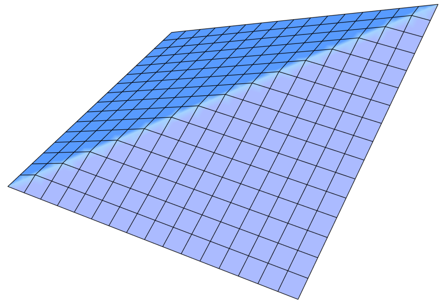

STA 2111H: Graduate Probability I

Professor Zhou Zhou • Fall 2018 • University of Toronto
M 2:00–5:00PM in SS 1069

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is `https://rob-zimmerman.github.io/coursenotes`.

Contents

Lecture 1 — Monday, September 10, 2018

1. Measure Spaces and Integration Foundations

Measurable Spaces and Measure Extension

Probability begins with a measurable space, which specifies the events to which probabilities may be assigned, and a measure, which assigns their sizes. We first develop these objects abstractly and then construct the Borel and Lebesgue measures needed throughout the course and daily life.

Definition 1.1 A pair $(\Omega, \mathscr{F})$ is a **measurable space** if Ω is a set and $\mathscr{F}$ is a σ-algebra on Ω. Thus, $\mathscr{F}$ is a nonempty collection of subsets of Ω such that the following hold:

1. The collection is closed under complements: if $A \in \mathscr{F}$, then $A^c \in \mathscr{F}$.
2. The collection is closed under countable unions: if $\{A_n\}_{n=1}^{\infty}$ is a countable collection

of sets in $\mathscr{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$.

Definition 1.2 A function $\mu\colon \mathscr{F} \to [0, \infty]$ is a **measure** on $(\Omega, \mathscr{F})$ if the following hold:

1. The empty set has measure zero: $\mu(\emptyset) = 0$.
2. The function is **countably additive**: if $A_1, A_2, \dots \in \mathscr{F}$ and $A_i \cap A_j = \emptyset$ whenever $i \neq j$, then
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

If $\mu(\Omega) = 1$, then μ is called a **probability measure**.

Lemma 1.3 $\Omega \in \mathscr{F}$ and $\emptyset \in \mathscr{F}$.

Proof. Since $\mathscr{F}$ is nonempty, there exists $A \in \mathscr{F}$. By (1), $A^c \in \mathscr{F}$, and then (2) gives $\Omega = A \cup A^c \in \mathscr{F}$. Applying (1) once more yields $\emptyset = \Omega^c \in \mathscr{F}$. $\square$

Remark 1.4 The collection $\{\Omega, \emptyset\}$ is the smallest possible σ-field on Ω and is called the **trivial σ-field**.

Corollary 1.5 Every measure μ satisfies the following properties:

1. If $A \subseteq B$, then $\mu(A) \leq \mu(B)$.
2. If $A \subseteq \bigcup_{i=1}^{\infty} A_i$, then
$$\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i).$$
3. If $A_i \uparrow A$, then $\mu(A_i) \to \mu(A)$.
4. If $A_i \downarrow A$ and $\mu(A_1) < \infty$, then $\mu(A_i) \to \mu(A)$.

Proof.

1. For (1), write $B = A \cup (B \setminus A)$ as a disjoint union. Countable additivity gives
$$\mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A).$$

2. For (2), set $B_i = A \cap A_i$. Then $A = \bigcup_{i=1}^{\infty} B_i$. Define $C_1 = B_1$ and
$$C_i = B_i \setminus \bigcup_{j=1}^{i-1} B_j, \quad i \geq 2.$$
The sets C_i are measurable and pairwise disjoint, $C_i \subseteq B_i \subseteq A_i$, and $A = \bigcup_{i=1}^{\infty} C_i$. Therefore,
$$\mu(A) = \sum_{i=1}^{\infty} \mu(C_i) \leq \sum_{i=1}^{\infty} \mu(B_i) \leq \sum_{i=1}^{\infty} \mu(A_i).$$

3. For (3), let $B_1 = A_1$ and $B_i = A_i \setminus A_{i-1}$ for $i \geq 2$. The sets B_i are pairwise disjoint, $A = \bigcup_{i=1}^{\infty} B_i$, and $A_n = \bigcup_{i=1}^{n} B_i$. Hence
$$\mu(A_n) = \sum_{i=1}^{n} \mu(B_i) \longrightarrow \sum_{i=1}^{\infty} \mu(B_i) = \mu(A).$$

4. For (4), assume that $A_i \downarrow A$ and $\mu(A_1) < \infty$. Define $B_i = A_1 \setminus A_i$. Then $B_i \uparrow A_1 \setminus A$, so (3) gives
$$\mu(B_i) \to \mu(A_1 \setminus A).$$
Since $\mu(A_i) = \mu(A_1) - \mu(B_i)$, it follows that
$$\mu(A_i) \to \mu(A_1) - \mu(A_1 \setminus A) = \mu(A).$$

□

Remark 1.6 The finiteness condition in (4) is essential. Let $\Omega = \mathbb{R}$, let $\mathscr{F} = \mathcal{B}_{\mathbb{R}}$, let μ be Lebesgue measure, and set $A_i = [i, \infty)$. Although $A_i \downarrow \emptyset$, we have $\mu(A_i) = \infty$ for every i, whereas $\mu(\emptyset) = 0$. Here Lebesgue measure is characterized on bounded intervals by
$$\mu([a,b]) = b - a, \quad -\infty < a \leq b < \infty.$$

Example 1.7 Let Ω be a countable set and let $\mathscr{F} = 2^\Omega$. If $f\colon \Omega \to [0,1]$ satisfies

$$\sum_{\omega\in\Omega} f(\omega) = 1,$$

then

$$\mu(A) = \sum_{\omega\in A} f(\omega), \quad A \subseteq \Omega,$$

defines a probability measure. If Ω is finite, the **uniform probability measure** is obtained by setting $f(\omega) = 1/|\Omega|$ for every $\omega \in \Omega$.

Lemma 1.8 If $\{\mathscr{F}_i\}_{i\in I}$ is any collection of σ-fields on Ω, then $\bigcap_{i\in I} \mathscr{F}_i$ is a σ-field. Consequently, for every collection $\mathcal{A}$ of subsets of Ω, there is a smallest σ-field containing $\mathcal{A}$: it is contained in every other σ-field that contains $\mathcal{A}$. This σ-field is called the **σ-field generated by $\mathcal{A}$** and is denoted by $\sigma(\mathcal{A})$.

Proof. Let $\mathcal{G} = \bigcap_{i\in I} \mathscr{F}_i$. Every $\mathscr{F}_i$ contains Ω, so $\Omega \in \mathcal{G}$. If $A \in \mathcal{G}$, then $A \in \mathscr{F}_i$ for every i, so $A^c \in \mathscr{F}_i$ for every i and therefore $A^c \in \mathcal{G}$. Similarly, if $A_n \in \mathcal{G}$ for every n, then $\bigcup_{n=1}^\infty A_n \in \mathscr{F}_i$ for every i, and hence $\bigcup_{n=1}^\infty A_n \in \mathcal{G}$. Thus $\mathcal{G}$ is a σ-field. $\square$

Definition 1.9 The **Borel σ-field** on $\mathbb{R}^d$ is the smallest σ-field containing every open subset of $\mathbb{R}^d$. We denote it by $\mathcal{R}^d$, so the corresponding measurable space is $(\mathbb{R}^d, \mathcal{R}^d)$.

Defining a measure directly on every Borel set is difficult. The extension procedure below instead begins with a simpler collection of sets.

Definition 1.10 A collection of subsets $\mathcal{S}$ is a **semialgebra** if it is closed under finite intersections and the complement of each $A \in \mathcal{S}$ can be expressed as a finite disjoint union of sets in $\mathcal{S}$.

Example 1.11 On $\Omega = \mathbb{R}$, the collection

$$\mathcal{S} = \{(a,b] : -\infty \le a \le b \le \infty\}$$

is a semialgebra. Lebesgue measure is first specified on finite intervals in this semialgebra by $\mu((a,b]) = b - a$ and is then extended to the Borel σ-field.

Suppose that a measure μ is defined on the semi-algebra in Example 1.11 and is finite on every bounded interval. Define

$$F(b) = \begin{cases} \mu((0,b]), & b \geq 0, \\ -\mu((b,0]), & b < 0. \end{cases}$$

Then $F(0) = 0$, $F(b) \geq 0$ for $b \geq 0$, and $F(b) \leq 0$ for $b < 0$. Monotonicity of μ shows that F is nondecreasing. Moreover, F is right continuous: if $y_n \downarrow y$, then $F(y_n) \downarrow F(y)$.

Proof. We prove the case $y \geq 0$; the case $y < 0$ is analogous. If $y_n \downarrow y$, then

$$(0, y_n] \downarrow (0, y].$$

Since $\mu((0, y_1]) < \infty$, continuity from above, (4), gives

$$F(y_n) = \mu((0, y_n]) \downarrow \mu((0, y]) = F(y).$$

$\square$

The next definitions provide the framework for extending this set function to the Borel σ-field.

Definition 1.12 A collection $\mathcal{A}$ of subsets of Ω is an **algebra** if the following hold:

1. The collection is closed under finite unions: $A, B \in \mathcal{A}$ implies $A \cup B \in \mathcal{A}$.
2. The collection is closed under complements: $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$.

Definition 1.13 The **σ-field generated by an algebra $\mathcal{A}$** is the smallest σ-field containing $\mathcal{A}$ and is denoted by $\sigma(\mathcal{A})$.

Theorem 1.14 (Extension theorem) Let $\mathcal{S}$ be a semialgebra, and let $\mu \colon \mathcal{S} \to [0, \infty]$ be a set function satisfying $\mu(\emptyset) = 0$. Suppose that

1. If $S \in \mathcal{S}$ is a finite disjoint union $S = \bigcup_{i=1}^n S_i$ of sets $S_i \in \mathcal{S}$, then

$$\mu(S) = \sum_{i=1}^n \mu(S_i).$$

2. If $S, S_1, S_2, \dots \in \mathcal{S}$ and $S \subseteq \bigcup_{i=1}^\infty S_i$, then

$$\mu(S) \leq \sum_{i=1}^\infty \mu(S_i).$$

Then μ has a unique extension $\bar{\mu}$ to the smallest algebra $\bar{\mathcal{S}}$ containing $\mathcal{S}$. If $\bar{\mu}$ is σ-finite, then it has a unique extension to $\sigma(\mathcal{S})$.

Definition 1.15 A measure α on $(\Omega, \mathscr{F})$ is ***σ*-finite** if there exists a sequence $A_i \uparrow \Omega$ such that $\alpha(A_i) < \infty$ for every i.

The first step in proving Theorem 1.14 is to identify the algebra generated by the semi-algebra.

Lemma 1.16 If $\mathcal{S}$ is a semialgebra, then

$$\bar{\mathcal{S}} = \{\text{finite disjoint unions of sets in } \mathcal{S}\}$$

is the smallest algebra containing $\mathcal{S}$.

Proof. We first show that $\bar{\mathcal{S}}$ is an algebra. If $A, B \in \bar{\mathcal{S}}$, write $A = \bigcup_{j=1}^n S_j$ and $B = \bigcup_{k=1}^m T_k$ as finite disjoint unions of sets in $\mathcal{S}$. Then

$$A \cap B = \bigcup_{j=1}^n \bigcup_{k=1}^m (S_j \cap T_k) \in \bar{\mathcal{S}},$$

because $\mathcal{S}$ is closed under finite intersections. For each j, express S_j^c as a finite disjoint union $\bigcup_{i=1}^{m_j} T_{ij}$ of sets in $\mathcal{S}$. Then

$$A^c = \bigcap_{j=1}^n S_j^c = \bigcap_{j=1}^n \bigcup_{i=1}^{m_j} T_{ij}.$$

Distributing the intersections over the finite unions expresses A^c as a finite union of finite intersections of sets in $\mathcal{S}$, hence as an element of $\bar{\mathcal{S}}$. Thus $\bar{\mathcal{S}}$ is closed under complements and finite intersections, and therefore it is an algebra.

If $\mathcal{A}$ is any algebra containing $\mathcal{S}$, then $\mathcal{A}$ contains every finite union of sets in $\mathcal{S}$. Consequently, $\bar{\mathcal{S}} \subseteq \mathcal{A}$, which proves minimality. $\square$

Lecture 2 — Monday, September 17, 2018

Definition 2.1 Let $\mathcal{A}$ be an algebra. A function $\mu\colon \mathcal{A} \to [0,\infty]$ is a **measure on $\mathcal{A}$** if

1. The empty set has measure zero: $\mu(\emptyset) = 0$.
2. The function is finitely additive: if $A_1, \ldots, A_n \in \mathcal{A}$ are pairwise disjoint, then

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i).$$

Unlike a measure on a σ-algebra, this definition initially requires only finite additivity because an algebra need not be closed under countable unions.

Lemma 2.2 Let μ be a set function on a semi-algebra $\mathcal{S}$ satisfying (1). Define $\bar{\mu}$ on $\bar{\mathcal{S}}$ by writing A as a finite disjoint union $A = \biguplus_{j=1}^{n} S_j$ with $S_j \in \mathcal{S}$ and setting

$$\bar{\mu}(A) = \sum_{j=1}^{n} \mu(S_j).$$

This definition is independent of the chosen representation, and the following properties hold:

1. If $A, B_1, \ldots, B_n \in \bar{\mathcal{S}}$ and $A = \biguplus_{i=1}^{n} B_i$, then

$$\bar{\mu}(A) = \sum_{i=1}^{n} \bar{\mu}(B_i).$$

2. If $A, B_1, \ldots, B_n \in \bar{\mathcal{S}}$ and $A \subseteq \bigcup_{i=1}^n B_i$, then
$$\bar{\mu}(A) \leq \sum_{i=1}^n \bar{\mu}(B_i).$$

Proof. To prove that $\bar{\mu}$ is well defined, suppose that a set has two finite disjoint representations,
$$A = \biguplus_{j=1}^n S_j = \biguplus_{k=1}^m T_k,$$
where every $S_j, T_k \in \mathcal{S}$. For each j, the sets $S_j \cap T_1, \ldots, S_j \cap T_m$ form a finite disjoint decomposition of S_j. Similarly, for each k, the sets $S_1 \cap T_k, \ldots, S_n \cap T_k$ form a finite disjoint decomposition of T_k. Therefore,
$$\sum_{j=1}^n \mu(S_j) = \sum_{j=1}^n \sum_{k=1}^m \mu(S_j \cap T_k) = \sum_{k=1}^m \sum_{j=1}^n \mu(S_j \cap T_k) = \sum_{k=1}^m \mu(T_k),$$
proving that $\bar{\mu}$ is well defined.

For (1), write
$$B_i = \biguplus_{j=1}^{k_i} S_{ij},$$
where $S_{ij} \in \mathcal{S}$. Since the sets B_i are pairwise disjoint,
$$A = \biguplus_{i=1}^n \biguplus_{j=1}^{k_i} S_{ij},$$
and hence
$$\bar{\mu}(A) = \sum_{i=1}^n \sum_{j=1}^{k_i} \mu(S_{ij}) = \sum_{i=1}^n \bar{\mu}(B_i).$$

For (2), define $C_i = A \cap B_i$, so $A = \bigcup_{i=1}^n C_i$. Disjointify this cover by setting $D_1 = C_1$ and
$$D_i = C_i \setminus \bigcup_{j=1}^{i-1} C_j, \quad 2 \leq i \leq n.$$

Then $A = \biguplus_{i=1}^n D_i$, and $D_i \subseteq C_i \subseteq B_i$. By (1) and monotonicity,

$$\bar{\mu}(A) = \sum_{i=1}^n \bar{\mu}(D_i) \leq \sum_{i=1}^n \bar{\mu}(B_i).$$

□

Proof of Theorem 1.14. Lemma 1.16 describes the generated algebra, and Lemma 2.2 constructs the unique extension to it. The countable covering condition (2) also transfers to $\bar{\mu}$. To see this, decompose every set in a countable cover into its finitely many semialgebra pieces, intersect those pieces with a disjoint semialgebra decomposition of the covered set, and apply (2) before summing over that finite decomposition. Consequently, if $A, A_1, A_2, \dots \in \bar{\mathcal{S}}$ and $A \subseteq \bigcup_n A_n$, then

$$\bar{\mu}(A) \leq \sum_{n=1}^\infty \bar{\mu}(A_n).$$

In particular, $\bar{\mu}$ is countably additive whenever a disjoint union remains in $\bar{\mathcal{S}}$.

For every $E \subseteq \Omega$, define

$$\mu^*(E) = \inf \left\{ \sum_{n=1}^\infty \bar{\mu}(A_n) : E \subseteq \bigcup_{n=1}^\infty A_n,\ A_n \in \bar{\mathcal{S}} \right\}.$$

The empty cover shows that $\mu^*(\varnothing) = 0$, and monotonicity is immediate. To prove countable subadditivity, cover each E_k by sets from $\bar{\mathcal{S}}$ whose total measure is within $\varepsilon 2^{-k}$ of $\mu^*(E_k)$, then combine all these covers and let $\varepsilon \downarrow 0$. Thus μ^* is an outer measure.

Let $\mathcal{M}$ be the class of sets $C \subseteq \Omega$ satisfying the Carathéodory identity

$$\mu^*(E) = \mu^*(E \cap C) + \mu^*(E \setminus C) \quad \text{for every } E \subseteq \Omega. \tag{2.1}$$

The class $\mathcal{M}$ contains Ω and is closed under complements. Applying (2.1) successively to two members $C, D \in \mathcal{M}$ and using outer subadditivity shows that $C \cup D \in \mathcal{M}$. Thus $\mathcal{M}$ is an algebra. If $C_1, C_2, \dots$ are pairwise disjoint members of $\mathcal{M}$, repeated use of (2.1) gives, for every r,

$$\mu^*(E) \geq \sum_{k=1}^r \mu^*(E \cap C_k) + \mu^* \left(E \setminus \bigcup_{k=1}^r C_k \right).$$

Letting $r \to \infty$ and using countable subadditivity proves (2.1) for $\bigcup_k C_k$. An arbitrary countable union can be disjointified within the algebra, so $\mathcal{M}$ is a σ-algebra. The same calculation shows that μ^* is countably additive on $\mathcal{M}$.

Every $A \in \bar{\mathcal{S}}$ belongs to $\mathcal{M}$. Indeed, if (B_n) is any algebra cover of E, finite additivity gives

$$\bar{\mu}(B_n) = \bar{\mu}(B_n \cap A) + \bar{\mu}(B_n \setminus A).$$

Summing and then taking the infimum over all covers yields the inequality in (2.1) that is opposite to outer subadditivity. Moreover, covering A by itself and using countable subadditivity of $\bar{\mu}$ for the reverse bound gives

$$\mu^*(A) = \bar{\mu}(A).$$

Therefore $\sigma(\mathcal{S}) = \sigma(\bar{\mathcal{S}}) \subseteq \mathcal{M}$, and μ^* restricted to this σ-algebra is the required extension.

Finally, suppose that Ω is covered by increasing sets $C_k \in \bar{\mathcal{S}}$ with $\bar{\mu}(C_k) < \infty$. If two extensions agree on $\bar{\mathcal{S}}$, the finite-measure λ–π uniqueness argument in Theorem 8.8 shows that they agree on $E \cap C_k$ for every $E \in \sigma(\mathcal{S})$. Continuity from below then gives agreement on E itself. This proves uniqueness under σ-finiteness. □

Recall that if μ is a measure on $(\mathbb{R}, \mathcal{R})$ that is finite on bounded intervals, then $\mu((a,b]) = F(b) - F(a)$ for a nondecreasing, right-continuous function F. The next theorem establishes the converse.

Theorem 2.3 (Lebesgue–Stieltjes measure) For every nondecreasing right-continuous function $F\colon \mathbb{R} \to \mathbb{R}$, there is a unique measure μ on $(\mathbb{R}, \mathcal{R})$ such that

$$\mu((a,b]) = F(b) - F(a)$$

whenever $a < b$ are finite. When $F(x) = x$, the resulting measure is called **Lebesgue measure**.

Proof. Define $\mu_0((a,b]) = F(b) - F(a)$ on the semi-algebra of half-open intervals. By Theorem 1.14, it is enough to verify finite additivity and countable subadditivity.

1. Suppose that

$$(a,b] = \biguplus_{j=1}^{n} (a_j, b_j].$$

After relabeling the intervals, we have $a_1 = a$, $b_j = a_{j+1}$ for $1 \le j < n$, and $b_n = b$. Therefore, the sum telescopes:

$$\sum_{j=1}^{n} \mu_0((a_j, b_j]) = \sum_{j=1}^{n} \big(F(b_j) - F(a_j)\big) = F(b) - F(a) = \mu_0((a, b]).$$

For unbounded intervals, define

$$F(\infty) = \lim_{x \to \infty} F(x) \qquad \text{and} \qquad F(-\infty) = \lim_{x \to -\infty} F(x).$$

These limits exist because F is nondecreasing, and the same telescoping argument applies.

2. Suppose that

$$(a, b] \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i].$$

First assume that a and b are finite. Fix $\varepsilon > 0$. Right continuity provides $\delta > 0$ and numbers $\delta_i > 0$ such that

$$\mu_0((a + \delta, b]) \geq \mu_0((a, b]) - \varepsilon$$

and

$$\mu_0((a_i, b_i + \delta_i]) \leq \mu_0((a_i, b_i]) + \frac{\varepsilon}{2^i}$$

for every i. The open intervals $(a_i, b_i + \delta_i)$ cover the compact interval $[a + \delta, b]$, so a finite subcollection indexed by I also covers it. Finite subadditivity, which follows from (1), now gives

$$\begin{aligned} \mu_0((a, b]) - \varepsilon &\leq \mu_0((a + \delta, b]) \\ &\leq \sum_{i \in I} \mu_0((a_i, b_i + \delta_i]) \\ &\leq \sum_{i=1}^{\infty} \mu_0((a_i, b_i]) + \varepsilon. \end{aligned}$$

Letting $\varepsilon \downarrow 0$ proves the desired inequality.

If either endpoint is infinite, apply the argument for finite endpoints to the intervals $(a, b] \cap [-k, k]$ and let $k \to \infty$. For example, when $a = -\infty$ and $b = \infty$,

$$F(k) - F(-k) \leq \sum_{i=1}^{\infty} \big(F(b_i) - F(a_i)\big),$$

and the left-hand side converges to $F(\infty) - F(-\infty)$.

Thus μ_0 extends to a measure on $\mathcal{R}$. The extension is σ-finite because $[-k,k] \uparrow \mathbb{R}$ and $\mu_0([-k,k]) < \infty$ for every k. Uniqueness now follows from Theorem 1.14. When $F(x) = x$, the resulting measure is Lebesgue measure. $\square$

The one-dimensional construction extends to $\mathbb{R}^d$, but there are some subtle differences; coordinatewise monotonicity and right continuity are no longer sufficient. For example, let

$$u(x) = \frac{1}{1+e^{-x}} \qquad \text{and} \qquad F(x,y) = u(x) + u(y) - u(x)u(y).$$

The function F is continuous, takes values in $(0,1)$, and is nondecreasing in each coordinate. However, its increment over a nondegenerate rectangle $(a_1,b_1] \times (a_2,b_2]$ is

$$-\big(u(b_1) - u(a_1)\big)\big(u(b_2) - u(a_2)\big) < 0,$$

so it cannot determine a measure through rectangular increments.

Thus, we need a stronger condition to ensure that every axis-aligned half-open rectangle receives nonnegative mass.

Remark 2.4 If $A = (a_1,b_1] \times (a_2,b_2] \subseteq \mathbb{R}^2$, inclusion–exclusion requires

$$\mu(A) = F(a_1,a_2) + F(b_1,b_2) - F(a_1,b_2) - F(b_1,a_2) \geq 0.$$

More generally, let $A = \prod_{i=1}^d (a_i,b_i]$ and let $V = \prod_{i=1}^d \{a_i,b_i\}$ be its set of 2^d vertices. For $v = (c_1,\dots,c_d) \in V$, let $\ell(v)$ be the number of coordinates for which $c_i = a_i$. Then

$$\mu(A) = \sum_{v \in V} (-1)^{\ell(v)} F(v).$$

Theorem 2.5 Suppose that $F\colon \mathbb{R}^d \to [0,1]$ satisfies the following conditions:

1. The function F is nondecreasing in each coordinate.
2. The function F is right continuous from above: if $\boldsymbol{y}_n \downarrow \boldsymbol{x}$ coordinatewise, then $F(\boldsymbol{y}_n) \downarrow F(\boldsymbol{x})$.

3. Every finite rectangle A with vertex set V has a nonnegative increment:

$$\sum_{v\in V}(-1)^{\ell(v)}F(v) \geq 0.$$

Then there is a unique measure μ on $(\mathbb{R}^d, \mathcal{R}^d)$ such that

$$\mu(A) = \sum_{v\in V}(-1)^{\ell(v)}F(v)$$

for every finite half-open rectangle A.

Proof. For $m \geq 1$, let $Q_m = (-m, m]^d$ and let $\mathcal{S}_m$ be the semialgebra of half-open rectangles contained in Q_m. Define $\mu_{0,m}$ on $\mathcal{S}_m$ by the displayed alternating sum. Condition (3) makes $\mu_{0,m}$ nonnegative. If a rectangle is partitioned by a hyperplane perpendicular to one coordinate axis, the two alternating sums telescope in that coordinate and add to the sum for the original rectangle. Repeating this argument proves finite additivity for every finite rectangular partition.

To verify countable subadditivity, enlarge the covering rectangles slightly in every coordinate. Condition (2) makes the change in $\mu_{0,m}$ arbitrarily small. After shrinking the covered rectangle slightly, compactness supplies a finite subcover. Finite subadditivity then gives the desired inequality, and letting the enlargements and shrinkage tend to zero removes the approximation error. Thus the hypotheses of Theorem 1.14 hold, so $\mu_{0,m}$ extends uniquely to a finite Borel measure μ_m on Q_m.

The measures are consistent: if $k < m$, then μ_m and μ_k agree on the rectangular semialgebra in Q_k, so uniqueness gives $\mu_m|_{Q_k} = \mu_k$. Put $D_1 = Q_1$, $D_m = Q_m \setminus Q_{m-1}$ for $m \geq 2$, and define

$$\mu(A) = \sum_{m=1}^{\infty} \mu_m(A \cap D_m), \qquad A \in \mathcal{R}^d.$$

This is a measure on $\mathcal{R}^d$. Every finite rectangle is contained in some Q_m, where consistency shows that μ has the required increment. Finally, any other such measure agrees with μ_m on every Q_m and hence agrees with μ on all Borel sets.

□

Random Variables and Distributions

Definition 2.6 Let $(\Omega, \mathcal{F})$ be a measurable space. A function $X\colon \Omega \to \mathbb{R}$ is a **random variable** if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for every $B \in \mathcal{R}$. This definition depends only on the measurable space; no probability measure is required.

Example 2.7 If $A \in \mathcal{F}$, then the indicator function

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A \end{cases}$$

is a random variable. Indeed, the inverse image of any Borel set is one of $\emptyset$, A, A^c, or Ω.

Definition 2.8 If X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, then X induces the **distribution measure** μ_X on $(\mathbb{R}, \mathcal{R})$ defined by

$$\mu_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)), \quad B \in \mathcal{R}.$$

The identities $\mu_X(\emptyset) = 0$ and $\mu_X(\mathbb{R}) = 1$ follow immediately from the corresponding properties of $\mathbb{P}$. If $A_1, A_2, \ldots$ are pairwise disjoint Borel sets, then their inverse images are also pairwise disjoint, and

$$\mu_X\left(\bigcup_{j=1}^{\infty} A_j\right) = \mathbb{P}\left(X^{-1}\left(\bigcup_{j=1}^{\infty} A_j\right)\right) = \mathbb{P}\left(\bigcup_{j=1}^{\infty} X^{-1}(A_j)\right) = \sum_{j=1}^{\infty} \mathbb{P}(X^{-1}(A_j)) = \sum_{j=1}^{\infty} \mu_X(A_j).$$

In applications, the underlying measurable space $(\Omega, \mathcal{F})$ is often not "observed" directly; rather, the distribution of X is the accessible object.

Definition 2.9 Let X be a random variable with distribution μ_X. Its **cumulative distribution function** is

$$F_X(x) = \mathbb{P}(X \leq x) = \mu_X((-\infty, x]).$$

This function determines the distribution uniquely because

$$\mu_X((a,b]) = F_X(b) - F_X(a),$$

and Theorem 2.3 gives a unique extension from half-open intervals to all Borel sets.

Lemma 2.10 Let F be the cumulative distribution function of X. Then the following hold:

1. F is nondecreasing.
2. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
3. F is right continuous.
4. If $F(x-) = \lim_{y\uparrow x} F(y)$, then $F(x-) = \mathbb{P}(X < x)$.
5. $\mathbb{P}(X = x) = F(x) - F(x-)$.

Proof. For (1), if $x \le y$, then $\{X \le x\} \subseteq \{X \le y\}$, so $F(x) \le F(y)$.

For (2), the events $\{X \le n\}$ increase to Ω, whereas the events $\{X \le -n\}$ decrease to $\emptyset$. Continuity of probability from below and above gives

$$F(n) \to 1 \qquad \text{and} \qquad F(-n) \to 0.$$

Monotonicity then yields the stated limits along the full real line.

For (3), if $x_n \downarrow x$, then $\{X \le x_n\} \downarrow \{X \le x\}$. Hence $F(x_n) \downarrow F(x)$.

For (4), choose any sequence $y_n \uparrow x$. Then $\{X \le y_n\} \uparrow \{X < x\}$, so

$$F(x-) = \lim_{n\to\infty} F(y_n) = \mathbb{P}(X < x).$$

Finally, (5) follows from the disjoint decomposition $\{X \le x\} = \{X < x\} \,\dot\cup\, \{X = x\}$. □

Theorem 2.11 If a function $F\colon \mathbb{R} \to [0,1]$ satisfies (1)–(3), then it is the cumulative distribution function of some random variable.

Proof. By Theorem 2.3, the rule

$$\mu((a,b]) = F(b) - F(a)$$

extends uniquely to a Borel measure on $\mathbb{R}$. The boundary limits in (2) give $\mu(\mathbb{R}) = 1$, so μ is a probability measure. On the probability space $(\mathbb{R}, \mathcal{R}, \mu)$, let $X(x) = x$. Then

$$\mathbb{P}(X \leq t) = \mu((-\infty, t]) = F(t),$$

so F is the cumulative distribution function of X. $\square$

On the probability space $\Omega = (0,1)$ with its Borel σ-field and Lebesgue probability measure, the identity map $U(t) = t$ has the uniform distribution on $(0,1)$.

The graph in Figure 1 displays its cumulative distribution function.

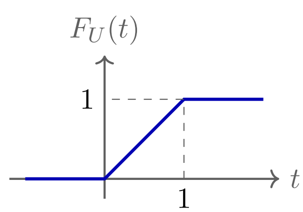

FIGURE 1

> **Definition 2.12** For $u \in (0,1)$, the **generalized inverse** of a cumulative distribution function F is
> $$F^{-1}(u) = \sup\{y \in \mathbb{R} : F(y) < u\}.$$

The construction is illustrated in Figure 2: move horizontally from the height u to the graph and then vertically to the real line.

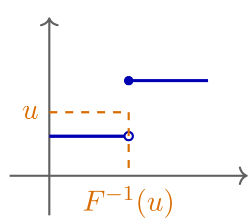

FIGURE 2

Theorem 2.13 (Probability integral transform) If U is uniformly distributed on $(0,1)$ and F is a cumulative distribution function, then $F^{-1}(U)$ has cumulative distribution function F.

Proof. We first prove that, for $u \in (0,1)$ and $x \in \mathbb{R}$,

$$F^{-1}(u) \le x \qquad \text{if and only if} \qquad u \le F(x).$$

Forward implication. Assume that $F^{-1}(u) \le x$. For every $\delta > 0$, the definition of the supremum implies $F(x+\delta) \ge u$. Letting $\delta \downarrow 0$ and using the right continuity of F gives $F(x) \ge u$.

Reverse implication. Assume that $u \le F(x)$. If $y > x$, then monotonicity gives $F(y) \ge F(x) \ge u$, so no such y belongs to $\{z : F(z) < u\}$. Hence its supremum is at most x, and $F^{-1}(u) \le x$.

The weak inequalities are essential; the analogous statement with strict inequalities need not hold at a jump of F. We now obtain

$$\mathbb{P}(F^{-1}(U) \le x) = \mathbb{P}(U \le F(x)) = F(x),$$

which proves the result. $\square$

Lecture 3 — Monday, September 24, 2018

The proof of Theorem 2.13 is constructive and gives the following inverse transform sampling procedure for simulating a random variable with cumulative distribution function F:

1. Generate a random variable U that is uniform on $(0,1)$.
2. Set $X = F^{-1}(U)$.

By Theorem 2.13, the resulting X has cumulative distribution function F.

Example 3.1 Construct a $\mathrm{Bin}(n,p)$ random variable from independent uniform random variables.

Solution. Generate independent random variables $U_1,\ldots,U_n$, each uniform on $(0,1)$, and define $X_i = \mathbb{1}_{\{U_i \le p\}}$. Then the X_i are independent and each has the Bernoulli distribution with success probability p. Consequently,

$$X = \sum_{i=1}^{n} X_i$$

has the binomial distribution with parameters n and p. This is the construction commonly used in computer simulation. $\square$

Example 3.2 Construct standard normal random variables from random variables whose cumulative distribution functions have closed-form inverses.

Solution. Direct inversion is inconvenient because the standard normal cumulative distribution function

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,\mathrm{d}y\,,$$

has no closed-form inverse. Instead, the Box–Muller construction uses polar coordinates. Generate independent random variables U_1 and U_2, each uniformly distributed on $(0,1)$, and proceed as follows:

1. Set $R = \sqrt{-2\log U_1}$, so R^2 has the chi-squared distribution with two degrees of freedom.
2. Set $\Theta = 2\pi U_2$, so Θ is uniformly distributed on $(0,2\pi)$ and is independent of R.
3. Set $X_1 = R\cos\Theta$ and $X_2 = R\sin\Theta$.

The joint change of variables from (R,Θ) to (X_1,X_2) gives the joint density

$$\frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right) = \frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}.$$

Hence X_1 and X_2 are independent standard normal random variables. $\square$

Other familiar distributions can be simulated by the same principles.

Definition 3.3 Suppose that X has cumulative distribution function F and there exists a nonnegative measurable function f such that

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y\,,$$

for every $x \in \mathbb{R}$. Then f is called a **probability density function** of X, or of F.

Lemma 3.4 If X has a probability density function, then $\mathbb{P}(X = t) = 0$ for every $t \in \mathbb{R}$.

Proof. For every $\delta > 0$,

$$0 \le \mathbb{P}(X = t) \le \mathbb{P}(t - \delta < X \le t + \delta) = \int_{t-\delta}^{t+\delta} f(y)\,\mathrm{d}y\,.$$

The final integral tends to 0 as $\delta \downarrow 0$, so the squeeze theorem gives $\mathbb{P}(X = t) = 0$. $\square$

Measurable Functions

Random variables are instances of measurable maps. Practical criteria based on generating classes and closure properties make it possible to verify measurability for the constructions used throughout probability theory.

Definition 3.5 Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be measurable spaces. A function $X\colon \Omega \to S$ is a **measurable map** if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for every $B \in \mathcal{S}$.

Theorem 3.6 Suppose that $\mathcal{A}$ is a collection of subsets of S such that $\sigma(\mathcal{A}) = \mathcal{S}$. If $X^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{A}$, then X is measurable.

Proof. Let

$$\mathcal{B} = \{B \subseteq S : X^{-1}(B) \in \mathcal{F}\}.$$

The proof has two steps:

1. First, show that $\mathcal{B}$ is a σ-algebra.
2. Second, show that $\mathcal{A} \subseteq \mathcal{B}$.

(This strategy is very useful in practice: it is often easier to check measurability on a generating class than on the full σ-algebra.)

To prove (1), note first that $X^{-1}(S) = \Omega \in \mathcal{F}$. If $C \in \mathcal{B}$, then $X^{-1}(C^c) = (X^{-1}(C))^c \in \mathcal{F}$. If $C_1, C_2, \dots \in \mathcal{B}$, then

$$X^{-1}\left(\bigcup_{i=1}^{\infty} C_i\right) = \bigcup_{i=1}^{\infty} X^{-1}(C_i) \in \mathcal{F}.$$

Hence $\mathcal{B}$ is a σ-algebra. Step (2) is precisely the hypothesis. Therefore $\mathcal{S} = \sigma(\mathcal{A}) \subseteq \mathcal{B}$, and X is measurable. □

Definition 3.7 If $X \colon \Omega \to S$ and $\mathcal{S}$ is a σ-algebra on S, then

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{S}\}$$

is the ***σ*-algebra generated by *X***. It is the smallest σ-algebra on Ω that makes X measurable.

Theorem 3.8 Let $(\Omega, \mathcal{F})$, $(S, \mathcal{S})$, and $(T, \mathcal{T})$ be measurable spaces. If $X \colon \Omega \to S$ and $f \colon S \to T$ are measurable, then $f \circ X \colon \Omega \to T$ is measurable.

Proof. For every $A \in \mathcal{T}$,

$$(f \circ X)^{-1}(A) = X^{-1}(f^{-1}(A)),$$

because

$$\omega \in (f \circ X)^{-1}(A) \iff f(X(\omega)) \in A \iff X(\omega) \in f^{-1}(A) \iff \omega \in X^{-1}(f^{-1}(A)).$$

Since f is measurable, $f^{-1}(A) \in \mathcal{S}$; since X is measurable, its inverse image belongs to $\mathcal{F}$. □

Theorem 3.9 If $X_1, \dots, X_n$ are random variables and $f \colon \mathbb{R}^n \to \mathbb{R}$ is Borel measurable, then $f(X_1, \dots, X_n)$ is a random variable.

Proof. Let $\boldsymbol{X} = (X_1, \dots, X_n)$. The open rectangles

$$\mathcal{A} = \left\{ \prod_{i=1}^{n} (a_i, b_i) : -\infty \leq a_i < b_i \leq \infty \right\}$$

generate $\mathcal{R}^n$, because every open set is a countable union of open rectangles. The inverse image of an open rectangle is

$$\boldsymbol{X}^{-1}\left(\prod_{i=1}^{n} (a_i, b_i) \right) = \bigcap_{i=1}^{n} X_i^{-1}((a_i, b_i)) \in \mathcal{F}.$$

Thus $\boldsymbol{X}$ is measurable by Theorem 3.6. Applying Theorem 3.8 to $f \circ \boldsymbol{X}$ completes the proof. □

Proposition 3.10 Every continuous function $f \colon \mathbb{R}^n \to \mathbb{R}$ is Borel measurable.

Proof. If f is continuous, then the inverse image of every open set is open. Since every open set is Borel, f is Borel measurable. □

Corollary 3.11 If $X_1, \dots, X_n$ are random variables and $f \colon \mathbb{R}^n \to \mathbb{R}$ is continuous, then $f(X_1, \dots, X_n)$ is a random variable. In particular, $\sum_{i=1}^{n} X_i$ is a random variable.

Proof. Every continuous function is Borel measurable, so the result follows from Theorem 3.9. □

Theorem 3.12 Let $X_1, X_2, \dots$ be random variables on $(\Omega, \mathcal{F})$. Whenever the resulting functions are finite valued,

$$\inf_{n \geq 1} X_n, \qquad \sup_{n \geq 1} X_n, \qquad \liminf_{n \to \infty} X_n, \qquad \text{and} \qquad \limsup_{n \to \infty} X_n$$

are random variables.

Proof. For every $x \in \mathbb{R}$,

$$\left\{ \inf_{n \geq 1} X_n \geq x \right\} = \bigcap_{n=1}^{\infty} \{X_n \geq x\} \in \mathcal{F}.$$

The half-lines $[x, \infty)$ generate the Borel σ-algebra, so Theorem 3.6 shows that $\inf_{n \geq 1} X_n$ is measurable. Since

$$\sup_{n\geq 1} X_n = -\inf_{n\geq 1}(-X_n),$$

the supremum is measurable as well. Finally,

$$\liminf_{n\to\infty} X_n = \sup_{k\geq 1}\inf_{n\geq k} X_n \qquad \text{and} \qquad \limsup_{n\to\infty} X_n = \inf_{k\geq 1}\sup_{n\geq k} X_n.$$

Each expression is measurable by the two cases already proved. $\square$

Definition 3.13 A sequence $(X_n)_{n\geq 1}$ **converges almost surely** to a random variable X if

$$\mathbb{P}\big(\{\omega : X_n(\omega) \to X(\omega)\}\big) = 1.$$

In particular, the set on which (X_n) has a finite limit is

$$\Omega_0 = \left\{\omega : \limsup_{n\to\infty} X_n(\omega) = \liminf_{n\to\infty} X_n(\omega) \in \mathbb{R}\right\} \in \mathcal{F}.$$

Definition 3.14 The set

$$\overline{\mathbb{R}} = [-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\}$$

is the **extended real line**. Its Borel σ-algebra is

$$\overline{\mathcal{R}} = \sigma\big(\{[-\infty, a), (a, b), (b, \infty] : a, b \in \mathbb{R}\}\big).$$

A measurable map $X\colon (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \overline{\mathcal{R}})$ is an **extended random variable**. Unless explicitly stated otherwise, a random variable is real valued.

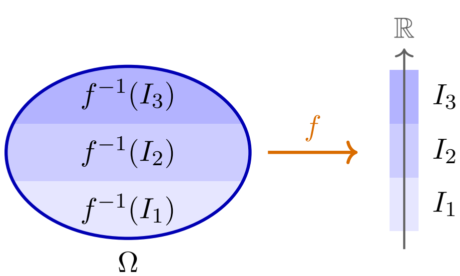

FIGURE 3

Construction of the Lebesgue Integral

Integration with respect to a measure is the basis of expectation, which appears ubiquitously in statistics. Unlike Riemann integration, which partitions a topological domain, Lebesgue integration partitions the range of a measurable function. Consequently, it applies even when the underlying sample space has no metric or topology. The schematic in Figure 3 emphasizes that the construction depends on measurable level sets rather than geometric intervals in Ω. This is particularly useful in statistics, since we often observe only the values of a random variable rather than the structure of the underlying probability space. We cannot simply chop up the stock market or the weather into intervals; we can only observe the outcomes of random variables. The Lebesgue integral is the natural tool for this situation.

Throughout our construction, let μ be a σ-finite measure on $(\Omega, \mathcal{F})$. The integral is defined successively for simple functions, bounded measurable functions with support of finite measure, nonnegative measurable functions, and finally integrable real-valued functions.

Definition 3.15 A measurable function $\varphi\colon \Omega \to \mathbb{R}$ is a **simple function** if

$$\varphi = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i},$$

where $a_1, \ldots, a_n \in \mathbb{R}$ and $A_1, \ldots, A_n \in \mathcal{F}$ are pairwise disjoint sets of finite measure. Its integral is

$$\int \varphi \, \mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

Exercise 3.16 Prove that the value in Definition 3.15 is independent of the chosen representation. More precisely, if

$$\varphi = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i} = \sum_{j=1}^{m} b_j \mathbb{1}_{B_j},$$

prove that

$$\sum_{i=1}^{n} a_i \mu(A_i) = \sum_{j=1}^{m} b_j \mu(B_j).$$

Solution. Let
$$C = \bigcup_{i=1}^{n} A_i \cup \bigcup_{j=1}^{m} B_j,$$
and complete both partitions on C by adjoining
$$A_0 = C \setminus \bigcup_{i=1}^{n} A_i, \qquad B_0 = C \setminus \bigcup_{j=1}^{m} B_j, \qquad \text{and} \qquad a_0 = b_0 = 0.$$
The sets $A_i \cap B_j$ form a finite partition of C. Whenever this intersection is nonempty, equality of the two representations gives $a_i = b_j$ on it. Consequently,
$$\sum_{i=0}^{n} a_i \mu(A_i) = \sum_{i=0}^{n} \sum_{j=0}^{m} a_i \mu(A_i \cap B_j) = \sum_{i=0}^{n} \sum_{j=0}^{m} b_j \mu(A_i \cap B_j) = \sum_{j=0}^{m} b_j \mu(B_j).$$
The terms with index zero vanish, proving the desired equality. □

Lecture 4 — Monday, October 01, 2018

Lemma 4.1 Let φ and ψ be simple functions. Then the following statements hold:

1. If $\varphi \geq 0$, then
$$\int \varphi \, \mathrm{d}\mu \geq 0.$$
2. For every $a \in \mathbb{R}$,
$$\int a\varphi \, \mathrm{d}\mu = a \int \varphi \, \mathrm{d}\mu \,.$$
3. The simple integral is additive:
$$\int (\varphi + \psi) \, \mathrm{d}\mu = \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\mu \,.$$

Proof. Statements (1) and (2) follow directly from Definition 3.15. For (3), write

$$\varphi = \sum_{i=1}^{p} a_i \mathbb{1}_{A_i} \qquad \text{and} \qquad \psi = \sum_{j=1}^{q} b_j \mathbb{1}_{B_j}.$$

Let $C = \bigcup_{i=1}^p A_i \cup \bigcup_{j=1}^q B_j$, and complete each partition on C by setting

$$A_0 = C \setminus \bigcup_{i=1}^{p} A_i, \qquad B_0 = C \setminus \bigcup_{j=1}^{q} B_j, \qquad \text{and} \qquad a_0 = b_0 = 0.$$

Because C has finite measure, all the sets in these completed partitions have finite measure. Refining the partitions by their pairwise intersections gives

$$\varphi + \psi = \sum_{i=0}^{p} \sum_{j=0}^{q} (a_i + b_j) \mathbb{1}_{A_i \cap B_j} \quad \text{on } C,$$

and all three functions vanish outside C. Therefore,

$$\int (\varphi + \psi) \, \mathrm{d}\mu = \sum_{i=0}^{p} \sum_{j=0}^{q} (a_i + b_j) \mu(A_i \cap B_j) = \sum_{i=0}^{p} a_i \mu(A_i) + \sum_{j=0}^{q} b_j \mu(B_j) = \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\mu \, .$$

□

Corollary 4.2 For simple functions φ and ψ, the following statements hold:

1. If $\varphi \leq \psi$ almost everywhere, then
$$\int \varphi \, \mathrm{d}\mu \leq \int \psi \, \mathrm{d}\mu \, .$$
2. If $\varphi = \psi$ almost everywhere, then
$$\int \varphi \, \mathrm{d}\mu = \int \psi \, \mathrm{d}\mu \, .$$
3. The absolute value of the integral is bounded by the integral of the absolute value:
$$\left| \int \varphi \, \mathrm{d}\mu \right| \leq \int |\varphi| \, \mathrm{d}\mu \, .$$

Proof. For (1), write the simple function $\psi - \varphi$ as

$$\psi - \varphi = \sum_{k=1}^{r} c_k \mathbb{1}_{C_k}$$

on pairwise disjoint sets C_k of finite measure. If $c_k < 0$, then C_k is contained in the null set on which $\psi - \varphi < 0$, so $\mu(C_k) = 0$. Every nonzero contribution to the integral is therefore nonnegative. By (3),

$$0 \leq \int (\psi - \varphi)\,\mathrm{d}\mu = \int \psi\,\mathrm{d}\mu - \int \varphi\,\mathrm{d}\mu\,.$$

Applying (1) in both directions proves (2). Finally, $-|\varphi| \leq \varphi \leq |\varphi|$, so two applications of (1) yield

$$\left|\int \varphi\,\mathrm{d}\mu\right| \leq \int |\varphi|\,\mathrm{d}\mu\,.$$

□

Definition 4.3 The **support** of f is

$$\mathrm{supp}(f) = \{\omega \in \Omega : f(\omega) \neq 0\}.$$

In this construction, a **bounded function** means a measurable function f for which $|f| \leq M$ for some finite $M > 0$ and $\mu(\mathrm{supp}(f)) < \infty$.

Definition 4.4 For a bounded measurable function f, define

$$\int f\,\mathrm{d}\mu = \sup_{\varphi \leq f} \int \varphi\,\mathrm{d}\mu\,,$$

where the supremum is over simple functions. Equivalently,

$$\int f\,\mathrm{d}\mu = \inf_{\psi \geq f} \int \psi\,\mathrm{d}\mu\,.$$

Here the infimum is likewise taken over simple functions.

To verify the equivalent formula in Definition 4.4, choose a finite-measure set E supporting f and a constant M such that $|f| \leq M$. For $n \geq 1$, let

$$E_{k,n} = \left\{\omega \in E : \frac{(k-1)M}{n} < f(\omega) \leq \frac{kM}{n}\right\},$$

and define

$$\psi_n = \sum_{k=-n}^{n} \frac{kM}{n} \mathbb{1}_{E_{k,n}} \qquad \text{and} \qquad \varphi_n = \sum_{k=-n}^{n} \frac{(k-1)M}{n} \mathbb{1}_{E_{k,n}}.$$

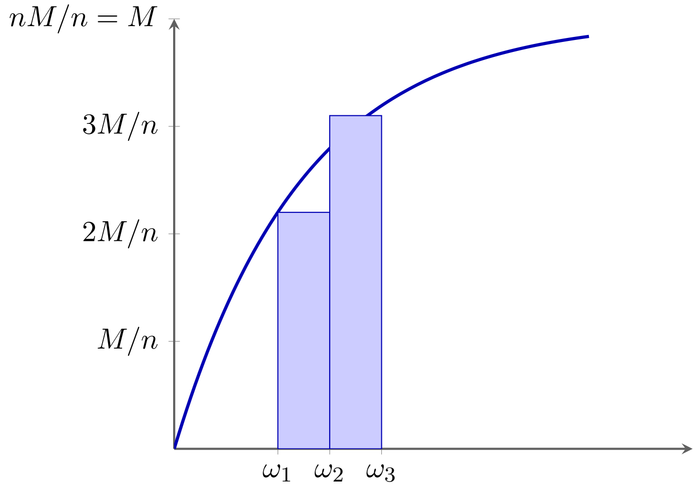

FIGURE 4

As illustrated in Figure 4, these functions approximate f from above and below. In fact, $\varphi_n \leq f \leq \psi_n$ on E, both functions vanish off E, and $\psi_n - \varphi_n = (M/n)\mathbb{1}_E$. Therefore

$$\inf_{\psi \geq f} \int \psi \, \mathrm{d}\mu \leq \int \psi_n \, \mathrm{d}\mu = \int \varphi_n \, \mathrm{d}\mu + \frac{M}{n}\mu(E) \leq \sup_{\varphi \leq f} \int \varphi \, \mathrm{d}\mu + \frac{M}{n}\mu(E).$$

Letting $n \to \infty$ proves the required reverse inequality. The opposite inequality follows from monotonicity because every pair satisfying $\varphi \leq f \leq \psi$ also satisfies $\varphi \leq \psi$.

Proposition 4.5 If f and g are bounded measurable functions and $a \in \mathbb{R}$, then the following hold:

1. If $f \geq 0$ almost everywhere, then $\int f \, \mathrm{d}\mu \geq 0$.
2. $\int af \, \mathrm{d}\mu = a \int f \, \mathrm{d}\mu$.
3. $\int (f+g) \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu$.

Proof. For (1), let $N = \{f < 0\}$. If $|f| \leq M$, then the simple function $-M\mathbb{1}_N$ is bounded above by f pointwise. Since $\mu(N) = 0$, its integral is zero, so the defining supremum is nonnegative.

If $a > 0$, the correspondence $\varphi \mapsto a\varphi$ is a bijection between simple minorants of f and simple minorants of af. Hence

$$\int af \, \mathrm{d}\mu = \sup_{\varphi \le f} \int a\varphi \, \mathrm{d}\mu = a \int f \, \mathrm{d}\mu \,.$$

The case $a = 0$ follows from (1). If $a < 0$, use the upper-approximation formula and the substitution $\psi = a\varphi$ to obtain the same identity. This proves (2).

For additivity, simple minorants $\varphi \le f$ and $\psi \le g$ satisfy $\varphi + \psi \le f + g$, and therefore

$$\int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu \le \int (f + g) \, \mathrm{d}\mu \,.$$

Apply this inequality to $-f$ and $-g$ and use (2) with $a = -1$ to obtain the reverse inequality. $\square$

The next step is to extend our integral to nonnegative functions. To this end, for $E \in \mathcal{F}$ and every measurable f for which the right-hand side is defined, we use the notation

$$\int_E f \, \mathrm{d}\mu = \int f \mathbb{1}_E \, \mathrm{d}\mu \,.$$

Definition 4.6 If $f \colon \Omega \to [0, \infty]$ is measurable, define

$$\int f \, \mathrm{d}\mu = \sup \left\{ \int h \, \mathrm{d}\mu : 0 \le h \le f, \ h \text{ is bounded} \right\}.$$

The supremum in Definition 4.6 always exists in $[0, \infty]$ (but a computer usually cannot find it). The following truncation formula makes the definition usable.

Lemma 4.7 (Truncation formula) Suppose that $E_n \nearrow \Omega$ and $\mu(E_n) < \infty$ for every n. If $f \ge 0$ is measurable, then

$$\int_{E_n} (f \wedge n) \, \mathrm{d}\mu \nearrow \int f \, \mathrm{d}\mu \,, \tag{4.1}$$

where $a \wedge b = \min\{a, b\}$.

Proof. Set $g_n = (f \wedge n)\mathbb{1}_{E_n}$. Then $0 \le g_n \le f$, each g_n is bounded, and (g_n) is nondecreasing.

Thus the left-hand side of (4.1) has a limit L satisfying

$$L \leq \int f \, \mathrm{d}\mu \, .$$

Conversely, let h be bounded, and suppose that $0 \leq h \leq f$. Choose $M < \infty$ such that $h \leq M$, and let $D = \operatorname{supp}(h)$. For $n \geq M$,

$$\int_{E_n} (f \wedge n) \, \mathrm{d}\mu \geq \int_{E_n} h \, \mathrm{d}\mu = \int h \, \mathrm{d}\mu - \int_{E_n^c} h \, \mathrm{d}\mu \geq \int h \, \mathrm{d}\mu - M\mu(D \cap E_n^c).$$

Since $D \cap E_n^c \searrow \varnothing$ and $\mu(D) < \infty$, continuity from above gives

$$\mu(D \cap E_n^c) \longrightarrow 0.$$

Taking lower limits on both sides gives

$$\liminf_{n\to\infty} \int_{E_n} (f \wedge n) \, \mathrm{d}\mu \geq \int h \, \mathrm{d}\mu$$

for all bounded h satisfying $0 \leq h \leq f$. Thus

$$\liminf_{n\to\infty} \int_{E_n} (f \wedge n) \, \mathrm{d}\mu \geq \sup \left\{ \int h \, \mathrm{d}\mu : 0 \leq h \leq f, \; h \text{ is bounded} \right\} = \int f \, \mathrm{d}\mu \, ,$$

and so

$$\lim_{n\to\infty} \int_{E_n} (f \wedge n) \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu \, .$$

□

This proof technique recurs throughout measure theory: continuity and monotonicity of measures often establish the corresponding properties of integrals. The next theorem collects the basic properties of the nonnegative integral.

Theorem 4.8 Let f and g be nonnegative measurable functions. Then

1. If $f = 0$ almost everywhere, then $\displaystyle\int f \, \mathrm{d}\mu = 0$.

2. For every $a \geq 0$,
$$\int af \, \mathrm{d}\mu = a \int f \, \mathrm{d}\mu \, .$$
3. $$\int (f+g) \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu \, .$$

Proof. Statement (1) follows from the definition. For $a > 0$, multiplication by a bijects the bounded minorants of f with those of af, proving (2); the case $a = 0$ follows from (1).

If $h \leq f$ and $\ell \leq g$ are nonnegative bounded functions, then $h + \ell \leq f + g$. Taking suprema and using (3) gives
$$\int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu \leq \int (f+g) \, \mathrm{d}\mu \, .$$
For the reverse inequality, choose $E_n \nearrow \Omega$ with $\mu(E_n) < \infty$. Since
$$(f+g) \wedge n \leq (f \wedge n) + (g \wedge n),$$
the bounded integral is monotone and additive, so
$$\int_{E_n} ((f+g) \wedge n) \, \mathrm{d}\mu \leq \int_{E_n} (f \wedge n) \, \mathrm{d}\mu + \int_{E_n} (g \wedge n) \, \mathrm{d}\mu \, .$$
Letting $n \to \infty$ and applying Lemma 4.7 to all three terms proves the reverse inequality. □

We are finally ready to define the Lebesgue integral for real-valued functions. The next definition is a natural extension of the nonnegative case.

Definition 4.9 For a measurable real-valued function f, define
$$f^+ = f \vee 0 \qquad \text{and} \qquad f^- = (-f) \vee 0.$$
Thus $f = f^+ - f^-$ and $|f| = f^+ + f^-$. The function f is **integrable** if
$$\int |f| \, \mathrm{d}\mu < \infty.$$

In that case, define
$$\int f \, \mathrm{d}\mu = \int f^+ \, \mathrm{d}\mu - \int f^- \, \mathrm{d}\mu \,.$$

Because $0 \le f^+, f^- \le |f|$, both terms in the last display are finite whenever f is integrable.

Theorem 4.10 If f and g are integrable and $a \in \mathbb{R}$, then

1. af is integrable and
$$\int af \, \mathrm{d}\mu = a \int f \, \mathrm{d}\mu \,.$$
2. $f + g$ is integrable and
$$\int (f+g) \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu \,.$$
3. The absolute value of the integral is bounded by the integral of the absolute value:
$$\left| \int f \, \mathrm{d}\mu \right| \le \int |f| \, \mathrm{d}\mu \,.$$

Proof. Integrability of af and $f+g$ follows from
$$|af| = |a||f| \qquad \text{and} \qquad |f+g| \le |f| + |g|.$$

Statement (1) follows directly from Definition 4.9 and Theorem 4.8, considering $a \ge 0$ and $a < 0$ separately.

Since
$$(f+g)^+ + f^- + g^- = (f+g)^- + f^+ + g^+,$$
additivity for nonnegative functions gives
$$\int (f+g)^+ \, \mathrm{d}\mu + \int f^- \, \mathrm{d}\mu + \int g^- \, \mathrm{d}\mu = \int (f+g)^- \, \mathrm{d}\mu + \int f^+ \, \mathrm{d}\mu + \int g^+ \, \mathrm{d}\mu \,.$$

All terms are finite, so rearrangement proves (2). Finally, $-|f| \le f \le |f|$ and monotonicity yield (3). $\square$

Notation 4.11 When μ is Lebesgue measure on $\mathbb{R}$, write

$$\int f(x)\,\mathrm{d}x$$

in place of $\int f\,\mathrm{d}\mu$. If μ is counting measure on a countable space $\Omega = \{\omega_1, \omega_2, \dots\}$, then

$$\int f\,\mathrm{d}\mu = \sum_{i=1}^{\infty} f(\omega_i)$$

whenever either side is defined.

Lecture 5 — Monday, October 15, 2018

Integral Inequalities and Convergence

The integral construction leads to inequalities that control nonlinear transformations and to convergence theorems that justify passing limits through integrals. These results will be used repeatedly for expectations and limit laws.

Definition 5.1 A function $\varphi\colon \mathbb{R} \to \mathbb{R}$ is **convex** if

$$\varphi(\lambda x + (1-\lambda)y) \le \lambda\varphi(x) + (1-\lambda)\varphi(y)$$

for every $x, y \in \mathbb{R}$ and $\lambda \in (0,1)$.

Theorem 5.2 (Jensen's inequality) Let μ be a probability measure, let f be integrable, and suppose that φ is convex and $\varphi \circ f$ is integrable. Then

$$\varphi\left(\int f\,\mathrm{d}\mu\right) \le \int \varphi(f)\,\mathrm{d}\mu\,.$$

If φ is concave, the inequality is reversed.

Proof. Put $c = \int f \, \mathrm{d}\mu$. Convexity implies that the left and right slopes at c satisfy

$$\lim_{h \searrow 0} \frac{\varphi(c) - \varphi(c-h)}{h} \leq \lim_{h \searrow 0} \frac{\varphi(c+h) - \varphi(c)}{h}.$$

Indeed, for $0 < h_1 < h_2$, the relevant secant slopes satisfy

$$\frac{\varphi(c+h_1) - \varphi(c)}{h_1} \leq \frac{\varphi(c+h_2) - \varphi(c)}{h_2},$$

with the analogous reversed monotonicity on the left. Choose a between the two one-sided slope limits. Then the affine function

$$\ell(x) = a(x-c) + \varphi(c)$$

is a supporting line: $\ell(x) \leq \varphi(x)$ for every $x \in \mathbb{R}$. For example, if $x = c + h$ with $h > 0$, the choice of a gives

$$\frac{\varphi(c+h) - \varphi(c)}{h} \geq a,$$

which is equivalent to $\varphi(x) \geq \ell(x)$; the case $x < c$ is analogous. Therefore

$$\int \varphi(f) \, \mathrm{d}\mu \geq \int \ell(f) \, \mathrm{d}\mu = a \int f \, \mathrm{d}\mu + \varphi(c) - ac = \varphi(c).$$

$\square$

Remark 5.3 A supporting line at a nondifferentiable point need not be unique, as illustrated in Figure 5:

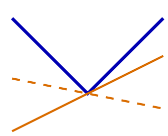

FIGURE 5

If φ is differentiable at the point, the supporting line is the tangent line and is unique.

Exercise 5.4 If μ is a probability measure, $f \geq 0$, and $0 < p < q$, then

$$\left(\int f^p \, \mathrm{d}\mu \right)^{\frac{1}{p}} \leq \left(\int f^q \, \mathrm{d}\mu \right)^{\frac{1}{q}}.$$

Consequently, if X is a random variable and $\mathbb{E}[|X|^q]$ is finite, then $\mathbb{E}[|X|^p]$ is finite for every $0 < p \le q$.

Solution. Apply Jensen's inequality to the convex function $\varphi(x) = x^{q/p}$ and the nonnegative function f^p. Since $\mu(\Omega) = 1$,

$$\left(\int f^p \, \mathrm{d}\mu\right)^{q/p} \le \int f^q \, \mathrm{d}\mu \,.$$

Raising both sides to the power $1/q$ gives the required inequality. The assertion about moments follows by taking $f = |X|$; the case $p = q$ is immediate. □

Lemma 5.5 If $h \ge 0$ and $\displaystyle\int h \, \mathrm{d}\mu = 0$, then $h = 0$ almost everywhere.

Proof. For $\varepsilon > 0$, let $A_\varepsilon = \{\omega : h(\omega) \ge \varepsilon\}$. Then

$$0 = \int h \, \mathrm{d}\mu \ge \int_{A_\varepsilon} h \, \mathrm{d}\mu \ge \varepsilon \mu(A_\varepsilon),$$

so $\mu(A_\varepsilon) = 0$. Since

$$\{h > 0\} = \bigcup_{n=1}^{\infty} A_{1/n},$$

the set on which h is nonzero has measure zero. □

Theorem 5.6 (Hölder's inequality) Let $p, q > 1$ satisfy $1/p + 1/q = 1$. Then

$$\int |fg| \, \mathrm{d}\mu \le \|f\|_p \|g\|_q,$$

where

$$\|f\|_p = \left(\int |f|^p \, \mathrm{d}\mu\right)^{1/p} \qquad \text{and} \qquad \|g\|_q = \left(\int |g|^q \, \mathrm{d}\mu\right)^{1/q}.$$

Proof. If either norm is infinite, the inequality is immediate. If either norm is zero, Lemma 5.5 implies that $f = 0$ almost everywhere or $g = 0$ almost everywhere, and the result follows.

Suppose both norms are finite and positive, and set

$$f^* = \frac{|f|}{\|f\|_p} \qquad \text{and} \qquad g^* = \frac{|g|}{\|g\|_q}.$$

Then

$$\int (f^*)^p \, \mathrm{d}\mu = \int (g^*)^q \, \mathrm{d}\mu = 1.$$

Young's inequality states that, for $x, y \geq 0$,

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}. \tag{5.1}$$

To verify it, fix y and minimize

$$H(x) = \frac{x^p}{p} + \frac{y^q}{q} - xy.$$

Since $H'(x) = x^{p-1} - y$ and $H''(x) = (p-1)x^{p-2} \geq 0$, the minimum occurs at $x = y^{1/(p-1)}$. Using $q = p/(p-1)$, its value is zero.

Apply (5.1) pointwise to f^* and g^* and integrate:

$$\int f^* g^* \, \mathrm{d}\mu \leq \frac{1}{p} \int (f^*)^p \, \mathrm{d}\mu + \frac{1}{q} \int (g^*)^q \, \mathrm{d}\mu = \frac{1}{p} + \frac{1}{q} = 1.$$

Multiplying by $\|f\|_p \|g\|_q$ gives the result. □

Corollary 5.7 (Minkowski's inequality) If $p \geq 1$, then

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Proof. The case $p = 1$ is the integral triangle inequality. Suppose that $p > 1$, and let q be conjugate to p. If either $\|f\|_p$ or $\|g\|_p$ is infinite, the result is immediate. Otherwise,

$$|f + g|^p \leq 2^{p-1}(|f|^p + |g|^p)$$

shows that $\|f + g\|_p$ is finite. If this norm is zero, there is nothing to prove. Hölder's inequality now gives

$$\|f + g\|_p^p = \int |f + g|^p \, \mathrm{d}\mu \leq \int (|f| + |g|)|f + g|^{p-1} \, \mathrm{d}\mu \leq (\|f\|_p + \|g\|_p) \left\| |f + g|^{p-1} \right\|_q.$$

Since $(p-1)q = p$,
$$\left\| |f+g|^{p-1} \right\|_q = \|f+g\|_p^{p-1}.$$
Dividing by this positive quantity proves the claim. $\square$

Remark 5.8 The functional
$$\|f\|_p = \left(\int |f|^p \, \mathrm{d}\mu \right)^{1/p}$$
defines a norm on equivalence classes modulo equality almost everywhere precisely when $p \geq 1$.

Definition 5.9 A sequence of measurable functions (f_n) **converges in measure** to a measurable function f if, for every $\varepsilon > 0$,
$$\mu\big(\{|f_n - f| > \varepsilon\}\big) \longrightarrow 0.$$
On a probability space, convergence in measure is also called **convergence in probability**.

Theorem 5.10 (Bounded convergence theorem) Let $E \in \mathcal{F}$ satisfy $\mu(E) < \infty$. Suppose that each f_n is supported on E, that $|f_n| \leq M$ for a common finite constant M, and that $f_n \to f$ in measure. Then f is integrable and
$$\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu \,.$$

Proof. Convergence in measure implies that $|f| \leq M$ almost everywhere and that $f = 0$ almost everywhere on E^c. Indeed, either failure would produce a set of positive measure on which $|f_n - f|$ remains bounded away from zero. Thus f is integrable.

Fix $\varepsilon > 0$ and set
$$A_n^\varepsilon = \{\omega : |f_n(\omega) - f(\omega)| > \varepsilon\}.$$
Then
$$\begin{aligned} \left| \int f_n \, \mathrm{d}\mu - \int f \, \mathrm{d}\mu \right| &\leq \int_E |f_n - f| \, \mathrm{d}\mu \\ &= \int_{A_n^\varepsilon \cap E} |f_n - f| \, \mathrm{d}\mu + \int_{(A_n^\varepsilon)^c \cap E} |f_n - f| \, \mathrm{d}\mu \end{aligned}$$

$$\leq 2M\mu(A_n^\varepsilon) + \varepsilon\mu(E).$$

Since $f_n \to f$ in measure, the first term tends to zero. Hence

$$\limsup_{n\to\infty} \left| \int f_n \,\mathrm{d}\mu - \int f \,\mathrm{d}\mu \right| \leq \varepsilon\mu(E).$$

Letting $\varepsilon \downarrow 0$ proves the result. □

Remark 5.11 The finite measure assumption in Theorem 5.10 is essential. For Lebesgue measure on $\mathbb{R}$, let

$$f_n = \frac{1}{n}\mathbb{1}_{(0,n)}.$$

Then $f_n \to 0$ pointwise and $|f_n| \leq 1$, but $\int f_n \,\mathrm{d}\mu = 1$ for every n.

Theorem 5.12 (Fatou's lemma) If $f_n \geq 0$ for every n, then

$$\int \liminf_{n\to\infty} f_n \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu\,.$$

Proof. Let

$$g_n = \inf_{k\geq n} f_k \qquad \text{and} \qquad g = \liminf_{n\to\infty} f_n.$$

Then $0 \leq g_n \nearrow g$ and $g_n \leq f_k$ whenever $k \geq n$. Choose $E_m \nearrow \Omega$ with $\mu(E_m) < \infty$. For fixed m, the functions $(g_n \wedge m)\mathbb{1}_{E_m}$ increase pointwise to $(g \wedge m)\mathbb{1}_{E_m}$ and are bounded by $m\mathbb{1}_{E_m}$. For every $\varepsilon > 0$, the sets

$$\{(g \wedge m) - (g_n \wedge m) > \varepsilon\} \cap E_m$$

decrease to the empty set. Since $\mu(E_m) < \infty$, continuity from above shows that their measures tend to zero. Thus the displayed functions converge in measure, and Theorem 5.10 gives

$$\int_{E_m} (g_n \wedge m) \,\mathrm{d}\mu \longrightarrow \int_{E_m} (g \wedge m) \,\mathrm{d}\mu\,.$$

For $k \geq n$,

$$\int f_k \,\mathrm{d}\mu \geq \int g_n \,\mathrm{d}\mu \geq \int_{E_m} (g_n \wedge m) \,\mathrm{d}\mu\,.$$

First let $k \to \infty$ through the lower limit, then let $n \to \infty$, to obtain

$$\liminf_{k\to\infty} \int f_k \,\mathrm{d}\mu \geq \int_{E_m} (g \wedge m) \,\mathrm{d}\mu\,.$$

Finally let $m \to \infty$ and apply Lemma 4.7. $\square$

Remark 5.13 Some uniform lower bound is necessary in Theorem 5.12. For Lebesgue measure, let

$$f_n = -n\mathbb{1}_{(0,1/n]}.$$

Then $f_n \to 0$ almost everywhere and $\int f_n \,\mathrm{d}\mu = -1$ for every n, so Fatou's inequality fails without a hypothesis that supplies a lower bound.

Corollary 5.14 (Monotone convergence theorem) If $0 \le f_n \nearrow f$, then

$$\int f \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu\,.$$

Proof. Since $f_n \le f$, monotonicity gives

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu \le \int f \,\mathrm{d}\mu\,.$$

Conversely, Theorem 5.12 yields

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu \ge \int \liminf_{n\to\infty} f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu\,.$$

$\square$

Corollary 5.15 (Dominated convergence theorem) Suppose that $f_n \to f$ almost everywhere and that $|f_n| \le g$ almost everywhere for an integrable function g. Then f is integrable and

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu\,.$$

Proof. Pointwise convergence gives $|f| \le g$ almost everywhere, so f is integrable. Since $f_n + g \ge 0$, Fatou's lemma gives

$$\liminf_{n\to\infty} \int (f_n + g) \,\mathrm{d}\mu \ge \int (f + g) \,\mathrm{d}\mu\,,$$

and hence

$$\liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu \ge \int f \,\mathrm{d}\mu\,.$$

Applying Fatou's lemma to $g - f_n$ similarly yields

$$\liminf_{n\to\infty} \int (g - f_n)\,\mathrm{d}\mu \geq \int (g - f)\,\mathrm{d}\mu\,,$$

which is equivalent to

$$\limsup_{n\to\infty} \int f_n\,\mathrm{d}\mu \leq \int f\,\mathrm{d}\mu\,.$$

The lower and upper limits therefore agree with $\int f\,\mathrm{d}\mu$. □

Corollary 5.16 (Fatou's lemma with a constant lower bound) Suppose that $\mu(\Omega) < \infty$ and $f_n \geq C$ almost everywhere for a finite constant C. Then

$$\int \liminf_{n\to\infty} f_n\,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int f_n\,\mathrm{d}\mu\,.$$

Proof. Apply Theorem 5.12 to $g_n = f_n - C \geq 0$. Since $\mu(\Omega) < \infty$, the finite quantity $C\mu(\Omega)$ may be subtracted from both sides. □

Remark 5.17 The finite-measure assumption in Corollary 5.16 cannot be omitted. For Lebesgue measure on $\mathbb{R}$, let

$$f_n = -\frac{1}{n}\mathbb{1}_{(0,n)}.$$

Then $f_n \to 0$ almost everywhere, but $\int f_n\,\mathrm{d}\mu = -1$ for every n.

Lecture 6 — Monday, October 22, 2018

2. Expectation, Product Measures, and Independence

Expectation

Expectation is the Lebesgue integral of a random variable with respect to its underlying probability measure. The preceding integral results therefore translate directly into algebraic, order, and convergence properties of expectation.

Definition 6.1 Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. If $X \geq 0$, its **expectation** is

$$\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P} \in [0, \infty].$$

For a general real-valued X, define

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

whenever at least one of the two terms is finite. If both terms are infinite, the expectation is undefined. The random variable is integrable precisely when $\mathbb{E}[|X|] < \infty$.

Theorem 6.2 Whenever the expressions involved are defined, the following statements hold:

1. If X and Y are nonnegative, or if both are integrable, then

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \qquad \text{and} \qquad \mathbb{E}[aX+b] = a\mathbb{E}[X] + b.$$

2. If $X \geq Y$ almost surely, then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.
3. If X and $\varphi(X)$ are integrable and φ is convex, then we have Jensen's inequality

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

4. If $p, q \in [1, \infty]$ satisfy $1/p + 1/q = 1$, then we have Hölder's inequality

$$\mathbb{E}[|XY|] \leq \|X\|_p \|Y\|_q,$$

   where, for finite p,

$$\|X\|_p = \big(\mathbb{E}[|X|^p]\big)^{1/p},$$

   and

$$\|X\|_\infty = \inf\{M \geq 0 : \mathbb{P}(|X| > M) = 0\}.$$

Proof. Statements (1) and (2) are Theorems 4.8 and 4.10 applied to the probability measure $\mathbb{P}$. Statement (3) is Theorem 5.2, and (4) follows from Theorem 5.6; the endpoint cases use the defining essential supremum bound. □

Theorem 6.3 (Generalized Chebyshev's inequality) Let $\varphi\colon \mathbb{R} \to [0, \infty]$ be measurable, let A be Borel measurable, and define

$$i_A = \inf\{\varphi(y) : y \in A\}.$$

Then

$$i_A \mathbb{P}(X \in A) \leq \mathbb{E}\big[\varphi(X) \mathbb{1}_{\{X \in A\}}\big] \leq \mathbb{E}[\varphi(X)].$$

Proof. Pointwise,

$$i_A \mathbb{1}_{\{X \in A\}} \leq \varphi(X) \mathbb{1}_{\{X \in A\}} \leq \varphi(X).$$

Take expectations and apply monotonicity. □

We obtain the classical Chebyshev inequality as a special case of Theorem 6.3:

Corollary 6.4 (Chebyshev's inequality) If $\mathbb{E}[X^2] < \infty$, then for every $x > 0$,

$$\mathbb{P}(|X| > x) \leq \frac{\mathbb{E}[X^2]}{x^2}.$$

Proof. Apply Theorem 6.3 with $\varphi(y) = y^2$ and $A = (-\infty, -x) \cup (x, \infty)$. □

Lecture 7 — Monday, October 29, 2018

Theorem 7.1 (Dominated convergence for expectations) If $X_n \to X$ almost surely and $|X_n| \leq Y$ almost surely, where $\mathbb{E}[Y] < \infty$, then

$$\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$

Proof. This is Corollary 5.15 applied to the probability measure $\mathbb{P}$. □

Remark 7.2 If Y is a finite constant, Theorem 7.1 is the bounded convergence theorem for expectations. The monotone convergence theorem and Fatou's lemma likewise transfer directly from integrals to expectations.

Theorem 7.3 Suppose that $X_n \to X$ almost surely and that $g, h\colon \mathbb{R} \to \mathbb{R}$ are continuous functions satisfying

1. The function g is nonnegative and coercive: $g(x) \to \infty$ as $|x| \to \infty$.
2. The function h grows strictly more slowly than g: $|h(x)|/g(x) \to 0$ as $|x| \to \infty$.
3. The g-moments are uniformly bounded: $\mathbb{E}[g(X_n)] \le K < \infty$ for every n.

Then

$$\mathbb{E}[h(X_n)] \longrightarrow \mathbb{E}[h(X)].$$

Proof. Fix a sufficiently large $M > 0$ such that $g(x) > 0$ for $|x| > M$ and $\mathbb{P}(|X| = M) = 0$. Such arbitrarily large choices are possible because g is coercive and a distribution function has only countably many discontinuities. Define

$$\bar{X}_n = X_n \mathbb{1}_{\{|X_n| \le M\}} \qquad \text{and} \qquad \bar{X} = X \mathbb{1}_{\{|X| \le M\}}.$$

Then $|\bar{X}_n|, |\bar{X}| \le M$ and $\bar{X}_n \to \bar{X}$ almost surely. Since h is continuous, $h(\bar{X}_n) \to h(\bar{X})$ almost surely and these random variables are bounded by a finite constant. The dominated convergence theorem gives

$$\mathbb{E}[h(\bar{X}_n)] \longrightarrow \mathbb{E}[h(\bar{X})].$$

Subtracting the constant $h(0)$ from h does not affect the conclusion, so assume $h(0) = 0$. The effect of truncating X_n is then bounded by

$$\begin{aligned} \left|\mathbb{E}[h(X_n)] - \mathbb{E}[h(\bar{X}_n)]\right| &\le \mathbb{E}\left[|h(X_n)| \mathbb{1}_{\{|X_n| > M\}}\right] \\ &\le K \sup_{|x| > M} \frac{|h(x)|}{g(x)}. \end{aligned}$$

Meanwhile, Fatou's lemma and (3) give $\mathbb{E}[g(X)] \le K$. Set

$$\varepsilon_M = \sup_{|x| > M} \frac{|h(x)|}{g(x)},$$

which tends to zero as $M \to \infty$. The same estimate for X gives

$$\left|\mathbb{E}[h(X)] - \mathbb{E}[h(\bar{X})]\right| \leq K\varepsilon_M.$$

The triangle inequality and the convergence of the truncated expectations now yield

$$\limsup_{n\to\infty} |\mathbb{E}[h(X_n)] - \mathbb{E}[h(X)]| \leq 2K\varepsilon_M.$$

Arbitrarily large choices of M satisfy the required continuity condition, and $\varepsilon_M \to 0$ as $M \to \infty$, so the result follows. □

Example 7.4 Take $h(x) = x$ and $g(x) = x^2$ in Theorem 7.3. If $\mathbb{E}[X_n^2] \leq K$ for every n, then

$$\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$

Thus a uniform second moment bound is enough to interchange this limit and expectation, even when there is no single dominating random variable.

How do we compute expectations in practice? We know that $\mathbb{E}[X] = \int_\Omega X \,\mathrm{d}\mathbb{P}$, but we usually do not know Ω or $\mathbb{P}$ explicitly. Instead, we often know the distribution of X, which is the pushforward measure $\mu = \mathbb{P} \circ X^{-1}$ on $(\mathbb{R}^d, \mathcal{R}^d)$. The following theorem allows us to compute expectations using only the distribution of X.

Theorem 7.5 (Law of the unconscious statistician) Let X be a random element of $(\mathbb{R}^d, \mathcal{R}^d)$ with distribution $\mu = \mathbb{P} \circ X^{-1}$. If $f\colon \mathbb{R}^d \to \mathbb{R}$ is measurable and either $f \geq 0$ or f is integrable with respect to μ, then

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^d} f(y)\mu(\mathrm{d}y).$$

Equivalently, for a measurable map $h\colon \Omega \to \mathbb{R}^d$,

$$\int_\Omega f(h(\omega))\mathbb{P}(\mathrm{d}\omega) = \int_{\mathbb{R}^d} f(y)(\mathbb{P} \circ h^{-1})(\mathrm{d}y).$$

Proof. If $f = \mathbb{1}_A$ for a Borel set A, then

$$\mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A) = \mu(A) = \int \mathbb{1}_A \,\mathrm{d}\mu\,.$$

Linearity proves the identity for nonnegative simple functions. If $f \geq 0$, choose simple functions $f_n \nearrow f$ and apply the monotone convergence theorem on both sides. Finally, apply the nonnegative result to f^+ and f^- and subtract whenever f is integrable. $\square$

Lecture 8 — Monday, November 12, 2018

Dynkin's λ–π Theorem

The λ–π theorem extends identities verified on a convenient generating class to the full generated σ-algebra.

Definition 8.1 A class $\mathcal{P}$ of subsets of Ω is a ***π*-system** if it is closed under finite intersections.

Example 8.2 The class $\{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system, although it is neither an algebra nor a σ-algebra.

Definition 8.3 A class $\mathcal{L}$ of subsets of Ω is a ***λ*-system** if the following hold:

λ1. The whole space belongs to the class: $\Omega \in \mathcal{L}$.

λ2. The class is closed under complements: $A \in \mathcal{L}$ implies $A^c \in \mathcal{L}$.

λ3. The class is closed under countable disjoint unions: if $A_1, A_2, \dots \in \mathcal{L}$ are pairwise disjoint, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$.

Example 8.4 Let $\Omega = \{1, 2, 3, 4\}$ and

$$\mathcal{L} = \{\varnothing, \Omega, \{1, 2\}, \{3, 4\}, \{1, 3\}, \{2, 4\}\}.$$

This class is a λ-system: its nontrivial members occur in complementary pairs, and the only disjoint unions that must be checked are unions of complementary pairs. It is not a σ-algebra because

$$\{1, 2\} \cap \{1, 3\} = \{1\} \notin \mathcal{L}.$$

Proposition 8.5 The definition of a λ-system is equivalent if we replace ($\lambda 2$) with the following:

$\lambda 2'$. The class is closed under nested differences: if $A, B \in \mathcal{L}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{L}$.

Proof. Assume ($\lambda 2$) and ($\lambda 3$). If $A \subseteq B$ and $A, B \in \mathcal{L}$, then $A \cup B^c$ is a disjoint union of two members of $\mathcal{L}$, and

$$B \setminus A = (A \cup B^c)^c \in \mathcal{L}.$$

Conversely, ($\lambda 1$) and ($\lambda 2'$) give $A^c = \Omega \setminus A \in \mathcal{L}$. □

Proposition 8.6 The definition of a λ-system is equivalent if we replace ($\lambda 3$) with the following:

$\lambda 3'$. In the presence of ($\lambda 2'$), closure under countable disjoint unions (i.e., ($\lambda 3$)) is equivalent to closure under increasing unions: if $A_n \in \mathcal{L}$ and $A_n \nearrow A$, then $A \in \mathcal{L}$.

Proof. Suppose that $A_n \nearrow A$. Define $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n \geq 2$. By ($\lambda 2'$), the sets B_n belong to $\mathcal{L}$; they are pairwise disjoint and have union A. Thus ($\lambda 3$) implies ($\lambda 3'$). Conversely, the partial unions of a disjoint sequence increase to its full union, so ($\lambda 3'$) implies ($\lambda 3$). □

Lemma 8.7 A class that is both a λ-system and a π-system is a σ-algebra.

Proof. Closure under complements is part of the definition of a λ-system. Given $A_1, A_2, \ldots$ in the class, define

$$B_1 = A_1 \qquad \text{and} \qquad B_n = A_n \cap \left(\bigcup_{k=1}^{n-1} A_k \right)^c \quad (n \geq 2).$$

Finite intersections and complements show that every B_n belongs to the class. The sets B_n are pairwise disjoint and

$$\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n.$$

Closure under disjoint countable unions proves the result. □

Theorem 8.8 (Dynkin's λ–π theorem) Let $\mathcal{P}$ be a π-system and let $\mathcal{L}$ be a λ-system. If

$$\mathcal{P} \subseteq \mathcal{L},$$

then

$$\sigma(\mathcal{P}) \subseteq \mathcal{L}.$$

Proof. Let $\mathcal{L}_0$ be the λ-system generated by $\mathcal{P}$. Then $\mathcal{L}_0$ is itself a λ-system. Moreover, since $\mathcal{L}$ is a λ-system containing $\mathcal{P}$, we have $\mathcal{L}_0 \subseteq \mathcal{L}$. Thus it suffices to show that $\mathcal{L}_0$ is a σ-algebra, for then

$$\sigma(\mathcal{P}) \subseteq \mathcal{L}_0 \subseteq \mathcal{L}.$$

By Lemma 8.7, it is enough to prove that $\mathcal{L}_0$ is a π-system. For $A \in \mathcal{L}_0$, define

$$\mathcal{L}_A := \{B \subseteq \Omega : A \cap B \in \mathcal{L}_0\}.$$

We first show that $\mathcal{L}_A$ is a λ-system.

Since $A \in \mathcal{L}_0$,

$$A \cap \Omega = A \in \mathcal{L}_0,$$

and hence $\Omega \in \mathcal{L}_A$ and so ($\lambda 1$) holds. Next, suppose that $B_1, B_2 \in \mathcal{L}_A$ with $B_1 \subseteq B_2$. Then

$$A \cap B_1,\ A \cap B_2 \in \mathcal{L}_0 \qquad \text{and} \qquad A \cap B_1 \subseteq A \cap B_2.$$

Since $\mathcal{L}_0$ is a λ-system,

$$(A \cap B_2) \setminus (A \cap B_1) \in \mathcal{L}_0.$$

But

$$(A \cap B_2) \setminus (A \cap B_1) = A \cap (B_2 \setminus B_1),$$

so $B_2 \setminus B_1 \in \mathcal{L}_A$ and hence ($\lambda 2'$) holds. Finally, suppose that $B_1, B_2, \ldots$ are pairwise disjoint members of $\mathcal{L}_A$. Then $A \cap B_k \in \mathcal{L}_0$ for every $k \in \mathbb{N}$, and the sets $A \cap B_k$ are pairwise disjoint. Therefore

$$A \cap \bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} (A \cap B_k) \in \mathcal{L}_0,$$

and hence

$$\bigcup_{k=1}^{\infty} B_k \in \mathcal{L}_A,$$

so ($\lambda 3$) holds. Thus $\mathcal{L}_A$ is a λ-system.

We now use this observation twice. Fix $A \in \mathcal{P}$. If $B \in \mathcal{P}$, then, since $\mathcal{P}$ is a π-system,

$$A \cap B \in \mathcal{P} \subseteq \mathcal{L}_0,$$

and hence $B \in \mathcal{L}_A$. Thus $\mathcal{P} \subseteq \mathcal{L}_A$. Since $\mathcal{L}_A$ is a λ-system and $\mathcal{L}_0$ is the smallest λ-system containing $\mathcal{P}$, it follows that

$$\mathcal{L}_0 \subseteq \mathcal{L}_A.$$

Consequently,

$$A \cap B \in \mathcal{L}_0 \quad \text{for every } A \in \mathcal{P},\ B \in \mathcal{L}_0.$$

Now fix $B \in \mathcal{L}_0$. By what we have just proved, for every $A \in \mathcal{P}$,

$$B \cap A = A \cap B \in \mathcal{L}_0,$$

and hence $A \in \mathcal{L}_B$. Thus $\mathcal{P} \subseteq \mathcal{L}_B$. Again, $\mathcal{L}_B$ is a λ-system, so the minimality of $\mathcal{L}_0$ gives $\mathcal{L}_0 \subseteq \mathcal{L}_B$. Therefore, for every $A, B \in \mathcal{L}_0$,

$$A \cap B \in \mathcal{L}_0.$$

Hence $\mathcal{L}_0$ is a π-system. By Lemma 8.7, $\mathcal{L}_0$ is therefore a σ-algebra, completing the proof. □

The Dynkin λ–π theorem is a powerful tool for proving that two measures agree on a σ-algebra. If $\mathcal{L}$ is the class of sets with some property, then it is often easy to show that $\mathcal{L}$ is a λ-system. If $\mathcal{P}$ is a π-system with $\mathcal{P} \subset \mathcal{L}$, then the λ–π theorem gives $\sigma(\mathcal{P}) \subset \mathcal{L}$. If $\sigma(\mathcal{P})$ is the σ-algebra of interest, then the property holds for all sets in that σ-algebra.

Product Measures and Fubini's Theorem

Product measures provide the natural measurable structure for several variables considered jointly. The Tonelli–Fubini theorem then reduces integrals on a product space to iterated integrals.

Definition 8.9 For σ-finite measure spaces $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$, the **product σ-algebra** is

$$\mathcal{A} \otimes \mathcal{B} = \sigma\{A \times B : A \in \mathcal{A},\ B \in \mathcal{B}\}.$$

The **product measure** $\mu_1 \times \mu_2$ is the unique measure on this σ-algebra satisfying

$$(\mu_1 \times \mu_2)(A \times B) = \mu_1(A)\mu_2(B).$$

Here and below, the product is interpreted using the convention $0 \cdot \infty = \infty \cdot 0 = 0$.

Theorem 8.10 (Tonelli–Fubini theorem) Let $f\colon X \times Y \to \overline{\mathbb{R}}$ be $\mathcal{A} \otimes \mathcal{B}$-measurable. If either $f \geq 0$ or

$$\int_{X\times Y} |f| \,\mathrm{d}(\mu_1 \times \mu_2) < \infty,$$

then the sections are measurable. If $f \geq 0$, both iterated integrals are defined as elements of $[0, \infty]$ and

$$\int_X \left(\int_Y f(x, y)\mu_2(\mathrm{d}y) \right) \mu_1(\mathrm{d}x) = \int_{X\times Y} f \,\mathrm{d}(\mu_1 \times \mu_2) = \int_Y \left(\int_X f(x, y)\mu_1(\mathrm{d}x) \right) \mu_2(\mathrm{d}y).$$

If f is integrable, almost every section is integrable, and the same equality holds after the exceptional inner integrals are defined arbitrarily.

Proof. The argument has three components:

1. For each fixed x, the map $y \mapsto f(x, y)$ is $\mathcal{B}$-measurable, and symmetrically for fixed y.
2. The map $x \mapsto \int_Y f(x, y)\mu_2(\mathrm{d}y)$ is $\mathcal{A}$-measurable.
3. The iterated and product-space integrals agree.

Begin with an indicator $f = \mathbb{1}_E$, where $E \in \mathcal{A} \otimes \mathcal{B}$, and write

$$E_x = \{y \in Y : (x, y) \in E\}.$$

For fixed x, the class

$$\mathcal{L}_x = \{E : E_x \in \mathcal{B}\}$$

is a λ-system containing every measurable rectangle. By Theorem 8.8, it contains $\mathcal{A} \otimes \mathcal{B}$, which proves (1).

First suppose that $\mu_1(X)$ and $\mu_2(Y)$ are finite. Let $\mathcal{D}$ be the class of measurable sets $E \subseteq X \times Y$ for which $x \mapsto \mu_2(E_x)$ is measurable and

$$\int_X \mu_2(E_x)\mu_1(\mathrm{d}x) = (\mu_1 \times \mu_2)(E).$$

This class contains $X \times Y$. If $E \in \mathcal{D}$, then

$$\mu_2((E^c)_x) = \mu_2(Y) - \mu_2(E_x),$$

so measurability and the integral identity also hold for E^c. If (E_n) are pairwise disjoint members of $\mathcal{D}$, then

$$\mu_2\left(\left(\bigcup_{n=1}^{\infty} E_n\right)_x\right) = \sum_{n=1}^{\infty} \mu_2((E_n)_x),$$

and the monotone convergence theorem proves both assertions for their union. Thus $\mathcal{D}$ is a λ-system. For a rectangle $E = A \times B$,

$$\mu_2(E_x) = \mu_2(B)\mathbb{1}_A(x)$$

and

$$\int_X \mu_2(E_x)\mu_1(\mathrm{d}x) = \mu_1(A)\mu_2(B) = (\mu_1 \times \mu_2)(E).$$

The measurable rectangles form a generating π-system, so Theorem 8.8 gives $\mathcal{D} = \mathcal{A} \otimes \mathcal{B}$.

For the σ-finite case, choose disjoint measurable partitions

$$X = \biguplus_{i=1}^{\infty} P_i \qquad \text{and} \qquad Y = \biguplus_{j=1}^{\infty} Q_j,$$

such that $\mu_1(P_i) < \infty$ and $\mu_2(Q_j) < \infty$. Apply the result for finite measures to each set $E \cap (P_i \times Q_j)$ and sum over i and j. The monotone convergence theorem then proves (2) and (3) for the indicator of every measurable E.

Linearity extends the result to nonnegative simple functions. For nonnegative measurable f, choose simple functions $f_n \nearrow f$ and use the monotone convergence theorem for the inner, outer, and product space integrals. This is Tonelli's theorem. Finally, if f is integrable, apply the nonnegative result to f^+ and f^-; both iterated absolute integrals are finite, so subtraction gives Fubini's theorem. $\square$

Independence

Independence asserts that joint probabilities factor into products of marginal probabilities. The λ–π theorem makes it sufficient to verify this factorization on convenient generating classes.

Definition 8.11

1. Events A and B are **independent** if
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$
2. Random variables X and Y are **independent** if, for all Borel sets $C, D \subseteq \mathbb{R}$,
$$\mathbb{P}(X \in C,\ Y \in D) = \mathbb{P}(X \in C)\mathbb{P}(Y \in D).$$
3. The σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ are **independent** if every $A \in \mathcal{F}_1$ is independent of every $B \in \mathcal{F}_2$.
4. The collections $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are **independent** if, for every nonempty $I \subseteq \{1, \ldots, n\}$ and every choice $A_i \in \mathcal{A}_i$,
$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Remark 8.12 Definition (1) is recovered from (2) by taking $X = \mathbb{1}_A$ and $Y = \mathbb{1}_B$. Moreover, X and Y are independent precisely when $\sigma(X)$ and $\sigma(Y)$ are independent. This equivalence makes Theorem 8.8 especially useful.

Theorem 8.13 If $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent π-systems, then $\sigma(\mathcal{A}_1), \ldots, \sigma(\mathcal{A}_n)$ are independent.

Proof. First adjoin Ω to each $\mathcal{A}_i$. The enlarged collections are still π-systems and remain independent: choosing Ω from an unused coordinate reduces the full intersection identity to the identity for the corresponding subcollection. It is now enough to establish the factorization for one set from every collection and to enlarge one system at a time.

Fix $A_i \in \mathcal{A}_i$ for $i = 2, \ldots, n$, put $B = \bigcap_{i=2}^n A_i$, and define
$$\mathcal{L}_B = \{A : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\}.$$
The class $\mathcal{L}_B$ is a λ-system. It contains Ω; if $A_1 \subseteq A_2$ are in $\mathcal{L}_B$, subtraction gives
$$\mathbb{P}((A_2 \setminus A_1) \cap B) = \mathbb{P}(A_2 \setminus A_1)\mathbb{P}(B).$$

Closure under increasing unions follows from continuity from below. Independence of the original π-systems gives $\mathcal{A}_1 \subseteq \mathcal{L}_B$. Hence Theorem 8.8 implies $\sigma(\mathcal{A}_1) \subseteq \mathcal{L}_B$.

This proves the full intersection identity for $\sigma(\mathcal{A}_1), \mathcal{A}_2, \ldots, \mathcal{A}_n$. Since every collection contains Ω, it also proves independence in the sense of all subcollections. Repeating the argument for each remaining coordinate proves the theorem. $\square$

Corollary 8.14 The random variables $X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \le x_i)$$

for every $x_1, \ldots, x_n \in \mathbb{R}$.

Proof. *Forward implication.* If the random variables are independent, apply the definition of independence to the Borel sets $(-\infty, x_i]$.

Reverse implication. For each i, let

$$\mathcal{A}_i = \{\{X_i \le x\} : x \in \mathbb{R}\}.$$

This is a π-system because the intersection of the events indexed by x and y is indexed by $\min\{x, y\}$. Moreover, $\sigma(\mathcal{A}_i) = \sigma(X_i)$. The assumed factorization says that the $\mathcal{A}_i$ are independent, so Theorem 8.13 completes the proof. $\square$

Theorem 8.15 Let $X_1, \ldots, X_n$ be independent, and let X_i have distribution μ_i. Then $(X_1, \ldots, X_n)$ has distribution $\mu_1 \times \cdots \times \mu_n$.

Proof. Let ν be the distribution of $(X_1, \ldots, X_n)$ and let $\mu = \mu_1 \times \cdots \times \mu_n$. Define

$$\mathcal{L} = \{E : \nu(E) = \mu(E)\}.$$

Since ν and μ are probability measures, $\mathcal{L}$ is a λ-system. For a measurable rectangle $A_1 \times \cdots \times A_n$, independence gives

$$\nu(A_1 \times \cdots \times A_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) = \prod_{i=1}^{n} \mu_i(A_i) = \mu(A_1 \times \cdots \times A_n).$$

The measurable rectangles form a π-system generating the product σ-algebra, so Theorem 8.8 implies that $\mathcal{L}$ contains the entire product σ-algebra. $\square$

Lecture 9 — Monday, November 19, 2018

Theorem 9.1 (Convolution formula) Suppose that X and Y are independent, with cumulative distribution functions F and G, respectively. Write $F(\mathrm{d}x)$ and $G(\mathrm{d}y)$ for their associated Lebesgue–Stieltjes probability measures. Then

$$\mathbb{P}(X+Y \le z) = \int_{\mathbb{R}} F(z-y)G(\mathrm{d}y).$$

Proof.

$$\begin{aligned}\mathbb{P}(X+Y \le z) &= \int_{\mathbb{R}^2} \mathbb{1}_{\{x+y\le z\}}(F\times G)(\mathrm{d}x\,,\mathrm{d}y)\\ &= \int_{\mathbb{R}}\left(\int_{\mathbb{R}} \mathbb{1}_{\{x\le z-y\}}F(\mathrm{d}x)\right)G(\mathrm{d}y)\\ &= \int_{\mathbb{R}} F(z-y)G(\mathrm{d}y),\end{aligned}$$

where Theorems 8.10 and 8.15 justify the first two equalities. $\square$

Repeated convolution determines the distribution of any finite sum of independent random variables. Infinite sums require separate convergence arguments.

Theorem 9.2 If X and Y are independent, X has density f, and Y has cumulative distribution function G, then $X+Y$ has density

$$h(x) = \int_{\mathbb{R}} f(x-y)G(\mathrm{d}y).$$

If Y also has density g, then

$$h(x) = \int_{\mathbb{R}} f(x-y)g(y)\,\mathrm{d}y\,.$$

Proof. By Theorem 9.1 and the representation of F through f,

$$\begin{aligned}\mathbb{P}(X+Y \le z) &= \int_{\mathbb{R}} \int_{-\infty}^{z-y} f(u)\,\mathrm{d}u\; G(\mathrm{d}y) \\ &= \int_{-\infty}^{z} \int_{\mathbb{R}} f(x-y)G(\mathrm{d}y)\,\mathrm{d}x \\ &= \int_{-\infty}^{z} h(x)\,\mathrm{d}x\,,\end{aligned}$$

where the Tonelli theorem permits the change in order. If $G(\mathrm{d}y) = g(y)\,\mathrm{d}y$, substitution gives the second formula. □

For example, convolution shows directly that sums of independent Gamma random variables with a common rate are Gamma distributed and that sums of independent Normal random variables are Normal. Note that product spaces construct *finite* collections of random variables; the Kolmogorov extension theorem supplies consistent countable collections and underlies constructions such as Brownian motion, but we will not cover it here.

3. Weak Laws and Convergence in Probability

Convergence in Probability and Finite Variance Laws

Weak laws of large numbers describe concentration of sample averages in probability. The finite variance argument rests on variance additivity and Chebyshev's inequality, while the later truncation argument removes the variance assumption.

Definition 9.3 A sequence $(X_n)_{n\ge 1}$ **converges in probability** to X, written $X_n \xrightarrow{\mathbb{P}} X$, if, for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| \ge \varepsilon) \longrightarrow 0.$$

In Figure 6, the error distributions are drawn schematically to emphasize the defining feature: for each fixed ε, the probability outside $[-\varepsilon, \varepsilon]$ vanishes. No assumption of normality is intended.

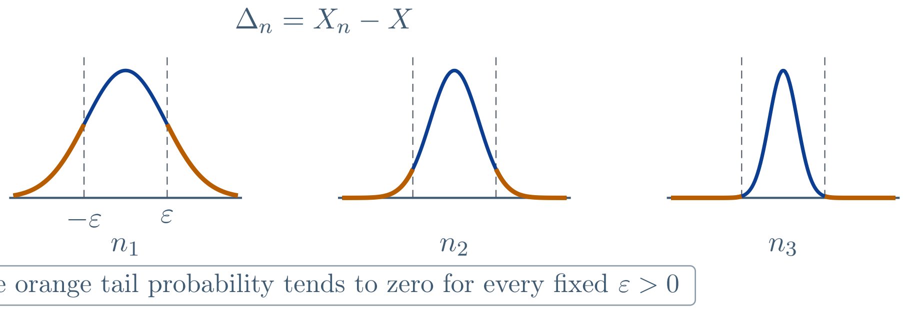

FIGURE 6

Definition 9.4 Integrable random variables X and Y with an integrable product are **uncorrelated** if

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

Theorem 9.5 (Variance of an uncorrelated sum) If $X_1, \dots, X_n$ are pairwise uncorrelated and have finite variances, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Proof. Expanding the centered square gives

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} \mathrm{Var}(X_i),$$

because every covariance in the second sum is zero. □

Lemma 9.6 If $\mathbb{E}[|X_n|^p] \to 0$ for some $p > 0$, then $X_n \xrightarrow{\mathbb{P}} 0$.

Proof. Chebyshev's inequality gives, for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n| \ge \varepsilon) = \mathbb{P}(|X_n|^p \ge \varepsilon^p) \le \frac{\mathbb{E}[|X_n|^p]}{\varepsilon^p} \longrightarrow 0.$$

□

Theorem 9.7 (Weak law for uncorrelated variables) Let $X_1, X_2, \dots$ be pairwise uncorrelated random variables such that $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) \le C < \infty$ for every i. If

$$S_n = \sum_{i=1}^{n} X_i,$$

then $S_n/n \xrightarrow{\mathbb{P}} \mu$.

Proof.

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

Now apply Lemma 9.6 with $p = 2$. □

Next, we have a very elegant application of the weak law to approximation theory. The Bernstein polynomials are a classical tool for approximating continuous functions on $[0,1]$ by polynomials.

Example 9.8 (Bernstein polynomial approximation) Let f be continuous on $[0,1]$, and define

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

Then $f_n \to f$ uniformly on $[0,1]$.

Proof. Fix $p \in [0,1]$. Let $X_1, X_2, \dots$ be independent Bernoulli random variables with success probability p (allowing the degenerate endpoint cases), and set

$$S_n = \sum_{i=1}^{n} X_i.$$

Since $S_n \sim \mathrm{Bin}(n,p)$, we have $\mathbb{P}(S_n = i) = \binom{n}{i} p^i (1-p)^{n-i}$, and hence

$$\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right] = \sum_{m=0}^{n} \binom{n}{m} p^m (1-p)^{n-m} f\left(\frac{m}{n}\right) = f_n(p).$$

Chebyshev's inequality gives, uniformly in p,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \delta\right) \leq \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

since $p(1-p) \leq 1/4$. Let $M = \sup_{x\in[0,1]}|f(x)|$. Given $\varepsilon > 0$, uniform continuity supplies $\delta > 0$ such that $|x-y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$. Splitting the expectation over the events where $|S_n/n - p|$ is below or above δ,

$$\begin{aligned}
\mathbb{E}\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\right] &\leq \mathbb{E}\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbb{1}_{\{|S_n/n - p| < \delta\}}\right] + \mathbb{E}\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbb{1}_{\{|S_n/n - p| \geq \delta\}}\right] \\
&\leq \varepsilon + 2\sup_{x\in[0,1]}|f(x)|\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \delta\right) \\
&\leq \varepsilon + \frac{M}{2n\delta^2},
\end{aligned}$$

which is independent of p (!). Taking the supremum over p, then letting $n \to \infty$ and $\varepsilon \downarrow 0$, proves uniform convergence. $\square$

Triangular Arrays and Truncation

A triangular array of random variables has the form

$$\begin{array}{llll}
X_{11} & & & \\
X_{21} & X_{22} & & \\
X_{31} & X_{32} & X_{33} & \\
X_{41} & X_{42} & X_{43} & X_{44} \\
\vdots & & &
\end{array}$$

The sum of the nth row is

$$S_n = \sum_{i=1}^{n} X_{n,i}.$$

Almost sure convergence is often inappropriate for triangular arrays because different rows may be defined on different probability spaces.

Theorem 9.9 Let $\mu_n = \mathbb{E}[S_n]$ and $\sigma_n^2 = \mathrm{Var}(S_n)$. If $b_n > 0$ and $\sigma_n^2/b_n^2 \to 0$, then

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{\mathbb{P}} 0.$$

Proof. Since

$$\mathbb{E}\left[\left(\frac{S_n - \mu_n}{b_n}\right)^2\right] = \frac{\sigma_n^2}{b_n^2} \longrightarrow 0,$$

the conclusion follows from Lemma 9.6. □

Example 9.10 (Coupon collector's problem) Sample independently and uniformly from n coupon types. Let

$$\tau_k^{(n)} = \inf\{m : |\{X_1, \dots, X_m\}| = k\}$$

be the first time that k distinct types have appeared, and define $X_{n,1} = 1$ and

$$X_{n,k} = \tau_k^{(n)} - \tau_{k-1}^{(n)} \quad (2 \le k \le n).$$

These waiting times are independent and

$$X_{n,k} \sim \text{Geometric}\left(1 - \frac{k-1}{n}\right).$$

The total collection time $S_n = \sum_{k=1}^n X_{n,k}$ satisfies

$$\mu_n = \mathbb{E}[S_n] = n \sum_{m=1}^n \frac{1}{m} = n \log n + O(n),$$

and, since the variance of a geometric random variable with parameter p is $(1-p)/p^2 \le 1/p^2$,

$$\mathrm{Var}(S_n) \le n^2 \sum_{m=1}^n \frac{1}{m^2} \le Cn^2.$$

Taking $b_n = n \log n$ in Theorem 9.9 gives

$$\frac{S_n - \mu_n}{n \log n} \xrightarrow{\mathbb{P}} 0.$$

Since $\mu_n/(n \log n) \to 1$, it follows that

$$\frac{S_n}{n \log n} \xrightarrow{\mathbb{P}} 1.$$

Truncation extends this method beyond random variables with finite variances by replacing each variable with its bounded part and controlling the probability that the truncation changes the row sum.

Theorem 9.11 (Weak law for triangular arrays) For each n, let $X_{n,1}, \dots, X_{n,n}$ be independent. Let $b_n > 0$ with $b_n \to \infty$, and define

$$\overline{X}_{n,k} = X_{n,k} \mathbb{1}_{\{|X_{n,k}| \le b_n\}}.$$

Suppose that

1. The probability of any truncation in a row tends to zero:

$$\sum_{k=1}^{n} \mathbb{P}(|X_{n,k}| > b_n) \longrightarrow 0.$$

2. The normalized sum of truncated second moments tends to zero:

$$\frac{1}{b_n^2} \sum_{k=1}^{n} \mathbb{E}[\overline{X}_{n,k}^2] \longrightarrow 0.$$

If

$$S_n = \sum_{k=1}^{n} X_{n,k} \qquad \text{and} \qquad a_n = \sum_{k=1}^{n} \mathbb{E}[\overline{X}_{n,k}],$$

then $(S_n - a_n)/b_n \xrightarrow{\mathbb{P}} 0$.

Proof. Let

$$\overline{S}_n = \sum_{k=1}^{n} \overline{X}_{n,k}.$$

By the union bound and (1),

$$\mathbb{P}(S_n \neq \overline{S}_n) \leq \sum_{k=1}^{n} \mathbb{P}(|X_{n,k}| > b_n) \longrightarrow 0.$$

For $\varepsilon > 0$, Chebyshev's inequality and independence give

$$\mathbb{P}\left(\left|\frac{\overline{S}_n - a_n}{b_n}\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(\overline{S}_n)}{\varepsilon^2 b_n^2} = \frac{1}{\varepsilon^2 b_n^2}\sum_{k=1}^{n} \mathrm{Var}(\overline{X}_{n,k}) \leq \frac{1}{\varepsilon^2 b_n^2}\sum_{k=1}^{n} \mathbb{E}[\overline{X}_{n,k}^2] \longrightarrow 0$$

by (2). Combining the last two estimates proves the result. □

Lemma 9.12 (Tail integral formula) If $Y \geq 0$ and $p > 0$, then

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

Proof. Since

$$Y^p = \int_0^\infty p y^{p-1} \mathbb{1}_{\{Y>y\}}\, \mathrm{d}y\,,$$

The Tonelli theorem gives

$$\begin{aligned}\mathbb{E}[Y^p] &= \int Y^p \,\mathrm{d}\mathbb{P} = \int \left(\int_0^Y p y^{p-1}\,\mathrm{d}y\right) \mathrm{d}\mathbb{P} \\ &= \int \int_0^\infty p y^{p-1} \mathbb{1}_{\{Y>y\}}\,\mathrm{d}y\,\mathrm{d}\mathbb{P} = \int_0^\infty p y^{p-1}\mathbb{P}(Y > y)\,\mathrm{d}y\,.\end{aligned}$$

□

Theorem 9.13 (Weak law under a tail condition) Let $X_1, X_2, \ldots$ be independent and identically distributed, and suppose that

$$x\mathbb{P}(|X_1| > x) \longrightarrow 0 \quad \text{as } x \to \infty.$$

Put

$$S_n = \sum_{k=1}^{n} X_k \qquad \text{and} \qquad \mu_n = \mathbb{E}\big[X_1 \mathbb{1}_{\{|X_1| \leq n\}}\big]\,.$$

Then
$$\frac{S_n - n\mu_n}{n} \xrightarrow{\mathbb{P}} 0.$$

Proof. Apply Theorem 9.11 with $X_{n,k} = X_k$ and $b_n = n$. Condition (1) becomes
$$n\mathbb{P}(|X_1| > n) \longrightarrow 0.$$

For the second condition, write
$$\overline{X}_{n,k} = X_k \mathbb{1}_{\{|X_k| \le n\}}.$$

By Lemma 9.12,
$$\mathbb{E}[\overline{X}_{n,1}^2] \le \int_0^n 2y\mathbb{P}(|X_1| > y)\,\mathrm{d}y\,.$$

Let
$$c_i = \int_i^{i+1} 2y\mathbb{P}(|X_1| > y)\,\mathrm{d}y$$

so that $\mathbb{E}[\overline{X}_{n,1}^2] \le \sum_{i=0}^{n-1} c_i$. Since $y \mapsto y\mathbb{P}(|X_1| > y)$ is nonnegative and tends to zero, the integrals c_i also tend to zero. By Cesàro convergence,
$$\frac{1}{n}\mathbb{E}[\overline{X}_{n,1}^2] \le \frac{1}{n}\sum_{i=0}^{n-1} c_i \longrightarrow 0.$$

Therefore
$$\frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}[\overline{X}_{n,k}^2] \longrightarrow 0,$$

and Theorem 9.11 gives the claim because its centering constant is $a_n = n\mu_n$. □

Corollary 9.14 (Weak law of large numbers) If $X_1, X_2, \ldots$ are independent and identically distributed with $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = \mu$, then
$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu.$$

Proof. For $x > 0$,
$$x\mathbb{P}(|X_1| > x) \le \mathbb{E}\big[|X_1|\mathbb{1}_{\{|X_1| > x\}}\big] \longrightarrow 0$$

by the dominated convergence theorem. Thus Theorem 9.13 applies. Its truncated means satisfy

$$\mu_n = \mathbb{E}\big[X_1 \mathbb{1}_{\{|X_1| \le n\}}\big] \longrightarrow \mathbb{E}[X_1] = \mu$$

by the dominated convergence theorem again. Hence

$$\frac{S_n}{n} - \mu = \frac{S_n - n\mu_n}{n} + (\mu_n - \mu) \xrightarrow{\mathbb{P}} 0.$$

□

Example 9.15 (Saint Petersburg paradox) Let $X_1, X_2, \ldots$ be independent and identically distributed with

$$\mathbb{P}(X_i = 2^j) = 2^{-j}, \quad j \ge 1.$$

Although $\mathbb{E}[X_i] = \infty$, the truncation argument gives

$$\frac{S_n}{n \log_2 n} \xrightarrow{\mathbb{P}} 1.$$

Thus, over a horizon of n plays, the typical payoff per play has order $\log_2 n$.

Proof. For $n \ge 2$, set

$$b_n = n \log_2 n \qquad \text{and} \qquad m_n = \lfloor \log_2 b_n \rfloor .$$

Since $2^{m_n} \le b_n < 2^{m_n+1}$,

$$\mathbb{P}(X_1 > b_n) = \sum_{j=m_n+1}^{\infty} 2^{-j} = 2^{-m_n} < \frac{2}{b_n}.$$

It follows that

$$n\mathbb{P}(X_1 > b_n) < \frac{2}{\log_2 n} \longrightarrow 0.$$

Moreover,

$$\mathbb{E}\left[X_1^2 \mathbb{1}_{\{X_1 \le b_n\}}\right] = \sum_{j=1}^{m_n} 2^{2j} 2^{-j} = 2^{m_n+1} - 2 \le 2b_n.$$

Therefore,

$$\frac{n}{b_n^2} \mathbb{E}\left[X_1^2 \mathbb{1}_{\{X_1 \le b_n\}}\right] \le \frac{2}{\log_2 n} \longrightarrow 0.$$

Both hypotheses of Theorem 9.11 hold. The corresponding truncated mean is

$$\mathbb{E}\left[X_1 \mathbb{1}_{\{X_1 \le b_n\}}\right] = \sum_{j=1}^{m_n} 2^j 2^{-j} = m_n.$$

Consequently,

$$\frac{S_n - n m_n}{b_n} \xrightarrow{\mathbb{P}} 0.$$

Finally,

$$\frac{n m_n}{b_n} = \frac{m_n}{\log_2 n} \longrightarrow 1,$$

which proves the asserted convergence. □

Lecture 10 — Monday, November 26, 2018

4. Strong Laws and Almost Sure Convergence

Event Limits and Borel–Cantelli Lemmas

Almost sure convergence is an eventual pathwise property and can therefore be expressed through limits of events. The Borel–Cantelli lemmas connect these event limits to summability and independence.

Definition 10.1 For events $A_1, A_2, \dots \in \mathcal{F}$, define

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$$

and

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m.$$

The first event consists of outcomes belonging to infinitely many A_n; it is also denoted $\{A_n \text{ infinitely often}\}$ or $\{A_n \text{ i.o.}\}$ for short. The second consists of outcomes belonging to all but finitely many A_n.

For a fixed outcome ω, Figure 7 contrasts recurring membership with eventual membership. The ellipses matter: the upper row represents occurrences that continue arbitrarily far along the sequence, whereas the lower row remains in every event after some index N.

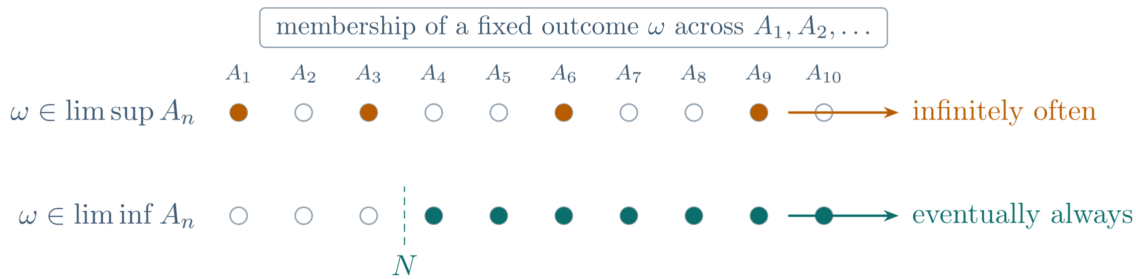

FIGURE 7

De Morgan's laws give the useful relation

$$\limsup_{n\to\infty} A_n = \left(\liminf_{n\to\infty} A_n^c\right)^c.$$

Lemma 10.2 A sequence (X_n) converges almost surely to 0 if and only if, for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n| > \varepsilon \text{ i.o}) = 0.$$

Proof. *Forward implication.* If $X_n \to 0$ almost surely, then for almost every ω and each $\varepsilon > 0$, there is an $N(\omega)$ such that $|X_n(\omega)| \leq \varepsilon$ for every $n \geq N(\omega)$. Hence the event occurs only finitely often.

Reverse implication. Apply the hypothesis with $\varepsilon = 1/k$ for $k \geq 1$. We have

$$\mathbb{P}\left(|X_i| > \frac{1}{k} \text{ i.o.}\right) = 0 \iff \mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty}\left\{|X_m| > \frac{1}{k}\right\}\right) = 0,$$

and the equality on the right implies

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty}\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty}\left\{|X_m| > \frac{1}{k}\right\}\right) = 0.$$

Taking complements, we have

$$\mathbb{P}\left(\bigcap_{k=1}^{\infty}\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty}\left\{|X_m| \le \frac{1}{k}\right\}\right) = 1.$$

This means that for each k, $|X_m| \le 1/k$ for sufficiently large m almost surely. Thus $X_m \to 0$ almost surely. ☐

Theorem 10.3 (Borel–Cantelli lemma I) If

$$\sum_{n=1}^{\infty}\mathbb{P}(A_n) < \infty,$$

then

$$\mathbb{P}(A_n \text{ i.o.}) = 0.$$

First proof. Let $X_i = \mathbb{1}_{A_i}$. By the monotone convergence theorem,

$$\mathbb{E}\left[\sum_{n=1}^{\infty}X_n\right] = \sum_{n=1}^{\infty}\mathbb{E}[X_n] = \sum_{n=1}^{\infty}\mathbb{P}(A_n) < \infty.$$

A nonnegative extended random variable with finite expectation is finite almost surely. Thus $\sum_{n=1}^{\infty}X_n < \infty$ almost surely, which implies that $X_n = 1$ only finitely often almost surely. Hence $\mathbb{P}(A_n \text{ i.o.}) = 0$. ☐

Second proof. Note that

$$\mathbb{P}\left(\bigcup_{m=n}^{\infty}A_m\right) \le \sum_{m=n}^{\infty}\mathbb{P}(A_m).$$

Let $B_n = \bigcup_{m=n}^{\infty}A_m$. Then B_n is a decreasing sequence of events, and $\mathbb{P}(B_n) \le \sum_{m=n}^{\infty}\mathbb{P}(A_m)$. Since $\sum_{n=1}^{\infty}\mathbb{P}(A_n) < \infty$, we have $\mathbb{P}(B_n) \to 0$ as $n \to \infty$. But $\limsup_{n\to\infty}A_n = \bigcap_{n=1}^{\infty}B_n$, and by continuity from above, we have

$$\mathbb{P}(\limsup_{n\to\infty}A_n) = \lim_{n\to\infty}\mathbb{P}(B_n) = 0.$$

☐

Remark 10.4 The converse of the first Borel–Cantelli lemma need not hold unless additional assumptions, such as independence, are imposed.

Theorem 10.5 (Subsequential characterization of convergence in probability) The sequence X_n converges to X in probability if and only if every subsequence (X_{n_m}) has a further subsequence $(X_{n_{m_k}})$ that converges to X almost surely.

The order of the quantifiers in this characterization is summarized in Figure 8: we begin with an arbitrary subsequence and may then select a further subsequence that converges almost surely.

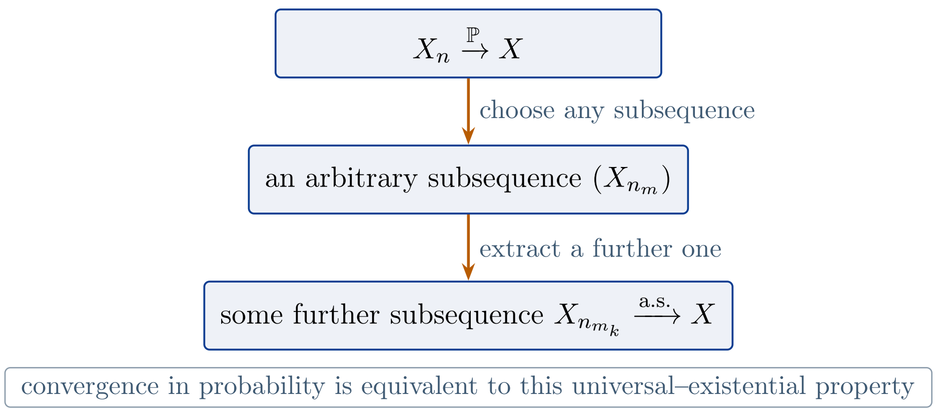

FIGURE 8

Proof. *Forward implication.* Suppose that $X_n \xrightarrow{\mathbb{P}} X$, and fix a subsequence (X_{n_m}). Choose m_k increasingly so that

$$\mathbb{P}(|X_{n_{m_k}} - X| > 1/k) \leq 2^{-k}.$$

The resulting series of probabilities is finite, so Theorem 10.3 and Lemma 10.2 imply $X_{n_{m_k}} \to X$ almost surely.

Reverse implication. Suppose every subsequence has a further almost surely convergent subsequence, but X_n does not converge to X in probability. Then there are $\varepsilon, \delta > 0$ and a subsequence (X_{n_m}) such that

$$\mathbb{P}(|X_{n_m} - X| > \varepsilon) \geq \delta$$

for every m. By assumption, a further subsequence converges almost surely to X. Bounded convergence applied to the corresponding indicator functions forces these probabilities to tend to zero, a contradiction. □

Corollary 10.6 Suppose that $X_n \xrightarrow{\mathbb{P}} X$ and $|X_n| \le Y$ for every n, with $\mathbb{E}[Y] < \infty$. Then X is integrable and
$$\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$

Proof. First note that X is integrable. Indeed, by Theorem 10.5, there exists a subsequence (X_{n_k}) such that $X_{n_k} \xrightarrow{\text{a.s.}} X$. Since $|X_{n_k}| \le Y$ almost surely for every k, it follows by taking limits that
$$|X| \le Y \quad \text{a.s.},$$
and hence $\mathbb{E}[|X|] < \infty$. We now show that the expectations converge. Let (X_{n_k}) be an arbitrary subsequence of (X_n). Since $X_{n_k} \xrightarrow{\mathbb{P}} X$, Theorem 10.5 gives a further subsequence $(X_{n_{k_j}})$ such that $X_{n_{k_j}} \xrightarrow{\text{a.s.}} X$. Moreover,
$$|X_{n_{k_j}}| \le Y \quad \text{a.s. for every } j,$$
with $\mathbb{E}[Y] < \infty$. Therefore, by the dominated convergence theorem,
$$\mathbb{E}[X_{n_{k_j}}] \longrightarrow \mathbb{E}[X].$$
Thus every subsequence of $(\mathbb{E}[X_n])$ has a further subsequence converging to $\mathbb{E}[X]$. This forces the whole sequence to converge to $\mathbb{E}[X]$. Indeed, if not, there would exist some $\varepsilon > 0$ and a subsequence (X_{n_k}) such that
$$\big|\mathbb{E}[X_{n_k}] - \mathbb{E}[X]\big| \ge \varepsilon \quad \text{for every } k.$$
But this subsequence has a further subsequence $(X_{n_{k_j}})$ satisfying
$$\mathbb{E}[X_{n_{k_j}}] \longrightarrow \mathbb{E}[X],$$
a contradiction. Hence
$$\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X].$$
□

Corollary 10.7 (Continuous mapping theorem) If $X_n \xrightarrow{\mathbb{P}} X$ and f is continuous, then
$$f(X_n) \xrightarrow{\mathbb{P}} f(X).$$

Proof. Let (X_{n_k}) be an arbitrary subsequence of (X_n). Since $X_{n_k} \xrightarrow{\mathbb{P}} X$, Theorem 10.5 gives a further subsequence $(X_{n_{k_j}})$ such that $X_{n_{k_j}} \xrightarrow{\text{a.s.}} X$. By continuity of f, $f(X_{n_{k_j}}) \xrightarrow{\text{a.s.}} f(X)$. In particular, $f(X_{n_{k_j}}) \xrightarrow{\mathbb{P}} f(X)$. We claim that this implies

$$f(X_n) \xrightarrow{\mathbb{P}} f(X).$$

Suppose otherwise. Then there exist $\varepsilon, \delta > 0$ and a subsequence (X_{n_k}) such that

$$\mathbb{P}\big(|f(X_{n_k}) - f(X)| > \varepsilon\big) \geq \delta \quad \text{for every } k.$$

By the preceding argument, this subsequence has a further subsequence $(X_{n_{k_j}})$ for which $f(X_{n_{k_j}}) \xrightarrow{\mathbb{P}} f(X)$. Hence

$$\mathbb{P}\big(|f(X_{n_{k_j}}) - f(X)| > \varepsilon\big) \longrightarrow 0,$$

contradicting the lower bound by δ. Therefore $f(X_n) \xrightarrow{\mathbb{P}} f(X)$ as claimed. □

Theorem 10.8 Let (y_n) be a sequence in a topological space. If every subsequence has a further subsequence converging to y, then $y_n \to y$.

Proof. If y_n does not converge to y, there is an open neighbourhood U of y such that infinitely many y_n lie outside U. Those terms form a subsequence, none of whose further subsequences can converge to y, which is a contradiction. □

Theorem 10.9 (Strong law under a fourth-moment condition) Let $X_1, X_2, \dots$ be independent and identically distributed with $\mathbb{E}[X_1] = \mu$ and $\mathbb{E}[|X_1|^4] < \infty$. Then

$$\frac{S_n}{n} \longrightarrow \mu \quad \text{almost surely.}$$

Proof. Let

$$\widetilde{X}_i := X_i - \mu,$$

so that the $\widetilde{X}_i$ are i.i.d. and centered. Expanding the fourth power, independence and centering imply that every term containing an odd power of some $\widetilde{X}_i$ has expectation zero. Hence

$$\mathbb{E}\left[\left(\sum_{i=1}^n \widetilde{X}_i\right)^4\right] = \sum_{i=1}^n \mathbb{E}[\widetilde{X}_i^4] + 6 \sum_{1\leq i<j\leq n} \mathbb{E}[\widetilde{X}_i^2]\mathbb{E}[\widetilde{X}_j^2] = n\,\mathbb{E}[\widetilde{X}_1^4] + 6\binom{n}{2}\big(\mathbb{E}[\widetilde{X}_1^2]\big)^2 \leq Cn^2$$

for some constant $C < \infty$ independent of n. Therefore

$$\mathbb{E}\left[\left(\frac{S_n - n\mu}{n}\right)^4\right] = \frac{1}{n^4}\mathbb{E}\left[\left(\sum_{i=1}^n \widetilde{X}_i\right)^4\right] \leq \frac{C}{n^2}.$$

Fix $\varepsilon > 0$. By the Markov inequality,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^4}\mathbb{E}\left[\left(\frac{S_n - n\mu}{n}\right)^4\right] \leq \frac{C}{\varepsilon^4 n^2}.$$

Consequently,

$$\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{C}{\varepsilon^4}\sum_{n=1}^\infty \frac{1}{n^2} < \infty.$$

By the first Borel–Cantelli lemma,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon \text{ i.o.}\right) = 0.$$

Applying this with $\varepsilon = 1/k$ for $k \in \mathbb{N}$ shows that, almost surely, for every k we eventually have $|S_n/n - \mu| \leq 1/k$. Hence

$$\frac{S_n}{n} \longrightarrow \mu \quad \text{almost surely.}$$

□

Theorem 10.10 (Second Borel–Cantelli lemma) If $A_1, A_2, \ldots$ are independent and

$$\sum_{n=1}^\infty \mathbb{P}(A_n) = \infty,$$

then

$$\mathbb{P}(A_n \text{ i.o.}) = 1.$$

Proof. Fix $N \in \mathbb{N}$. By continuity of probability from above and independence, for every $M \geq N$,

$$\mathbb{P}\left(\bigcap_{m=N}^M A_m^c\right) = \prod_{m=N}^M \mathbb{P}(A_m^c) = \prod_{m=N}^M \big(1 - \mathbb{P}(A_m)\big).$$

Using $1 - x \le e^{-x}$ for $x \ge 0$, we obtain

$$\mathbb{P}\left(\bigcap_{m=N}^{\infty} A_m^c\right) = \lim_{M\to\infty} \mathbb{P}\left(\bigcap_{m=N}^{M} A_m^c\right) \le \lim_{M\to\infty} \exp\left(-\sum_{m=N}^{M} \mathbb{P}(A_m)\right) = 0,$$

since every tail of the divergent series $\sum_m \mathbb{P}(A_m)$ also diverges. Therefore

$$\mathbb{P}\left(\bigcup_{m=N}^{\infty} A_m\right) = 1 \quad \text{for every } N.$$

The event that the A_n occur infinitely often is

$$\{A_n \text{ i.o.}\} = \bigcap_{N=1}^{\infty} \bigcup_{m=N}^{\infty} A_m.$$

The events $\bigcup_{m=N}^{\infty} A_m$ decrease as $N \to \infty$, so continuity from above gives

$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{N\to\infty} \mathbb{P}\left(\bigcup_{m=N}^{\infty} A_m\right) = 1.$$

□

Theorem 10.11 (Necessity of integrability for a finite strong-law limit) If $X_1, X_2, \dots$ are independent and identically distributed and $\mathbb{E}[|X_1|] = \infty$, then

$$\mathbb{P}(|X_n| \ge n \text{ i.o.}) = 1$$

and

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists and is finite}\right) = 0.$$

Proof. The **layer-cake identity** for the integer part gives

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \ge n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1| \ge n) = \mathbb{E}\big[\lfloor |X_1| \rfloor\big].$$

Since $|X_1| \le \lfloor |X_1| \rfloor + 1$ and $\mathbb{E}[|X_1|] = \infty$, we must have $\mathbb{E}[\lfloor |X_1| \rfloor] = \infty$. Thus

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \ge n) = \infty.$$

The events $\{|X_n| \geq n\}$ are independent, so the second Borel–Cantelli lemma gives

$$\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1.$$

Now let A and B be the events

$$A := \left\{\omega : \frac{S_n(\omega)}{n} \text{ converges to a finite limit}\right\},$$

$$\text{and} \qquad B := \{\omega : |X_n(\omega)| \geq n \text{ infinitely often}\}.$$

We have just shown that $\mathbb{P}(B) = 1$. We claim that $A \cap B = \varnothing$. Indeed, suppose that $\omega \in A$. For $n \geq 2$,

$$\frac{X_n(\omega)}{n} = \frac{S_n(\omega)}{n} - \frac{n-1}{n}\frac{S_{n-1}(\omega)}{n-1}.$$

Since $S_n(\omega)/n$ converges to a finite limit, both terms on the right converge to the same limit, and therefore

$$\frac{X_n(\omega)}{n} \longrightarrow 0.$$

In particular, $|X_n(\omega)|/n < 1$ for all sufficiently large n. Thus $\omega \notin B$, proving that $A \cap B = \varnothing$. Since $\mathbb{P}(B) = 1$, it follows that $\mathbb{P}(A) = 0$, which proves the second assertion. □

Theorem 10.12 (Strong law for independent event counts) If $A_1, A_2, \ldots$ are independent and

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$$

then

$$\frac{\sum_{i=1}^n \mathbb{1}_{A_i}}{\sum_{i=1}^n \mathbb{P}(A_i)} \longrightarrow 1 \quad \text{almost surely.}$$

In particular, infinitely many A_i occur almost surely.

Proof. Set

$$S_n := \sum_{i=1}^n \mathbb{1}_{A_i} \qquad \text{and} \qquad m_n := \mathbb{E}[S_n] = \sum_{i=1}^n \mathbb{P}(A_i).$$

By assumption, $m_n \to \infty$. Since the events are independent, the random variables $\mathbb{1}_{A_i}$ are independent, and hence

$$\operatorname{Var}(S_n) = \sum_{i=1}^n \operatorname{Var}(\mathbb{1}_{A_i}) = \sum_{i=1}^n \mathbb{P}(A_i)\big(1 - \mathbb{P}(A_i)\big) \leq m_n.$$

Therefore, for every $\delta > 0$, Chebyshev's inequality gives

$$\mathbb{P}\left(\left|\frac{S_n}{m_n} - 1\right| > \delta\right) = \mathbb{P}\left(|S_n - m_n| > \delta m_n\right) \leq \frac{\mathrm{Var}(S_n)}{\delta^2 m_n^2} \leq \frac{1}{\delta^2 m_n}.$$

Since $m_n \to \infty$, this already shows that

$$\frac{S_n}{m_n} \xrightarrow{\mathbb{P}} 1.$$

The bound is not necessarily summable over all n, so we pass to a suitable subsequence. For each $k \in \mathbb{N}$, let n_k be the first index for which

$$m_{n_k} \geq k^2.$$

Such an index exists because $m_n \to \infty$. Moreover, since $m_n - m_{n-1} = \mathbb{P}(A_n) \leq 1$, the minimality of n_k gives

$$k^2 \leq m_{n_k} < k^2 + 1.$$

Hence, for every $\delta > 0$,

$$\mathbb{P}\left(\left|\frac{S_{n_k}}{m_{n_k}} - 1\right| > \delta\right) \leq \frac{1}{\delta^2 m_{n_k}} \leq \frac{1}{\delta^2 k^2}.$$

Consequently,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{n_k}}{m_{n_k}} - 1\right| > \delta\right) < \infty.$$

By the first Borel–Cantelli lemma, for each fixed $\delta > 0$ the event

$$\left\{\left|\frac{S_{n_k}}{m_{n_k}} - 1\right| > \delta \text{ i.o.}\right\}$$

has probability zero. Applying this with $\delta = 1/r$, $r \in \mathbb{N}$, yields

$$\frac{S_{n_k}}{m_{n_k}} \longrightarrow 1 \quad \text{almost surely.}$$

It remains to extend the convergence from (n_k) to the full sequence. If $n_k \leq n < n_{k+1}$, then the monotonicity of S_n and m_n gives

$$\frac{S_{n_k}}{m_{n_{k+1}}} \leq \frac{S_n}{m_n} \leq \frac{S_{n_{k+1}}}{m_{n_k}}.$$

Equivalently,

$$\frac{S_{n_k}}{m_{n_k}} \frac{m_{n_k}}{m_{n_{k+1}}} \leq \frac{S_n}{m_n} \leq \frac{S_{n_{k+1}}}{m_{n_{k+1}}} \frac{m_{n_{k+1}}}{m_{n_k}}.$$

From the choice of n_k,

$$\frac{k^2}{(k+1)^2+1} \le \frac{m_{n_k}}{m_{n_{k+1}}} \le 1,$$

so

$$\frac{m_{n_k}}{m_{n_{k+1}}} \longrightarrow 1 \qquad \text{and} \qquad \frac{m_{n_{k+1}}}{m_{n_k}} \longrightarrow 1.$$

Since $S_{n_k}/m_{n_k} \to 1$ almost surely, both the lower and upper bounds above converge almost surely to 1. The squeeze theorem therefore gives

$$\frac{S_n}{m_n} \longrightarrow 1 \quad \text{almost surely.}$$

Finally, $m_n \to \infty$, so the almost sure convergence $S_n/m_n \to 1$ implies $S_n \to \infty$ almost surely. Since S_n counts the number of events among $A_1, \dots, A_n$ that occur, this means that infinitely many of the A_i occur almost surely. $\square$

The moral of the proof is that the strong law for independent event counts is a consequence of the first Borel–Cantelli lemma applied to a suitable subsequence.

Example 10.13 (Longest run of successes) Let $(X_i)_{i\in\mathbb{Z}}$ be independent and identically distributed with

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}.$$

For every $n \ge 1$, let

$$\ell_n = \sup\{m \ge 0 : X_{n-m+1} = X_{n-m+2} = \cdots = X_n = 1\}.$$

Thus ℓ_n is the length of the run of successes ending at time n; it is infinite only on a null event. Let

$$L_n = \max_{1\le m\le n} \ell_m.$$

Thus L_n is the longest run observed by time n.

$$\begin{array}{cccccccccc} \dots & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 \\ & & & n=0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \qquad \begin{array}{ll} \ell_0 = 0 & \\ \ell_1 = 1 & L_1 = 1 \\ \ell_2 = 2 & L_2 = 2 \\ \ell_3 = 3 & L_3 = 3 \\ \ell_4 = 4 & L_4 = 4 \\ \ell_5 = 0 & L_5 = 4 \end{array}$$

Then

$$\frac{L_n}{\log_2 n} \longrightarrow 1 \quad \text{almost surely.}$$

Thus the longest run among 1024 fair trials is typically close to 10.

Proof. For every $k \geq 0$,

$$\mathbb{P}(\ell_n = k) = 2^{-(k+1)},$$

because this event requires k successive successes immediately preceded by a failure. In particular,

$$\mathbb{P}(\ell_n \geq k) = 2^{-k}.$$

Fix $\varepsilon \in (0,1)$ and set

$$k_n = \lceil (1+\varepsilon) \log_2 n \rceil .$$

Then

$$\sum_{n=1}^{\infty} \mathbb{P}(\ell_n \geq k_n) = \sum_{n=1}^{\infty} 2^{-k_n} \leq \sum_{n=1}^{\infty} n^{-(1+\varepsilon)} < \infty.$$

By the first Borel–Cantelli lemma, almost surely there is a random N such that $\ell_n < k_n$ whenever $n \geq N$. Hence, for sufficiently large n,

$$L_n \leq \max\left\{ \max_{1 \leq m < N} \ell_m, \lceil (1+\varepsilon) \log_2 n \rceil \right\}.$$

Dividing by $\log_2 n$ gives

$$\limsup_{n\to\infty} \frac{L_n}{\log_2 n} \leq 1 + \varepsilon \quad \text{almost surely.}$$

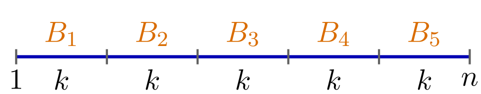

FIGURE 9

For all sufficiently large n, set

$$r_n = \lfloor (1-\varepsilon)\log_2 n \rfloor$$

and divide $\{1,\dots,n\}$ into

$$m_n = \left\lfloor \frac{n}{r_n} \right\rfloor$$

disjoint blocks of length r_n, discarding any remainder. This decomposition is illustrated in Figure 9. If at least one block consists entirely of successes, then $L_n \geq r_n$. Independence between the blocks and the inequality $1 - x \leq e^{-x}$ therefore give

$$\mathbb{P}(L_n < r_n) \leq \left(1 - 2^{-r_n}\right)^{m_n} \leq \exp\left(-m_n 2^{-r_n}\right).$$

For all sufficiently large n,

$$m_n \geq \frac{n}{2r_n} \qquad \text{and} \qquad 2^{-r_n} \geq n^{-(1-\varepsilon)}.$$

Consequently, for a constant $c_\varepsilon > 0$,

$$\mathbb{P}(L_n < r_n) \leq \exp\left(-c_\varepsilon \frac{n^\varepsilon}{\log_2 n}\right).$$

The series of these upper bounds converges. Another application of the first Borel–Cantelli lemma shows that $L_n \geq r_n$ almost surely for all sufficiently large n, and therefore

$$\liminf_{n\to\infty} \frac{L_n}{\log_2 n} \geq 1 - \varepsilon \quad \text{almost surely.}$$

Apply the upper and lower estimates simultaneously for every positive rational ε and then let $\varepsilon \downarrow 0$. The required almost sure limit follows. $\square$

Lecture 11 — Monday, December 03, 2018

Tail Events and Kolmogorov's Zero–One Law

Tail events are unaffected by changing any finite initial segment of a sequence. For independent random variables, this invariance forces every tail event to have probability zero or one.

Definition 11.1 Let

$$\mathscr{F}_n' = \sigma(X_n, X_{n+1}, \dots).$$

The **tail σ-algebra** of (X_n) is

$$\mathscr{T} = \bigcap_{n=1}^{\infty} \mathscr{F}_n'.$$

Anything in $\mathscr{T}$ happens in the *distant* future. Any event A belongs to $\mathscr{T}$ if and only if changing finitely many X_n does not affect whether A occurs. In particular, if $A \in \mathscr{T}$, then $\mathbb{1}_A$ is a tail random variable.

Example 11.2 The event $\{A_n \text{ i.o.}\}$ is a tail event for the indicator sequence $(\mathbb{1}_{A_n})$.

Example 11.3 The event that the series of partial sums $S_n = \sum_{i=1}^n X_i$ has a finite limit belongs to $\mathscr{T}$, because changing finitely many summands does not affect convergence.

Theorem 11.4 (Kolmogorov's zero–one law) If $X_1, X_2, \dots$ are independent and $A \in \mathscr{T}$, then $\mathbb{P}(A) \in \{0, 1\}$.

Proof. For each n, independence implies that $\sigma(X_1, \dots, X_{n-1})$ is independent of $\mathscr{F}_n'$, and hence of $\mathscr{T}$. Applying Theorem 8.13 to the increasing union gives

$$\mathscr{T} \text{ independent of } \sigma\left(\bigcup_{n=2}^{\infty} \sigma(X_1, \dots, X_{n-1})\right) = \sigma(X_1, X_2, \dots).$$

Since $\mathscr{T} \subseteq \sigma(X_1, X_2, \dots)$, the tail σ-algebra is independent of itself. Thus A is independent of itself, and

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2.$$

□

Corollary 11.5 If $X_1, X_2, \dots$ are independent, then, for every $\mu \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{S_n}{n} \to \mu\right) \in \{0, 1\}.$$

Note that in Corollary 11.5, the X_i need not be identically distributed, and the limit μ need not be the mean of the X_i. The corollary simply states that the event of convergence to a given limit has probability zero or one.

Proof. Changing finitely many variables changes S_n/n by a quantity that tends to zero. Thus the event $\{S_n/n \to \mu\}$ belongs to the tail σ-algebra, and Theorem 11.4 applies. $\square$

Maximal Inequalities and Random Series

Maximal inequalities control all partial sums at once and are consequently stronger than bounds for a single terminal sum. They provide the main tool for establishing almost sure convergence of random series.

Theorem 11.6 (Kolmogorov's maximal inequality) Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$ and finite variances. If $S_k = \sum_{i=1}^k X_i$, then for every $x > 0$,

$$\mathbb{P}\left(\max_{1 \le k \le n} |S_k| \ge x\right) \le \frac{\operatorname{Var}(S_n)}{x^2}.$$

This strengthens Chebyshev's inequality by controlling every partial sum simultaneously, at the cost of independence.

Proof. Fix $x > 0$. For $1 \le k \le n$, define

$$A_k := \big\{|S_k| \ge x,\ |S_j| < x \text{ for every } j < k\big\}.$$

Thus A_k is the event that time k is the first time the partial sums leave the interval $(-x, x)$; see Figure 10. In particular, the events $A_1, \ldots, A_n$ are pairwise disjoint, and

$$\left\{\max_{1 \le j \le n} |S_j| \ge x\right\} = \bigcup_{k=1}^n A_k.$$

We now estimate the contribution from each possible first hitting time. Fix k. Since

$$S_n = S_k + (S_n - S_k),$$

we have

$$\mathbb{E}\big[S_n^2 \mathbb{1}_{A_k}\big] = \mathbb{E}\big[S_k^2 \mathbb{1}_{A_k}\big] + 2\mathbb{E}\big[S_k \mathbb{1}_{A_k}(S_n - S_k)\big] \qquad + \mathbb{E}\big[(S_n - S_k)^2 \mathbb{1}_{A_k}\big].$$

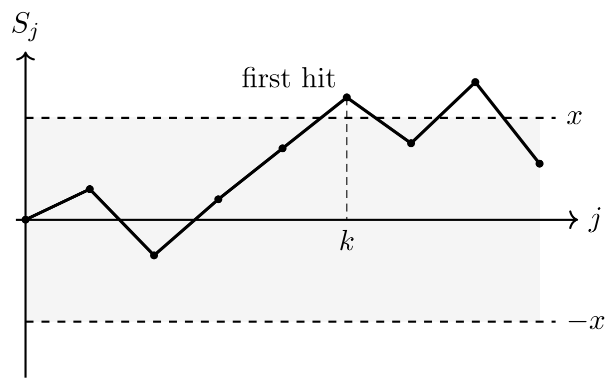

FIGURE 10

The event A_k depends only on $X_1, \ldots, X_k$, so $S_k \mathbb{1}_{A_k}$ is $\sigma(X_1, \ldots, X_k)$-measurable. On the other hand, $S_n - S_k = X_{k+1} + \cdots + X_n$ is independent of $\sigma(X_1, \ldots, X_k)$ and has mean zero. Hence

$$\mathbb{E}\big[S_k \mathbb{1}_{A_k}(S_n - S_k)\big] = \mathbb{E}\big[S_k \mathbb{1}_{A_k}\big]\mathbb{E}[S_n - S_k] = 0.$$

Since the remaining last term is nonnegative, it follows that

$$\mathbb{E}\big[S_n^2 \mathbb{1}_{A_k}\big] \geq \mathbb{E}\big[S_k^2 \mathbb{1}_{A_k}\big].$$

Because $\mathbb{E}[S_n] = 0$, we have $\mathrm{Var}(S_n) = \mathbb{E}[S_n^2]$. Using the disjointness of the A_k,

$$\mathrm{Var}(S_n) = \mathbb{E}[S_n^2] \geq \mathbb{E}\left[S_n^2 \mathbb{1}_{\bigcup_{k=1}^n A_k}\right] = \sum_{k=1}^n \mathbb{E}\big[S_n^2 \mathbb{1}_{A_k}\big] \geq \sum_{k=1}^n \mathbb{E}\big[S_k^2 \mathbb{1}_{A_k}\big].$$

But $S_k^2 \geq x^2$ on A_k, so

$$\mathrm{Var}(S_n) \geq x^2 \sum_{k=1}^n \mathbb{P}(A_k) = x^2\, \mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| \geq x\right).$$

Dividing by x^2 gives

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| \geq x\right) \leq \frac{\mathrm{Var}(S_n)}{x^2},$$

as required. $\square$

Theorem 11.7 (Kolmogorov convergence criterion) Let $X_1, X_2, \dots$ be independent random variables with $\mathbb{E}[X_i] = 0$. If

$$\sum_{i=1}^{\infty} \mathrm{Var}(X_i) < \infty,$$

then $\sum_{i=1}^{\infty} X_i$ converges almost surely and in probability.

Proof. Let $S_n = \sum_{i=1}^{n} X_i$. By Theorem 11.6, for $M \geq 1$ and $x > 0$,

$$\mathbb{P}\left(\max_{N<m\leq N+M} |S_m - S_N| \geq x\right) \leq \frac{\sum_{i=N+1}^{N+M} \mathrm{Var}(X_i)}{x^2}.$$

Letting $M \to \infty$ gives

$$\mathbb{P}\left(\sup_{m>N} |S_m - S_N| \geq x\right) \leq \frac{\sum_{i=N+1}^{\infty} \mathrm{Var}(X_i)}{x^2} \longrightarrow 0.$$

Define the decreasing sequence

$$W_N = \sup_{m,n\geq N} |S_m - S_n|.$$

Since $W_N \leq 2\sup_{m\geq N}|S_m - S_N|$, we have $W_N \to 0$ in probability. By Theorem 10.5, some subsequence converges to zero almost surely; monotonicity of W_N then forces the full sequence to converge to zero almost surely. Thus (S_n) is almost surely Cauchy and hence converges almost surely. Almost sure convergence implies convergence in probability. □

Theorem 11.8 (Kolmogorov's three-series theorem) Let $X_1, X_2, \dots$ be independent. Fix $A > 0$ and set

$$Y_n = X_n \mathbb{1}_{\{|X_n| \leq A\}}.$$

The series $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if all three of the following numerical series converge:

1. The tail-probability series converges:

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A).$$

2. The truncated-mean series converges:

$$\sum_{n=1}^{\infty} \mathbb{E}[Y_n].$$

3. The truncated-variance series converges:

$$\sum_{n=1}^{\infty} \mathrm{Var}(Y_n).$$

Proof. *Forward implication.* Suppose that $\sum_n X_n$ converges almost surely. Then $X_n \to 0$ almost surely, so the events $\{|X_n| > A\}$ occur only finitely often almost surely. These events are independent; therefore, the second Borel–Cantelli lemma gives

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty.$$

The first series consequently converges, and $X_n = Y_n$ eventually almost surely. It follows that $\sum_n Y_n$ also converges almost surely.

Let (Y_n') be an independent copy of (Y_n) on a product probability space, and put $Z_n = Y_n - Y_n'$. The series $\sum_n Z_n$ converges almost surely, the variables Z_n are independent and symmetric, and $|Z_n| \le 2A$. Set $t = 1/(2A)$. For $|u| \le 1$,

$$1 - \cos u \ge \frac{u^2}{4}.$$

Hence

$$\mathbb{E}[\cos(tZ_n)] = 1 - \mathbb{E}[1 - \cos(tZ_n)] \le 1 - \frac{t^2}{4}\mathbb{E}[Z_n^2] \le \exp\left(-\frac{t^2}{4}\mathbb{E}[Z_n^2]\right).$$

The factors on the left are positive because $|tZ_n| \le 1$. Almost sure convergence makes the tail sums $\sum_{n=N+1}^{M} Z_n$ tend to zero in probability as $M > N \to \infty$. Independence and symmetry therefore give

$$\prod_{n=N+1}^{M} \mathbb{E}[\cos(tZ_n)] = \mathbb{E}\left[\cos\left(t \sum_{n=N+1}^{M} Z_n\right)\right] \longrightarrow 1.$$

Comparing the last two displays shows that $\sum_n \mathbb{E}[Z_n^2] < \infty$. Since

$$\mathbb{E}[Z_n^2] = 2\,\mathrm{Var}(Y_n),$$

the variance series in (3) converges. By Theorem 11.7, the centered series

$$\sum_{n=1}^{\infty}\big(Y_n - \mathbb{E}[Y_n]\big)$$

converges almost surely. Subtracting its partial sums from those of the almost surely convergent series $\sum_n Y_n$ shows that the deterministic series in (2) converges as well.

Reverse implication. Suppose that all three numerical series converge. By (3) and Theorem 11.7,

$$\sum_{n=1}^{\infty}(Y_n - \mathbb{E}[Y_n])$$

converges almost surely. Adding the convergent numerical series in (2) shows that $\sum_n Y_n$ converges almost surely. Moreover, (1) and the first Borel–Cantelli lemma imply

$$\mathbb{P}(X_n \neq Y_n \text{ infinitely often}) = 0.$$

Thus the two series differ in only finitely many terms almost surely, and $\sum_n X_n$ converges almost surely. □

Theorem 11.9 (Kronecker's lemma) Let (a_n) be positive with $a_n \nearrow \infty$. If

$$\sum_{n=1}^{\infty}\frac{x_n}{a_n}$$

converges, then

$$\frac{1}{a_n}\sum_{m=1}^{n} x_m \longrightarrow 0.$$

Proof. Put $b_n = x_n/a_n$ and $B_n = \sum_{k=1}^{n} b_k$, so $B_n \to B$. Summation by parts gives

$$\sum_{k=1}^{n} a_k b_k = a_n B_n - \sum_{k=1}^{n-1}(a_{k+1} - a_k)B_k.$$

Write $B_k = B + r_k$, where $r_k \to 0$. After division by a_n, the contribution from B equals $a_1 B/a_n \to 0$. Given $\varepsilon > 0$, choose N with $|r_k| < \varepsilon$ for $k \geq N$. The finitely many earlier terms vanish after division by a_n, while the remaining terms have absolute value at most

$$\frac{\varepsilon}{a_n}\sum_{k=N}^{n-1}(a_{k+1}-a_k) \leq \varepsilon.$$

The term r_n also tends to zero, proving the result. □

Strong Laws of Large Numbers

The preceding results on truncation, random series, and deterministic summation now combine to prove the strong law under the optimal first moment assumption. We first require a lemma.

Lemma 11.10 Let $X_1, X_2, \dots$ be identically distributed with $\mathbb{E}[|X_1|] < \infty$, and define

$$Y_n := X_n \mathbb{1}_{\{|X_n| \leq n\}}, \qquad S_n := \sum_{k=1}^{n} X_k, \qquad \text{and} \qquad T_n := \sum_{k=1}^{n} Y_k.$$

If

$$\frac{T_n}{n} \longrightarrow \mu \quad \text{almost surely,}$$

then

$$\frac{S_n}{n} \longrightarrow \mu \quad \text{almost surely.}$$

Proof. Since the X_n are identically distributed,

$$\sum_{n=1}^{\infty}\mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty}\mathbb{P}(|X_1| > n).$$

Using the tail integral formula and the fact that $t \mapsto \mathbb{P}(|X_1| > t)$ is decreasing,

$$\sum_{n=1}^{\infty}\mathbb{P}(|X_1| > n) \leq \int_0^{\infty}\mathbb{P}(|X_1| > t)\,\mathrm{d}t = \mathbb{E}[|X_1|] < \infty.$$

Hence, by the first Borel–Cantelli lemma,

$$\mathbb{P}(X_n \neq Y_n \text{ infinitely often}) = 0.$$

Thus, almost surely, there is some N such that $X_n = Y_n$ for all $n \geq N$. On this event,

$$S_n - T_n = \sum_{k=1}^{N-1} (X_k - Y_k)$$

for every $n \geq N$, so $S_n - T_n$ is eventually constant. Therefore

$$\frac{S_n - T_n}{n} \longrightarrow 0 \quad \text{almost surely.}$$

Consequently, if $T_n/n \to \mu$ almost surely, then also $S_n/n \to \mu$ almost surely. □

Theorem 11.11 (Strong law of large numbers) Let $X_1, X_2, \ldots$ be independent and identically distributed with $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = \mu$. If

$$S_n = \sum_{k=1}^{n} X_k,$$

then

$$\frac{S_n}{n} \longrightarrow \mu \quad \text{almost surely.}$$

Proof. Define

$$Y_n := X_n \mathbb{1}_{\{|X_n| \leq n\}} \qquad \text{and} \qquad T_n := \sum_{k=1}^{n} Y_k.$$

By Lemma 11.10, it is enough to prove that

$$\frac{T_n}{n} \longrightarrow \mu \quad \text{almost surely.}$$

We first show that

$$\sum_{n=1}^{\infty} \frac{\operatorname{Var}(Y_n)}{n^2} < \infty.$$

Since $\operatorname{Var}(Y_n) \leq \mathbb{E}[Y_n^2]$ and $|Y_n| \leq n$, the layer-cake formula gives

$$\operatorname{Var}(Y_n) \leq \mathbb{E}[Y_n^2] = \int_0^n 2y \, \mathbb{P}(|Y_n| > y) \, \mathrm{d}y.$$

For $0 \leq y \leq n$, the event $\{|Y_n| > y\}$ is contained in $\{|X_n| > y\}$, and hence, by identical distribution,

$$\mathbb{P}(|Y_n| > y) \leq \mathbb{P}(|X_1| > y).$$

Therefore

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(Y_n)}{n^2} \le 2\sum_{n=1}^{\infty} \frac{1}{n^2} \int_0^n y\,\mathbb{P}(|X_1| > y)\,\mathrm{d}y = 2\int_0^{\infty} \left(y \sum_{n=1}^{\infty} \frac{\mathbb{1}_{\{y \le n\}}}{n^2} \right) \mathbb{P}(|X_1| > y)\,\mathrm{d}y$$

where the Tonelli theorem justifies the interchange of the sum and integral. We claim that

$$y \sum_{n=1}^{\infty} \frac{\mathbb{1}_{\{y \le n\}}}{n^2} \le 2 \quad \text{for every } y \ge 0.$$

If $0 \le y \le 1$, then

$$y \sum_{n=1}^{\infty} \frac{1}{n^2} \le \frac{\pi^2}{6} < 2.$$

If $y > 1$ and $m = \lceil y \rceil$, then

$$y \sum_{n=m}^{\infty} \frac{1}{n^2} \le y \left(\frac{1}{m^2} + \int_m^{\infty} \frac{\mathrm{d}t}{t^2} \right) \le 2.$$

It follows that

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(Y_n)}{n^2} \le 4 \int_0^{\infty} \mathbb{P}(|X_1| > y)\,\mathrm{d}y = 4\mathbb{E}[|X_1|] < \infty.$$

Now set $Z_n := Y_n - \mathbb{E}[Y_n]$. Since each Y_n is a function of X_n, the random variables $Z_1, Z_2, \ldots$ are independent and centered. Moreover,

$$\sum_{n=1}^{\infty} \mathrm{Var}\left(\frac{Z_n}{n} \right) = \sum_{n=1}^{\infty} \frac{\mathrm{Var}(Y_n)}{n^2} < \infty.$$

Hence Theorem 11.7 implies that

$$\sum_{n=1}^{\infty} \frac{Z_n}{n} = \sum_{n=1}^{\infty} \frac{Y_n - \mathbb{E}[Y_n]}{n}$$

converges almost surely. By Kronecker's lemma,

$$\frac{1}{n} \sum_{k=1}^{n} \big(Y_k - \mathbb{E}[Y_k] \big) \longrightarrow 0 \quad \text{almost surely.}$$

It remains to identify the limiting mean. Note that $\mathbb{E}[Y_n] = \mathbb{E}\left[X_1 \mathbb{1}_{\{|X_1| \le n\}}\right]$. As $n \to \infty$, $X_1 \mathbb{1}_{\{|X_1| \le n\}} \to X_1$ almost surely, while $\left| X_1 \mathbb{1}_{\{|X_1| \le n\}} \right| \le |X_1|$, and $\mathbb{E}[|X_1|] < \infty$. Thus the dominated convergence theorem gives

$$\mathbb{E}[Y_n] \longrightarrow \mathbb{E}[X_1] = \mu.$$

By Cesàro convergence,

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[Y_k] \longrightarrow \mu.$$

Combining the centered and deterministic parts,

$$\frac{T_n}{n} = \frac{1}{n}\sum_{k=1}^{n}\big(Y_k - \mathbb{E}[Y_k]\big) + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Y_k] \longrightarrow \mu \quad \text{almost surely.}$$

The conclusion now follows from Lemma 11.10. □

Corollary 11.12 If $X_1, X_2, \ldots$ are independent and identically distributed with $\mathbb{E}[X_1] = 0$ and $\operatorname{Var}(X_1) = \sigma^2 < \infty$, then, for every $\varepsilon > 0$,

$$\frac{S_n}{\sqrt{n}(\log n)^{1/2+\varepsilon}} \longrightarrow 0 \quad \text{almost surely.}$$

Proof. Fix $\varepsilon > 0$, and for $n \geq 2$ define

$$a_n := \sqrt{n}(\log n)^{1/2+\varepsilon}.$$

We will apply Kronecker's lemma to the series

$$\sum_{n=2}^{\infty} \frac{X_n}{a_n}.$$

Since the X_n are independent and centered, so are the random variables X_n/a_n. Moreover,

$$\sum_{n=2}^{\infty} \operatorname{Var}\left(\frac{X_n}{a_n}\right) = \sigma^2 \sum_{n=2}^{\infty} \frac{1}{n(\log n)^{1+2\varepsilon}}.$$

We verify that the series on the right converges. The function

$$x \longmapsto \frac{1}{x(\log x)^{1+2\varepsilon}}$$

is positive and decreasing for $x > 1$, and

$$\int_2^\infty \frac{\mathrm{d}x}{x(\log x)^{1+2\varepsilon}} = \int_{\log 2}^\infty \frac{\mathrm{d}u}{u^{1+2\varepsilon}} = \frac{1}{2\varepsilon(\log 2)^{2\varepsilon}} < \infty,$$

where we used the substitution $u = \log x$. Thus, by the integral test,

$$\sum_{n=2}^\infty \frac{1}{n(\log n)^{1+2\varepsilon}} < \infty.$$

Consequently,

$$\sum_{n=2}^\infty \operatorname{Var}\left(\frac{X_n}{a_n}\right) < \infty.$$

By Theorem 11.7, it follows that

$$\sum_{n=2}^\infty \frac{X_n}{a_n}$$

converges almost surely. Since (a_n) is increasing and $a_n \to \infty$, Kronecker's lemma gives

$$\frac{1}{a_n}\sum_{k=2}^n X_k \longrightarrow 0 \quad \text{almost surely.}$$

Finally,

$$\frac{S_n}{a_n} = \frac{X_1}{a_n} + \frac{1}{a_n}\sum_{k=2}^n X_k.$$

Since $X_1/a_n \to 0$ almost surely, we conclude that

$$\frac{S_n}{\sqrt{n}(\log n)^{1/2+\varepsilon}} \longrightarrow 0 \quad \text{almost surely.}$$

□

Theorem 11.13 (Strong law with one infinite mean component) If $X_1, X_2, \ldots$ are independent and identically distributed, $\mathbb{E}[X_1^+] = \infty$, and $\mathbb{E}[X_1^-] < \infty$, then

$$\frac{S_n}{n} \longrightarrow \infty \quad \text{almost surely.}$$

Proof. For $M > 0$, set

$$X_i^{+,M} = X_i^+ \wedge M.$$

These variables are integrable and identically distributed, so Theorem 11.11 gives

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i^+ \geq \mathbb{E}[X_1^+ \wedge M] \quad \text{almost surely.}$$

Letting $M \to \infty$ through the positive integers and using the monotone convergence theorem shows that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^+ \longrightarrow \infty$$

almost surely. On the other hand, the usual strong law applied to X_i^- gives

$$\frac{1}{n}\sum_{i=1}^{n} X_i^- \longrightarrow \mathbb{E}[X_1^-] < \infty$$

almost surely. Subtracting proves the claim. □

Theorem 11.14 (Elementary renewal theorem, almost-sure form) Let $X_1, X_2, \ldots$ be independent and identically distributed with $0 < X_i < \infty$ almost surely. Interpret X_i as the waiting time between the $(i-1)$st and ith arrivals. Let

$$T_n = \sum_{i=1}^{n} X_i$$

be the time of the nth arrival, and define the number of arrivals by time t as

$$T_0 = 0 \qquad \text{and} \qquad N_t = \max\{n \geq 0 : T_n \leq t\}.$$

If $\mathbb{E}[X_1] = \mu \in (0, \infty]$, then

$$\frac{N_t}{t} \longrightarrow \frac{1}{\mu} \quad \text{almost surely as } t \to \infty,$$

with the convention $1/\infty = 0$.

This result is very important if you manage a McDonald's. How will you best deploy your minimum wage staff to maximize the number of customers served per hour?

Proof. By Theorems 11.11 and 11.13,

$$\frac{T_n}{n} \longrightarrow \mu \quad \text{almost surely.}$$

In particular, $T_n \to \infty$ almost surely, so $N_t < \infty$ for every t almost surely. We will also need that

$$N_t \longrightarrow \infty \quad \text{as } t \to \infty$$

almost surely. Indeed, since each X_i is finite almost surely, every fixed partial sum T_m is finite almost surely. Thus, for every fixed m, once $t \geq T_m$ we have $N_t \geq m$. Hence $N_t \to \infty$ as $t \to \infty$.

Now work on an event of probability one on which

$$\frac{T_n}{n} \longrightarrow \mu \qquad \text{and} \qquad N_t \longrightarrow \infty.$$

Since N_t is integer-valued and tends to infinity, the sequence T_{N_t}/N_t is obtained by evaluating the convergent sequence T_n/n along indices tending to infinity. Therefore

$$\frac{T_{N_t}}{N_t} \longrightarrow \mu.$$

Similarly, since $N_t + 1 \to \infty$,

$$\frac{T_{N_t+1}}{N_t + 1} \longrightarrow \mu.$$

Moreover,

$$\frac{N_t + 1}{N_t} = 1 + \frac{1}{N_t} \longrightarrow 1.$$

By the definition of N_t,

$$T_{N_t} \leq t < T_{N_t+1}.$$

Dividing by N_t gives

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t} = \frac{T_{N_t+1}}{N_t + 1}\,\frac{N_t + 1}{N_t}.$$

The left-hand side converges to μ, while the product on the right converges to μ. Hence, by the squeeze theorem,

$$\frac{t}{N_t} \longrightarrow \mu \quad \text{almost surely.}$$

Equivalently,

$$\frac{N_t}{t} \longrightarrow \frac{1}{\mu} \quad \text{almost surely,}$$

with the usual convention $1/\infty = 0$ if $\mu = \infty$. $\square$

Remark 11.15 The proof uses the pathwise fact that a convergent sequence remains convergent along any random index tending to infinity. The analogous assertion can fail when the original convergence is only in probability and the random index depends on the sequence.

Appendix: Lecture and Tutorial Index