.* Write $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \sum_{j=1}^{m} a_j \mathbf{1}\{X_i = a_j\}$ and then apply Borel's strong law of large numbers to the Bernoulli random variables $1\{X_i = a_j\}$. □

It is only a bit more work to prove the strong law of large numbers for *bounded* random variables. Since we haven't yet defined expectation for arbitrary random variables, we cannot yet express the limit as an expectation; nevertheless, the proof will yield an expression for the limit.

**Theorem 2.13.** *If $X_1, X_2, \dots$ are independent, identically distributed* bounded *random variables then with probability one,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.7}$$

*exists and is constant.*

*Proof.* For simplicity let's assume that $0 < X_i < 1$; the general case can then be deduced by a simple scaling and translation. For each integer $m \geq 1$, define functions $g_m$ and $h_m$ on $[0,1]$ as follows:

$$g_m(x) = 2^{-m}\lfloor 2^m x\rfloor \quad \text{and} \quad h_m(x) = g_m(x) + 2^{-m};$$

thus, $g_m(x)$ is the largest $k/2^m$ less or equal to $x$, and $h_m(x)$ is the smallest $k/2^m$ greater than $x$. Clearly, the random variables $g_m(X_i)$ and $h_m(X_i)$ are *simple*, as they take values in the finite set $\{k/2^m\}_{0\leq k\leq 2^m}$. Moreover, since the random variables $X_1, X_2, \dots$ are independent and identically distributed, then for any fixed $m \geq 1$ so are the random variables $g_m(X_1), g_m(X_2), \dots$, and so are the random variables $h_m(X_1), h_m(X_2), \dots$. Therefore, by Theorem 2.12, for each $m \geq 1$, with probability one,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g_m(X_i) = Eg_m(X_1) \quad \text{and}$$
$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h_m(X_i) = Eh_m(X_1).$$

Since $h_m - g_m = 2^{-m}$, the difference between the two limits $Eh_m(X_1)$ and $Eg_m(X_1)$ is exactly $2^{-m}$. Now each random variable $X_i$ is bounded above and below by $h_m(X_i)$ and $g_m(X_i)$; consequently, for each $n$

$$\frac{1}{n}\sum_{i=1}^{n} g_m(X_i) \leq \frac{1}{n}\sum_{i=1}^{n} X_i \leq \frac{1}{n}\sum_{i=1}^{n} h_m(X_i),$$

and so with probability one the liminf and limsup of the sequence $n^{-1}\sum_{i=1}^{n} X_i$ are between $Eg_m(X_1)$ and $Eh_m(X_1)$. Since this is true for *every* $m \geq 1$, and since $Eh_m(X_1) - Eg_m(X_1) = 2^{-m}$, it follows that with probability one the limit (2.7) exists and equals

$$\lim_{m\to\infty} Eg_m(X_1).$$

□

## 2.4 Glivenko-Cantelli Theorem

**Definition 2.14.** If $X_1, X_2, \dots, X_n$ are any real random variables defined n a common probability space $(\Omega, \mathscr{F}, P)$ then their *empirical c.d.f* is the cumulative distribution function

$$F_n(y) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le y\}.$$

**Theorem 2.15.** *(Glivenko-Cantelli) If $X_1, X_2, \dots$ are independent, identically distributed real random variables with common cumulative distribution function $F$ then with probability one,*

$$\lim_{n\to\infty} \sup_{y\in\mathbb{R}} |F_n(y) - F(y)| = 0. \tag{2.8}$$

*Proof.* First observe that for each *fixed* $y \in \mathbb{R}$ the random variables $\mathbf{1}\{X_i \le y\}$ are independent, identically distributed Bernoulli$-p$ with $p = F(y)$. Hence, Borel's strong law of large numbers implies that for each $y$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le y\} = F(y) \quad \text{with probability 1.}$$

This statement is weaker than (2.8), however, which states that the convergence is uniform over all real numbers $y$. For uniformity an additional argument is needed.

Consider first the case where $F$ is everywhere continuous. In this case, for every $\varepsilon > 0$ there are real numbers $y_1, y_2, \dots, y_m$ such that (with the notational convention $y_0 = -\infty$ and $y_{m+1} = +\infty$)

$$F(y_{i+1}) - F(y_i) < \varepsilon \quad \text{for every } 0 \le i \le m.$$

Since there are only finitely many points $y_i$ involved, Borel's SLLN implies that with probability one,

$$\lim_{n\to\infty} \max_{i\le m} |F_n(y_i) - F(y_i)| = 0$$

But both $F_n$ and $F$ are *monotone* functions of $y$, since they are cumulative distribution functions. Consequently, if $|F_n(y_i) - F(y_i)| < \varepsilon$ for both $i = j$ and $i = j+1$ then by the triangle inequality,

$$\sup_{y_i \le y \le y_{i+1}} |F_n(y) - F(y)| < 2\varepsilon.$$

The uniform convergence (2.8) now follows. □

**Exercise 2.16.** Finish the proof by showing how to handle distribution functions $F$ with points of discontinuity. HINTS: (a) For any $y \in \mathbb{R}$ the random variables $\mathbf{1}\{X_i < y\}$ are independent, identically distributed Bernoulli$-p$ with $p = F(y-) = \lim_{x\uparrow y} F(x)$. (b) For any $\varepsilon > 0$ there are at most finitely many real numbers $y$ at which $F$ has a jump discontinuity of size $\varepsilon$ or greater.

**Exercise 2.17.** Let $X_1, X_2, \ldots$ be independent, identically distributed, each with the uniform distribution on $[0,1]$. Explain why

$$\lim_{n\to\infty} \sup_{B\in\mathscr{B}_{[0,1]}} |\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_B(X_i) - \lambda(B)| = 1.$$