Probability and Measure Theory

Amit Rajaraman

Summer 2020

We primarily use "Probability Theory: A Comprehensive Course"[2] by Achim Klenke as reference for this course

Contents

§0. Notation

$\mathbb{N}$ represents the set $\{1, 2, \ldots\}$.

$\mathbb{N}_0$ represents the set $\{0, 1, 2, \ldots\}$.

For $x \in \mathbb{R}$ and $n \in \mathbb{N}$,

$$\binom{x}{r} = \frac{x(x-1)\cdots(x-r+1)}{r!}$$

is the generalised binomial coefficient.

For $n \in \mathbb{N}$, we denote $\{1, 2, \ldots, n\}$ as $[n]$ and $\{0, 1, 2, \ldots, n\}$ as $[n]_0$

For $a \in \mathbb{R}^n$, we denote the ith coordinate of a by a_i for each $i = 1, 2, \ldots, n$.

For $a, b \in \mathbb{R}^n$, we write $a < b$ if $a_i < b_i$ for each $i = 1, 2, \ldots, n$.

Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of reals. Then

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \left(\sup_{m\geq n} a_n \right) = \inf_{n\geq 0} \left(\sup_{m\geq n} a_n \right)$$

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \left(\inf_{m\geq n} a_n \right) = \sup_{n\geq 0} \left(\inf_{m\geq n} a_n \right)$$

Let A and B be two sets. We denote by

$$A \triangle B = (A \setminus B) \cup (B \setminus A)$$

the *symmetric difference* of A and B.

If sets A and B are disjoint, we represent their union as $A \uplus B$ (similar to the $\oplus$ notation in linear algebra).

§1. Measure Theory

Before beginning a rigorous study of probability theory, it is necessary to understand some parts of basic measure theory.

1.1. Classes of Sets

Let Ω be a non-empty set and $\mathcal{A} \subseteq 2^\Omega$, where 2^Ω is the power set of Ω. Then

Definition 1.1. $\mathcal{A}$ is called

- $\cap$-closed (closed under intersections) or a π-system if $A \cap B \in \mathcal{A}$ for all $A, B \in \mathcal{A}$.
- σ-$\cap$-closed (closed under countable intersections) if $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$ for any choice of countably many sets $A_1, A_2, \ldots \in \mathcal{A}$.
- $\cup$-closed (closed under unions) if $A \cup B \in \mathcal{A}$ for all $A, B \in \mathcal{A}$.
- σ-$\cup$-closed (closed under countable unions) if $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ for any choice of countably many sets $A_1, A_2, \ldots \in \mathcal{A}$.
- $\setminus$-closed (closed under differences) if $A \setminus B \in \mathcal{A}$ for all $A, B \in \mathcal{A}$.
- closed under complements if $A^c = \Omega \setminus A \in \mathcal{A}$ for all $A \in \mathcal{A}$.

Theorem 1.1. Let $\mathcal{A}$ be closed under complements. Then $\mathcal{A}$ is $\cup$-closed (σ-$\cup$-closed) if and only if $\mathcal{A}$ is closed $\cap$-closed (σ-$\cap$-closed).

The above is relatively straightforward to prove using De Morgan's Laws.

Theorem 1.2. Let $\mathcal{A}$ be $\setminus$-closed. Then

(a) $\mathcal{A}$ is $\cap$-closed,

(b) if $\mathcal{A}$ is σ-$\cup$-closed, then $\mathcal{A}$ is σ-$\cap$-closed.

(c) Any countable union of sets in $\mathcal{A}$ can be expressed as a countable union of pairwise disjoint sets in $\mathcal{A}$.

Proof.

(a) For $A, B \in \mathcal{A}$, $A \cap B = A \setminus (A \setminus B) \in \mathcal{A}$.

(b) Let $A_1, A_2, \ldots \in \mathcal{A}$. Then

$$\begin{aligned}\bigcap_{i=1}^{\infty} A_i &= \bigcap_{i=1}^{\infty} (A_1 \cap A_i) \\ &= \bigcap_{i=1}^{\infty} A_1 \setminus (A_1 \setminus A_i) \\ &= A_1 \setminus \bigcup_{i=1}^{\infty} (A_1 \setminus A_i).\end{aligned}$$

(c) Let $A_1, A_2, \ldots \in \mathcal{A}$. We then have

$$\bigcup_{i=1}^{\infty} A_i = A_1 \uplus (A_2 \setminus A_1) \uplus ((A_3 \setminus A_2) \setminus A_1) \uplus \cdots$$

The result follows.

■

This equivalence between $\cap$ and $\cup$ if the class is $\setminus$-closed is apparent from De Morgan's laws.

Definition 1.2 (Algebra)**.** A class of sets $\mathcal{A} \subseteq 2^\Omega$ is called an *algebra* if

(i) $\Omega \in \mathcal{A}$,

(ii) $\mathcal{A}$ is $\setminus$-closed, and

(iii) $\mathcal{A}$ is $\cup$-closed.

Definition 1.3 (σ-algebra)**.** A class of sets $\mathcal{A} \subseteq 2^\Omega$ is called a *σ-algebra* if

(i) $\Omega \in \mathcal{A}$,

(ii) $\mathcal{A}$ is closed under complements, and

(iii) $\mathcal{A}$ is σ-$\cup$-closed.

σ-algebras are also known as *σ-fields*.
Note that any σ-algebra is an algebra (but the converse is not true).

Theorem 1.3. A class of sets $\mathcal{A} \subseteq 2^\Omega$ is an algebra if and only if

(a) $\Omega \in \mathcal{A}$,

(b) $\mathcal{A}$ is closed under complements, and

(c) $\mathcal{A}$ is $\cap$-closed.

The proof of the above is left as an exercise to the reader.

Definition 1.4 (Ring)**.** A class of sets $\mathcal{A} \subseteq 2^\Omega$ is called a *ring* if

(i) $\varnothing \in \mathcal{A}$,

(ii) $\mathcal{A}$ is $\setminus$-closed, and

(iii) $\mathcal{A}$ is $\cup$-closed.

Further, a ring is a *σ-ring* if it is σ-$\cup$-closed.

Definition 1.5 (Semiring)**.** A class of sets $\mathcal{A} \subseteq 2^\Omega$ is called a *semiring* if

(i) $\varnothing \in \mathcal{A}$,

(ii) for any $A, B \in \mathcal{A}$, $A \setminus B$ is a finite union of mutually disjoint sets in $\mathcal{A}$, and

(iii) $\mathcal{A}$ is $\cap$-closed.

Definition 1.6 (λ-system)**.** A class of sets $\mathcal{A} \subseteq 2^\Omega$ is called a *λ-system* (or *Dynkin's λ-system*) if

(i) $\Omega \in \mathcal{A}$,

(ii) for any $A, B \in \mathcal{A}$ with $B \subseteq A$, $A \setminus B \in \mathcal{A}$, and

(iii) $\biguplus_{i=1}^{\infty} A_i \in \mathcal{A}$ for any choice of countably many pairwise disjoint sets $A_1, A_2, \ldots \in \mathcal{A}$.

Among the above classes of sets, σ-algebras in particular are extremely important as we shall use them when defining probability.

Theorem 1.4.

(a) Every σ-algebra is also a λ-system, an algebra and a σ-ring.

(b) Every σ-ring is a ring, and every ring is a semiring.

(c) Every algebra is a ring. An algebra on a finite set Ω is a σ-algebra.

Proof.

(a) Let $\mathcal{A}$ be a σ-algebra. Then for any $A, B \in \mathcal{A}$, $A \setminus B = (A^c \cup B)^c \in \mathcal{A}$ and $A \cap B = (A^c \cup B^c)^c \in \mathcal{A}$, that is, $\mathcal{A}$ is $\setminus$-closed and $\cup$-closed. The result follows.

(b) Let $\mathcal{A}$ be a ring. Then theorem 1.1 implies that $\mathcal{A}$ is $\cap$-closed. The result follows.

(c) Let $\mathcal{A}$ be an algebra. With proof similar to the first part of this theorem, it is seen that $\mathcal{A}$ is $\setminus$-closed. We have $\varnothing = \Omega \setminus \Omega \in \mathcal{A}$ and thus, it is a ring. If Ω is finite, then $\mathcal{A}$ is finite. Thus any countable union of sets is a finite union of sets and the result follows.

■

Definition 1.7. Let $A_1, A_2, \ldots$ be subsets of Ω. Then

$$\liminf_{n\to\infty} A_n := \bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} A_j \text{ and } \limsup_{n\to\infty} A_n := \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j$$

are respectively called the *limit inferior* and *limit superior*, of the sequence $(A_n)_{n\in\mathbb{N}}$.

The above may be rewritten as

$$A_* := \liminf_{n\to\infty} A_n = \{\omega \in \Omega : |n \in \mathbb{N} : \omega \notin A_n| < \infty\}$$
$$A^* := \limsup_{n\to\infty} A_n = \{\omega \in \Omega : |n \in \mathbb{N} : \omega \in A_n| = \infty\}$$

That is, A_* represents the set of elements that do not appear in a finite number of sets and A^* represents the set of elements that appear in an infinite number of sets. This implies that $A_* \subseteq A^*$. (Why is the opposite not necessarily true?)

Definition 1.8 (Indicator function)**.** Let A be a subset of Ω. The *indicator function on* A is defined by

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

With the above notation, it may be shown that

$$\mathbb{1}_{A_*} = \liminf_{n\to\infty} \mathbb{1}_{A_n} \text{ and } \mathbb{1}_{A^*} = \limsup_{n\to\infty} \mathbb{1}_{A_n}.$$

If $\mathcal{A} \subseteq 2^\Omega$ is a σ-algebra and if $A_n \in \mathcal{A}$ for every $n \in \mathbb{N}$, then $A_* \in \mathcal{A}$ and $A^* \in \mathcal{A}$. This is clear from the fact that σ-algebras are closed under countable unions and intersections.
Proving the above statements is left as an exercise to the reader.

Theorem 1.5. Let I be some index set and $\mathcal{A}_i$ be a σ-algebra for each $i \in I$. Then the intersection $\mathcal{A}_I = \bigcap_{i\in I} \mathcal{A}_i$ is also a σ-algebra.

Proof. We can prove this by using the three conditions in the definition of a σ-algebra.

(i) Since $\Omega \in \mathcal{A}_i$ for every $i \in I$, $\Omega \in \mathcal{A}_I$.

(ii) Let $A \in \mathcal{A}_I$. Then $A \in \mathcal{A}_i$ for each $i \in I$ and thus $A^c \in \mathcal{A}_i$ for each $i \in I$. Therefore, $A^c \in \mathcal{A}_I$.

(iii) Let $A_1, A_2, \ldots \in \mathcal{A}_I$. Then $A_n \in \mathcal{A}_i$ for each $n \in \mathbb{N}$ and $i \in I$. Thus $A = \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}_i$ for each i as well. The result follows.

■

A similar statement holds for λ-systems.

Theorem 1.6. Let $\mathcal{E} \subseteq 2^\Omega$. Then there exists a smallest σ-algebra $\sigma(\mathcal{E})$ with $\mathcal{E} \subseteq \sigma(\mathcal{E})$:

$$\sigma(\mathcal{E}) = \bigcap_{\substack{\mathcal{A} \subseteq 2^\Omega \text{ is a } \sigma\text{-algebra} \\ \mathcal{E} \subseteq \mathcal{A}}} \mathcal{A}.$$

$\sigma(\mathcal{E})$ is called the *σ-algebra generated by $\mathcal{E}$* and $\mathcal{E}$ is called a *generator of $\sigma(\mathcal{E})$*.

Proof. 2^Ω is a σ-algebra that contains $\mathcal{E}$ so the intersection is non-empty. By theorem 1.5, $\sigma(\mathcal{E})$ is a σ-algebra. ■

Similar to the above, $\delta(\mathcal{E})$ is defined as the λ-system generated by $\mathcal{E}$.

We always have the following:

1. $\mathcal{E} \subseteq \sigma(\mathcal{E})$.
2. If $\mathcal{E}_1 \subseteq \mathcal{E}_2$, then $\sigma(\mathcal{E}_1) \subseteq \sigma(\mathcal{E}_2)$.
3. $\mathcal{A}$ is a σ-algebra if and only if $\sigma(\mathcal{A}) = \mathcal{A}$.

Similar statements hold for λ-systems. Further, $\delta(\mathcal{E}) \subseteq \sigma(\mathcal{E})$. This is to be expected as σ-algebras have more "structure" than λ-systems.

Theorem 1.7 ($\cap$-closed λ-system)**.** Let $\mathcal{D} \subseteq 2^\Omega$ be a λ-system. Then $\mathcal{D}$ is a π-system if and only if $\mathcal{D}$ is a σ-algebra.

Proof. If $\mathcal{D}$ is a σ-algebra, then it is obviously a π-system. Let $\mathcal{D}$ be a π-system. Then

(a) As $\mathcal{D}$ is a λ-system, $\Omega \in \mathcal{D}$.

(b) Let $A \in \mathbb{D}$. Since $\Omega \in \mathcal{D}$ and $\mathcal{D}$ is a λ-system, $A^c = \Omega \setminus A \in \mathcal{D}$.

(c) Let $A, B \in \mathcal{D}$. We have $A \cap B \in \mathcal{D}$. We now have $A \setminus B = A \setminus (A \cap B) \in \mathcal{D}$, that is, $\mathcal{D}$ is $\setminus$-closed.
Let $A_1, A_2, \ldots \in \mathcal{D}$. Then by theorem 1.2, there exist $B_1, B_2, \ldots \in \mathcal{D}$ such that

$$\bigcup_{i=1}^{\infty} A_i = \biguplus_{i=1}^{\infty} B_i \in \mathcal{D}.$$

This completes the proof. ■

Theorem 1.8 (Dynkin's π-λ theorem)**.** If $\mathcal{E} \subseteq 2^\Omega$ is a π-system, then $\delta(\mathcal{E}) = \sigma(\mathcal{E})$.

Proof. We already have $\delta(\mathcal{E}) \subseteq \sigma(\mathcal{E})$. We must now prove the reverse inclusion. We shall show that $\delta(\mathcal{E})$ is a π-system.
For each $E \in \delta(\mathcal{E})$, let

$$\mathcal{D}_E = \{A \in \delta(\mathcal{E}) : A \cap E \in \delta(\mathcal{E})\}.$$

To show that $\delta(\mathcal{E})$ is a π-system, it suffices to show that $\delta(\mathcal{E}) \subseteq \mathcal{D}_E$ for all $E \in \delta(\mathcal{E})$. We shall first show that $\mathcal{D}_E$ is a λ-system for each $E \in \mathcal{E}$ by checking each of the conditions in definition 1.6.

(a) We clearly have $\Omega \in \mathcal{D}_E$ as $\Omega \cap E = E$.

(b) For any $A, B \in \mathcal{D}_E$ with $A \subseteq B$,

$$(B \setminus A) \cap E = (B \cap E) \setminus (A \cap E) \in \delta(\mathcal{E}).$$

(c) Let $A_1, A_2, \ldots \in \mathcal{D}_E$ be mutually disjoint sets. Then

$$\left(\biguplus_{i=1}^{\infty} A_i\right) \cap E = \biguplus_{i=1}^{\infty} (A_i \cap E) \in \delta(\mathcal{E}).$$

Now since $\mathcal{D}_E$ is a λ-system and $\mathcal{E} \subseteq \mathcal{D}_E$ (Why?), $\delta(\mathcal{E}) \subseteq \mathcal{D}_E$.
Now that we have shown that $\delta(\mathcal{E})$ is a π-system, the result follows by theorem 1.7. ■

Definition 1.9 (Topology)**.** Let $\Omega \neq \varnothing$ be an arbitrary set. A class of sets $\tau \subseteq 2^\Omega$ is called a *topology* on 2^Ω if

(i) $\varnothing, \Omega \in \tau$,

(ii) τ is $\cap$-closed, and

(iii) for any $\mathcal{F} \subseteq \tau$, $\bigcup_{A \in \mathcal{F}} A \in \tau$.

In the above case, the pair (Ω, τ) is called a *topological space*. The sets $A \in \tau$ are called *open* and the sets $A \subseteq \Omega$ with $A^c \in \tau$ are called *closed*.
Note that in contrast with σ-algebras, topologies are closed under only finite intersections but are also closed under arbitrary unions.

For example, consider the natural topology on $\mathbb{R}$ which consists of all open intervals in $\mathbb{R}$ and any arbitrary union of them.

Definition 1.10 (Borel σ-algebra)**.** Let (Ω, τ) be a topological space. The σ-algebra

$$\mathcal{B}(\Omega) = \mathcal{B}(\Omega, \tau) = \sigma(\tau)$$

that is generated by the open sets is called the *Borel σ-algebra on Ω*. The elements $A \in \mathcal{B}(\Omega, \tau)$ are called *Borel sets* or *Borel measurable sets*.

A Borel σ-algebra that we shall often encounter is $\mathcal{B}(\mathbb{R}^n)$ for $n \in \mathbb{N}$. Consider the following classes of sets:

$$\begin{aligned}
\mathcal{A}_1 &= \{A \subseteq \mathbb{R}^n : A \text{ is open}\} \\
\mathcal{A}_2 &= \{A \subseteq \mathbb{R}^n : A \text{ is closed}\} \\
\mathcal{A}_3 &= \{A \subseteq \mathbb{R}^n : A \text{ is compact}\} \\
\mathcal{A}_4 &= \{(a, b) : a, b \in \mathbb{Q}^n \text{ and } a < b\} \\
\mathcal{A}_5 &= \{(a, b] : a, b \in \mathbb{Q}^n \text{ and } a < b\} \\
\mathcal{A}_6 &= \{[a, b) : a, b \in \mathbb{Q}^n \text{ and } a < b\} \\
\mathcal{A}_7 &= \{[a, b] : a, b \in \mathbb{Q}^n \text{ and } a < b\} \\
\mathcal{A}_8 &= \{(-\infty, b) : b \in \mathbb{Q}^n\} \\
\mathcal{A}_9 &= \{(-\infty, b] : b \in \mathbb{Q}^n\} \\
\mathcal{A}_{10} &= \{(a, \infty) : a \in \mathbb{Q}^n\} \\
\mathcal{A}_{11} &= \{[a, \infty) : a \in \mathbb{Q}^n\}
\end{aligned}$$

It may be proved that $\mathcal{B}(\mathbb{R}^n)$ is generated by any of the classes of sets $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{11}$.

For $A \in \mathcal{B}(\mathbb{R})$, we represent by $\mathcal{B}(\mathbb{R})|_A$ the restriction of $\mathcal{B}(\mathbb{R})$ to A. It may be proved that this is equal to $\mathcal{B}(A)$, the σ-algebra generated by the open subsets of A.

1.2. Measure

Definition 1.11. Let $\mathcal{A} \subseteq 2^\Omega$ and let $\mu : \mathcal{A} \to [0, \infty]$ be a set function. We say that μ is

(i) *monotone* if for any $A, B \in \mathcal{A}$, $A \subseteq B$ implies that $\mu(A) \leq \mu(B)$,

(ii) *additive* if for any choice of finitely many mutually disjoint sets $A_1, \ldots, A_n \in \mathcal{A}$ with $\biguplus_{i=1}^n A_i \in \mathcal{A}$,

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

(iii) σ-additive if for any choice of countably many mutually disjoint sets $A_1, A_2, \ldots \in \mathcal{A}$ with $\biguplus_{i=1}^{\infty} A_i \in \mathcal{A}$,

$$\mu\left(\biguplus_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i),$$

(iv) subadditive if for any choice of finitely many sets $A, A_1, A_2, \ldots, A_n \in \mathcal{A}$ with $A \subseteq \bigcup_{i=1}^{n} A_i$, we have

$$\mu(A) \leq \sum_{i=1}^{n} \mu(A_i), \text{ and}$$

(v) σ-subadditive if for any choice of countably many sets $A, A_1, A_2, \ldots \in \mathcal{A}$ with $A \subseteq \bigcup_{i=1}^{\infty} A_i$, we have

$$\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i).$$

Definition 1.12. Let $\mathcal{A}$ be a semiring and $\mu : \mathcal{A} \to [0, \infty]$ be a set function with $\mu(\varnothing) = 0$. μ is called a

(i) *content* if μ is additive,

(ii) *premeasure* if μ is σ-additive, and

(iii) *measure* if μ is σ-additive and $\mathcal{A}$ is a σ-algebra.

Theorem 1.9 (Properties of contents)**.** Let $\mathcal{A}$ be a semiring and μ be a content on $\mathcal{A}$. Then

(a) If $\mathcal{A}$ is a ring, then $\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$ for any $A, B \in \mathcal{A}$.

(b) μ is monotone. If $\mathcal{A}$ is a ring, then $\mu(B) = \mu(A) + \mu(B \setminus A)$ for any $A, B \in \mathcal{A}$ with $A \subseteq B$.

(c) μ is subadditive. If μ is σ-additive, then it is also σ-subadditive.

(d) If $\mathcal{A}$ is a ring, then

$$\sum_{n=1}^{\infty} \mu(A_n) \leq \mu\left(\bigcup_{n=1}^{\infty} A_n\right)$$

for any choice of countably many mutually disjoint sets $A_1, A_2, \ldots \in \mathcal{A}$ with $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Proof.

(a) Note that $A \cup B = A \uplus (B \setminus A)$ and $B = (A \cap B) \uplus (B \setminus A)$. As μ is additive,

$$\mu(A \cup B) = \mu(A) + \mu(B \setminus A) \text{ and } \mu(B) = \mu(A \cap B) + \mu(B \setminus A).$$

The result follows.

(b) Let $A \subseteq B$. If $B \setminus A \in \mathcal{A}$ (which is true in the case of a ring), we have $B = A \uplus (B \setminus A)$ and thus

$$\mu(B) = \mu(A) + \mu(B \setminus A).$$

If $\mathcal{A}$ is just a semiring, then there exist $n \in \mathbb{N}$ and mutually disjoint sets $C_1, C_2, \ldots, C_n \in \mathcal{A}$ such that

$$B \setminus A = \biguplus_{i=1}^{n} C_i.$$

In either case, we have $\mu(A) \leq \mu(B)$.

(c) Let $A, A_1, A_2, \ldots, A_n \in \mathcal{A}$ such that $A \subseteq \bigcup_{i=1}^n A_i$. Let $B_1 = A_1$ and for each $k = 2, 3, \ldots, n$, let

$$B_k = A_k \setminus \left(\bigcup_{i=1}^{k-1} A_i \right).$$

Note that any two B_is are disjoint. As μ is additive and monotone, we have

$$\begin{aligned} \mu(A) &\leq \mu\left(\bigcup_{i=1}^n A_i \right) \\ &= \mu\left(\biguplus_{i=1}^n B_i \right) \\ &= \sum_{i=1}^n \mu(B_i) \leq \sum_{i=1}^n \mu(A_i). \end{aligned}$$

We can similarly prove that if μ is σ-additive, then it is σ-subadditive.

(d) Let $A = \bigcup_{i=1}^\infty A_i \in \mathcal{A}$. Since μ is additive and monotone,

$$\sum_{i=1}^m \mu(A_i) = \mu\left(\biguplus_{i=1}^m A_i \right) \leq \mu(A) \text{ for any } m \in \mathbb{N}.$$

The result follows.

■

Note that if equality holds in the fourth part of the above theorem, μ is a premeasure.

Definition 1.13 (Finite content)**.** Let $\mathcal{A}$ be a semiring. A content μ on A is called

(i) *finite* if $\mu(A) < \infty$ for all $A \in \mathcal{A}$ and

(ii) σ*-finite* if there exists a sequence of sets $\Omega_1, \Omega_2, \ldots \in \mathcal{A}$ such that $\Omega = \bigcup_{i=1}^\infty \Omega_i$ and $\mu(\Omega_i) < \infty$ for every $i \in \mathbb{N}$.

Definition 1.14. Let $A, A_1, A_2, \ldots$ be sets. We write

(i) $A_n \uparrow A$ if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ and $\bigcup_{i=1}^\infty A_i = A$. In this case, we say that A_n increases to A.

(ii) $A_n \downarrow A$ if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ and $\bigcap_{i=1}^\infty A_i = A$. In this case, we say that A_n decreases to A.

For example, if $A_n = \left(-\frac{1}{n}, \frac{1}{n}\right)$ for $n \in \mathbb{N}$, then $A_n \downarrow \{0\}$.

Definition 1.15 (Continuity of contents)**.** Let μ be a content on the ring $\mathcal{A}$. μ is called

(i) *lower semicontinuous* if $\lim_{n\to\infty} \mu(A_n) = \mu(A)$ for any $A \in \mathcal{A}$ and sequence $(A_n)_{n\in\mathbb{N}}$ such that $A_n \uparrow A$,

(ii) *upper semicontinuous* if $\lim_{n\to\infty} \mu(A_n) = \mu(A)$ for any $A \in \mathcal{A}$ and sequence $(A_n)_{n\in\mathbb{N}}$ such that $\mu(A_n) < \infty$ for some n (this implies that it holds for all $n \in \mathbb{N}$) and $A_n \downarrow A$,

(iii) $\varnothing$*-continuous* if (ii) holds for $A = \varnothing$.

Theorem 1.10. Let μ be a content on the ring $\mathcal{A}$. The following properties are equivalent:

(a) μ is σ-additive (and hence a premeasure).

(b) μ is σ-subadditive.

(c) μ is lower semicontinuous.

(d) μ is $\varnothing$-continuous.

(e) μ is upper semicontinuous.

Then (a) $\iff$ (b) $\iff$ (c) $\implies$ (d) $\iff$ (e). If μ is finite, then all five statements are equivalent.

Proof.

- (a) $\implies$ (b) (σ-additivity implies σ-subadditivity).
  This follows from theorem 1.9(c).
- (b) $\implies$ (a) (σ-subadditivity implies σ-additivity).
  This follows from theorem 1.9(d).
- (a) $\implies$ (c) (σ-additivity implies lower semicontinuity).
  Let μ be a premeasure and $A \in \mathcal{A}$. Let $A_1, A_2, \ldots \in \mathcal{A}$ such that $A_n \uparrow A$ and let $A_0 = \varnothing$. Then
  $$\mu(A) = \sum_{i=1}^{\infty} \mu(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \sum_{i=1}^{n} \mu(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \mu(A_n).$$
- (c) $\implies$ (a) (lower semicontinuity implies σ-additivity).
  Let $B_1, B_2, \ldots \in \mathcal{A}$ be mutually disjoint and let $B = \biguplus_{n=1}^{\infty} B_n \in \mathcal{A}$. Let $A_n = \biguplus_{i=1}^{n} B_i$ for each $n \in \mathbb{N}$. Then
  $$\mu(B) = \lim_{n\to\infty} \mu(A_n) = \sum_{i=1}^{\infty} \mu(B_i).$$
  Thus μ is σ-additive.
- (d) $\implies$ (e) ($\varnothing$-continuity implies upper semicontinuity).
  Let $A, A_1, A_2, \ldots \in \mathcal{A}$ with $A_n \downarrow A$ and $\mu(A_1) < \infty$. Define $B_n = A_n \setminus A \in \mathcal{A}$ for valid n. Then $B_n \downarrow \varnothing$. Thus
  $$\lim_{n\to\infty} \mu(A_n) - \mu(A) = \lim_{n\to\infty} \mu(B_n) = 0$$
  and the result is proved.
- (e) $\implies$ (d) (upper semicontinuity implies $\varnothing$-continuity).
  This is obvious.
- (c) $\implies$ (d) (lower semicontinuity implies $\varnothing$-continuity).
  Let $A_1, A_2, \ldots \in \mathcal{A}$ with $A_n \downarrow \varnothing$ and $\mu(A_1) < \infty$. Then $A_1 \setminus A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$ and $A_1 \setminus A_n \uparrow A_1$. Thus
  $$\mu(A_1) = \lim_{n\to\infty} \mu(A_1) - \mu(A_n).$$
  Since $\mu(A_1) < \infty$, $\lim_{n\to\infty} A_n = 0$ and the result is proved.
- (d) $\implies$ (c) ($\varnothing$-continuity implies lower semicontinuity) if μ is finite.
  Let $A, A_1, A_2, \ldots \in \mathcal{A}$ with $A_n \uparrow A$. Then $A \setminus A_n \downarrow \varnothing$ and
  $$\lim_{n\to\infty} \mu(A) - \mu(A_n) = \lim_{n\to\infty} \mu(A \setminus A_n) = 0.$$
  The result follows.

■

Definition 1.16 (Measurable spaces)**.**

(i) A pair $(\Omega, \mathcal{A})$ consisting of a nonempty set Ω and a σ-algebra $\mathcal{A} \subseteq 2^{\Omega}$ is called a *measurable space*. The sets $A \in \mathcal{A}$ are called *measurable sets*. If Ω is countable and $\mathcal{A} = 2^{\Omega}$, then the space $(\Omega, 2^{\Omega})$ is called *discrete*.

(ii) A triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space* if $(\Omega, \mathcal{A})$ is a measurable space and μ is a measure on $\mathcal{A}$.

1.3. Measurable Maps

In measure theory, the homomorphisms (structure-preserving maps between objects) are studied as measurable maps.

Definition 1.17 (Measurable map)**.** Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces. A map $X : \Omega \to \Omega'$ is called $\mathcal{A} - \mathcal{A}'$-*measurable* (or just measurable) if

$$X^{-1}(A') \in \mathcal{A} \text{ for any } A' \in \mathcal{A}'.$$

In this case, we write $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$.

Theorem 1.11 (Generated σ-algebra)**.** Let $(\Omega', \mathcal{A}')$ be a measurable space and Ω be a nonempty set. Let $X : \Omega \to \Omega'$ be a map. Then

$$X^{-1}(\mathcal{A}') = \{X^{-1}(A') : A' \in \mathcal{A}'\}$$

is the smallest σ-algebra with respect to which X is measurable. We call $X^{-1}(\mathcal{A}')$ the *σ-algebra generated by X* and denote it as $\sigma(X)$.

Proof. Let X be measurable with respect to some σ-algebra $\mathcal{A}$. Then $X^{-1}(A') \in \mathcal{A}$ for any $A' \in \mathcal{A}'$, that is, $\sigma(X) \subseteq \mathcal{A}$. Let us now prove that $\sigma(X)$ is a σ-algebra by checking each of the axioms in definition 1.3.

1. As $\Omega' \in \mathcal{A}'$ and $X^{-1}(\Omega') = \Omega$, $\Omega \in \sigma(X)$.
2. Let $A \in \sigma(X)$ and $A' \in \mathcal{A}'$ such that $X^{-1}(A') = A$. Then as $\mathcal{A}'$ is closed under complements,

$$\Omega \setminus A = X^{-1}(\Omega') \setminus X^{-1}(A') = X^{-1}(\Omega' \setminus A') \in \sigma(X).$$

   Therefore, $\sigma(X)$ is closed under complements.

3. Let $A_1, A_2 \ldots \in \sigma(X)$ and $A_1', A_2', \ldots \in \mathcal{A}'$ such that $A_i = X^{-1}(A_i')$ for each $i \in \mathbb{N}$. Then as $\mathcal{A}'$ is σ-$\cup$-closed,

$$\bigcup_{i\in\mathbb{N}} A_i = \bigcup_{i\in\mathbb{N}} X^{-1}(A_i') = X^{-1}\left(\bigcup_{i\in\mathbb{N}} A_i'\right) \in \sigma(X)$$

Therefore, $\sigma(X)$ is a σ-algebra. ■

Theorem 1.12. Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces and $X : \Omega \to \Omega'$ be a map. Let $\mathcal{E}' \subseteq \mathcal{A}'$ be a class of sets. Then $\sigma(X^{-1}(\mathcal{E}')) = X^{-1}(\sigma(\mathcal{E}'))$.

Proof. We have that $X^{-1}(\mathcal{E}) \subseteq X^{-1}(\sigma(\mathcal{E})) = \sigma(X^{-1}(\sigma(\mathcal{E})))$. This implies that

$$\sigma(X^{-1}(\mathcal{E})) \subseteq X^{-1}(\sigma(\mathcal{E})).$$

To establish the reverse inclusion, consider

$$\mathcal{A}_0' = \{A' \in \sigma(\mathcal{E}') : X^{-1}(A') \in \sigma(X^{-1}(\mathcal{E}'))\}$$

We shall show that $\mathcal{A}_0'$ is a σ-algebra.

(a) Clearly, $\Omega' \in \mathcal{A}_0'$ as $\Omega \in \sigma(X^{-1}(\mathcal{E}'))$ and $\Omega' \in \sigma(\mathcal{E}')$.

(b) Let $A_0' \in \mathcal{A}_0'$. Then

$$X^{-1}((A_0')^c) = (X^{-1}(A_0'))^c \in \sigma(X^{-1}(\mathcal{E}'))$$

and thus $\mathcal{A_\prime}'$ is closed under complements.

(c) Let $A_1', A_2', \ldots \in \mathcal{A}_0'$. Then

$$X^{-1}\left(\bigcup_{i=1}^{\infty} A_i'\right) = \bigcup_{i=1}^{\infty} X^{-1}\left(A_i'\right) \in \sigma(X^{-1}(\mathcal{E}')).$$

Thus, $\mathcal{A}_0'$ is σ-$\cup$-closed.

Now, note that $\mathcal{E}' \subseteq \mathcal{A}_0'$ and $\mathcal{A}_0' \subseteq \sigma(\mathcal{E}')$. This implies that $\mathcal{A}_0' = \sigma(\mathcal{E}')$, and thus $X^{-1}(\sigma(\mathcal{E}')) \subseteq \sigma(X^{-1}(\mathcal{E}'))$. This proves the result. ■

Corollary 1.13. Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces and $X : \Omega \to \Omega'$ be a map. Let $\mathcal{E}' \subseteq \mathcal{A}'$ be a class of sets. Then X is $\mathcal{A}$-$\sigma(\mathcal{E}')$ measurable if and only if $X^{-1}(\mathcal{E}') \in \mathcal{A}$. If in particular $\sigma(\mathcal{E}') = \mathcal{A}'$, then X is $\mathcal{A} - \mathcal{A}'$-measurable if and only if $X^{-1}(\mathcal{E}') \subseteq \mathcal{A}'$.

Theorem 1.14. Let $(\Omega, \mathcal{A})$, $(\Omega', \mathcal{A}')$, and $(\Omega'', \mathcal{A}'')$ be measurable spaces and let $X : \Omega \to \Omega'$ and $X' : \Omega' \to \Omega''$ be measurable. Then $Y = X' \circ X : \Omega \to \Omega''$ is $\mathcal{A} - \mathcal{A}''$-measurable.

Proof. This is due to the fact that

$$Y^{-1}(\mathcal{A}'') \subseteq X^{-1}((X^{-1})(\mathcal{A}'')) \subseteq X^{-1}(\mathcal{A}') \subseteq \mathcal{A}.$$

■

The above theorem just states that the composition of two measurable maps is measurable.

Theorem 1.15 (Measurability of Continuous Maps)**.** Let (Ω, τ) and (Ω', τ') be topological spaces and let $f : \Omega \to \Omega'$ be a continuous map. Then f is $\mathcal{B}(\Omega) - \mathcal{B}(\Omega')$-measurable.

Proof. As $\mathcal{B}(\Omega') = \sigma(\tau')$, by corollary 1.13 it is enough to show that $f^{-1}(A') \in \sigma(\tau)$ for all $A' \in \tau'$. However, since f is continuous, $f^{-1}(A') \in \tau$ for all $A' \in \tau'$ so the result follows. ■

Theorem 1.16. Let $X_1, X_2, \ldots$ be measurable maps $(\Omega, \mathcal{A}) \to (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$. Then $\inf_{n\in\mathbb{N}} X_n$, $\sup_{n\in\mathbb{N}} X_n$, $\liminf_{n\in\mathbb{N}} X_n$, and $\limsup_{n\in\mathbb{N}} X_n$ are also measurable.

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

$$\left(\inf_{n\in\mathbb{N}} X_n\right)^{-1}([-\infty, x)) = \bigcup_{n=1}^{\infty} (X_n)^{-1}([-\infty, x)) \in \mathcal{A}.$$

The first part of the result follows by corollary 1.13. The proof for $\sup_{n\in\mathbb{N}} X_n$ is similar.
For $\limsup_{n\in\mathbb{N}}$, consider the sequence $(Y_n)_{n\in\mathbb{N}}$ where $Y_n = \sup_{m\geq n} X_m$. Each Y_m is measurable. Then since $\inf_{n\in\mathbb{N}} Y_m$ is measurable, the result follows. ■

Definition 1.18 (Simple Function)**.** Let $(\Omega, \mathcal{A})$ be a measurable space. A map $f : \Omega \to \mathbb{R}$ is called *simple* if there exists some $n \in \mathbb{N}$, mutually disjoint sets $A_1, \ldots, A_n \in \mathcal{A}$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}.$$

Definition 1.19. Let $f, f_1, f_2, \ldots$ be maps $\Omega \to \overline{\mathbb{R}}$ such that

$$f_1(\omega) \leq f_2(\omega) \leq \cdots \text{ and } \lim_{n\to\infty} f_n(\omega) = \omega \text{ for all } \omega \in \Omega.$$

We then write $f_n \uparrow f$. Similarly, we write $f_n \downarrow f$ if $(-f_n) \uparrow (-f)$.

Theorem 1.17. Let $(\Omega, \mathcal{A})$ be a measurable space and let $f : \Omega \to [0, \infty]$ be measurable. Then

(a) There exists a sequence $(f_n)_{n\in\mathbb{N}}$ of non-negative simple functions such that $f_n \uparrow f$.

(b) There are measurable sets $A_1, A_2, \ldots \in \mathcal{A}$ and $\alpha_1, \alpha_2, \ldots \in [0, \infty)$ such that $f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$.

Proof.

(a) For $n \in \mathbb{N}_0$, define

$$f_n = \min\{n, 2^{-n}\lfloor 2^n f \rfloor\}.$$

Each f_n is measurable. (Why?) Since it can take at most $n2^n + 1$ distinct values, each f_n is simple. Clearly, $f_n \uparrow f$.

(b) Let f_n be the same as above. For $n \in \mathbb{N}$ and $i \in [2^n]$, define

$$B_{n,i} = \{\omega : f_n(\omega) - f_{n-1}(\omega) = i2^{-n}\} \text{ and } \beta_{n,i} = i2^{-n}.$$

Then $f_n - f_{n-1} = \sum_{i=1}^{2^n} \beta_{n,i} \mathbb{1}_{B_{n,i}}$. Changing the enumeration from (n,i) to m, we get some $(\alpha_m)_{m\in\mathbb{N}}$ and $(A_m)_{m\in\mathbb{N}}$ such that

$$f = f_0 + \sum_{n=1}^{\infty}(f_n - f_{n-1}) = \sum_{m=1}^{\infty} \alpha_m \mathbb{1}_{A_m}.$$

■

1.4. Outer Measure

Lemma 1.18. Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space and $\mathcal{E} \subseteq \mathcal{A}$ be a π-system that generates $\mathcal{A}$. Assume there exists sequence $\Omega_1, \Omega_2 \ldots \in \mathcal{E}$ such that $\bigcup_{i=1}^{\infty} \Omega_i = \Omega$ and $\mu(\Omega_i) < \infty$ for all $i \in \mathbb{N}$. Then μ is uniquely determined by the values $\mu(E)$, $E \in \mathcal{E}$.
If $\Omega \in \mathcal{A}$ and $\mu(\Omega) = 1$, then the existence of the sequence $(\Omega_n)_{n\in\mathbb{N}}$ is not required.

Proof. Let ν be a σ-finite measure on $(\Omega, \mathcal{A})$ such that $\mu(E) = \nu(E)$ for all $E \in \mathcal{E}$.
Let $E \in \mathcal{E}$ with $\mu(E) < \infty$. Consider

$$\mathcal{D}_E = \{A \in \mathcal{A} : \mu(A \cap E) = \nu(A \cap E)\}.$$

We claim that $\mathcal{D}_E$ is a λ-system. We shall prove this by checking each of the conditions of definition 1.6.

(a) Clearly, $\Omega \in \mathcal{D}_E$.

(b) Let $A, B \in \mathcal{D}_E$ with $B \subseteq A$. Then

$$\begin{aligned}\mu((A \setminus B) \cap E) &= \mu(A \cap E) - \mu(B \cap E) \quad \text{(using theorem 1.9)} \\ &= \nu(A \cap E) - \nu(B \cap E) \\ &= \nu((A \setminus B) \cap E).\end{aligned}$$

That is, $(A \setminus B) \in \mathcal{D}_E$.

(c) Let $A_1, A_2, \ldots \in \mathcal{D}_E$ be mutually disjoint sets. Then

$$\begin{aligned}\mu\left(\left(\biguplus_{i=1}^{\infty} A_i\right) \cap E\right) &= \sum_{i=1}^{\infty} \mu(A_i \cap E) \\ &= \sum_{i=1}^{\infty} \nu(A_i \cap E) \\ &= \nu\left(\left(\biguplus_{i=1}^{\infty} A_i\right) \cap E\right).\end{aligned}$$

Therefore, $\biguplus_{i=1}^{\infty} A_i \in \mathcal{D}_E$ and $\mathcal{D}_E$ is a λ-system.

As $\mathcal{E} \subseteq \mathcal{D}_E$ (Why?), $\delta(\mathcal{E}) \subseteq \mathcal{D}_E$. Since $\mathcal{E}$ is a π-system, theorem 1.8 implies that

$$\mathcal{A} \supseteq \mathcal{D}_E \supseteq \delta(\mathcal{E}) = \sigma(\mathcal{E}) = \mathcal{A}.$$

Hence $\mathcal{D}_E = \mathcal{A}$.
Therefore, $\mu(A \cap E) = \nu(A \cap E)$ for any $A \in \mathcal{A}$ and $E \in \mathcal{E}$ with $\mu(E) < \infty$.
Now, let $\Omega_1, \Omega_2, \ldots \in \mathcal{E}$ be a sequence such that $\bigcup_{i=1}^{\infty} \Omega_i = \Omega$ and $\mu(\Omega_i) < \infty$ for all $i \in \mathbb{N}$. Let $E_0 = \varnothing$ and $E_n = \bigcup_{i=1}^{n} \Omega_i$ for each $n \in \mathbb{N}$. Note that

$$E_n = \biguplus_{i=1}^{n} (E_{i-1}^c \cap \Omega_i).$$

Therefore for any $A \in \mathcal{A}$ and $n \in \mathbb{N}$,

$$\begin{aligned}\mu(A \cap E_n) &= \sum_{i=1}^{n} \mu((A \cap E_{i-1}^c) \cap \Omega_i) \\ &= \sum_{i=1}^{n} \nu((A \cap E_{i-1}^c) \cap \Omega_i) = \nu(A \cap E_n).\end{aligned}$$

Now, since $E_n \uparrow \Omega$ and μ, ν are lower semicontinuous (by theorem 1.10),

$$\begin{aligned}\mu(A) &= \lim_{n\to\infty} \mu(A \cap E_n) \\ &= \lim_{n\to\infty} \nu(A \cap E_n) = \nu(A)\end{aligned}$$

This proves the result.

The second part of the theorem is trivial as $\mathcal{E} \cup \{\Omega\}$ is a π-system that generates $\mathcal{A}$. Hence one can choose the constant sequence $E_n = \Omega, n \in \mathbb{N}$. ■

Definition 1.20 (Outer Measure)**.** A function $\mu^* : 2^\Omega \to [0, \infty]$ is called an *outer measure* if

(i) $\mu^*(\varnothing) = 0$,

(ii) μ^* is monotone, and

(iii) μ^* is σ-subadditive.

Lemma 1.19. Let $\mathcal{A} \subseteq 2^\Omega$ be an arbitrary class of sets with $\varnothing \in \mathcal{A}$ and let μ be a nonnegative set function on $\mathcal{A}$ with $\mu(\varnothing) = 0$. For $A \subseteq \Omega$, define the set of countable coverings $\mathcal{F}$ with sets $F \in \mathcal{A}$

$$\mathcal{U}(A) = \left\{ \mathcal{F} \subseteq \mathcal{A} : \mathcal{F} \text{ is countable and } A \subseteq \bigcup_{F\in\mathcal{F}} F \right\}.$$

Define

$$\mu^*(A) = \inf \left\{ \sum_{F\in\mathcal{F}} \mu(F) : \mathcal{F} \in \mathcal{U}(A) \right\}$$

where $\inf \varnothing = \infty$. Then μ^* is an outer measure. If μ is σ-subadditive then $\mu^*(A) = \mu(A)$ for all $A \in \mathcal{A}$.

Proof. Let us check each of the three conditions in the definition of an outer measure.

(a) Since $\varnothing \in \mathcal{A}$, we have $\{\varnothing\} \in \mathcal{U}(\varnothing)$ and hence $\mu(\varnothing) = 0$.

(b) If $A \subseteq B$, then $\mathcal{U}(A) \subseteq \mathcal{U}(B)$, and hence $\mu^*(A) \leq \mu^*(B)$.

(c) Let $A, A_1, A_2, \ldots \subseteq \Omega$ such that $A \subseteq \bigcup_{i=1}^\infty A_i$. We claim that $\mu^*(A) \leq \sum_{i=1}^\infty \mu^*(A_i)$.

Without loss of generality, assume that $\mu^*(A_i) < \infty$ and hence $\mathcal{U}(A_i) \neq \varnothing$ for all $i \in \mathbb{N}$. Fix some $\varepsilon > 0$. Now, for every $n \in \mathbb{N}$, we may choose a covering $\mathcal{F}_n \in \mathcal{U}(A_n)$ such that

$$\sum_{F\in\mathcal{F}_n} \mu(F) \leq \mu^*(A_n) + \varepsilon 2^{-n}.$$

Then let $\mathcal{F} = \bigcup_{n=1}^\infty \mathcal{F}_n \in \mathcal{U}(A)$.

$$\mu^*(A) \leq \sum_{F\in\mathcal{F}} \mu(F) \leq \sum_{n=1}^\infty \sum_{F\in\mathcal{F}_n} \mu(F) \leq \sum_{n=1}^\infty \mu^*(A_n) + \varepsilon.$$

This proves the first part of the result.

To prove the next part of the result, first note that since $\{A\} \in \mathcal{U}(A)$, we have $\mu^*(A) \leq \mu(A)$. If μ is σ-subadditive, then for any $\mathcal{F} \in \mathcal{U}(A)$,

$$\sum_{F \in \mathcal{F}} \mu(F) \geq \mu(A).$$

It follows that $\mu^*(A) \geq \mu(A)$. ■

Definition 1.21 (μ^*-measurable sets)**.** Let μ^* be an outer measure. A set $A \in 2^\Omega$ is called μ^*-*measurable* if

$$\mu^*(A \cap E) + \mu^*(A^c \cap E) = \mu^*(E) \text{ for any } E \in 2^\Omega.$$

We write $\mathcal{M}(\mu^*) = \{A \subseteq \Omega : A \text{ is } \mu^*\text{-measurable}\}$.

Lemma 1.20. $A \in \mathcal{M}(\mu^*)$ if and only if

$$\mu^*(A \cap E) + \mu^*(A^c \cap E) \leq \mu^*(E) \text{ for any } E \in 2^\Omega.$$

Proof. As μ^* is subadditive, we trivially have

$$\mu^*(A \cap E) + \mu^*(A^c \cap E) \geq \mu^*(E) \text{ for any } E \in 2^\Omega.$$

The result follows. ■

Lemma 1.21. $\mathcal{M}(\mu^*)$ is an algebra.

Proof. We shall check the conditions given in the definition of an algebra definition 1.2.

(a) We clearly have $\Omega \in \mathcal{M}(\mu^*)$.

(b) By definition, $\mathcal{M}(\mu^*)$ is closed under complements.

(c) We must check that $\mathcal{M}(\mu^*)$ is closed under intersections. Let $A, B \in \mathcal{M}(\mu^*)$ and $E \subseteq \Omega$. Then

$$\begin{aligned}
\mu^*((A \cap B) \cap E) + \mu^*((A \cap B)^c \cap E) &= \mu^*((A \cap B) \cap E) \\
&\quad + \mu^*((A \cap B^c \cap E) \cup (A^c \cap B \cap E) \cup (A^c \cap B^c \cap E)) \\
&\leq \mu^*(A \cap (B \cap E)) + \mu^*(A \cap (B^c \cap E)) \\
&\quad + \mu^*(A^c \cap (B \cap E)) + \mu^*(A^c \cap (B^c \cap E)) \\
&= \mu^*(B \cap E) + \mu^*(B^c \cap E) \quad (\text{since } A \in \mathcal{M}(\mu^*)) \\
&= \mu^*(E). \quad (\text{since } B \in \mathcal{M}(\mu^*))
\end{aligned}$$

This proves the result. ■

Lemma 1.22. An outer measure μ^* is σ-additive on $\mathcal{M}(\mu^*)$.

Proof. Let $A, B \in \mathcal{M}(\mu^*)$ with $A \cap B = \varnothing$. Then

$$\begin{aligned}
\mu^*(A \cup B) &= \mu^*(A \cap (A \cup B)) + \mu^*(A^c \cap (A \cup B)) \\
&= \mu^*(A) + \mu^*(B).
\end{aligned}$$

That is, μ^* is additive (and thus a content) on $\mathcal{M}(\mu^*)$. Since μ^* is σ-subadditive, theorem 1.10 gives the required result. ■

Lemma 1.23. If μ^* is an outer measure, $\mathcal{M}(\mu^*)$ is a σ-algebra.

Proof. We have already shown that $\mathcal{M}(\mu^*)$ is an algebra (and thus a π-system). Using theorem 1.7, it is sufficient to show that $\mathcal{M}(\mu^*)$ is a λ-system.
Let $A_1, A_2, \ldots \in \mathcal{M}(\mu^*)$ be mutually disjoint sets and let $A = \biguplus_{i=1}^{\infty} A_i$. Further, for each $n \in \mathbb{N}$, let $B_n = \biguplus_{i=1}^{n} A_i$. We must show that $M \in \mathcal{M}(\mu^*)$.
For any E and valid $n \in \mathbb{N}$, we have

$$\begin{aligned}\mu^*(E \cap B_{n+1}) &= \mu^*((E \cap B_{n+1}) \cap B_n) + \mu^*((E \cap B_{n+1}) \cap B_n^c) \\ &= \mu^*(E \cap B_n) + \mu^*(E \cap A_{n+1}).\end{aligned}$$

By a simple induction, it follows that

$$\mu(E \cap B_n) = \sum_{i=1}^{n} \mu^*(E \cap A_i).$$

Since μ^* is monotonic, we have

$$\begin{aligned}\mu^*(E) &= \mu^*(E \cap B_n) + \mu^*(E \cap B_n^c) \\ &\geq \mu^*(E \cap B_n) + \mu^*(E \cap A^c) \\ &= \sum_{i=1}^{n} \mu^*(E \cap A_i) + \mu^*(E \cap A^c).\end{aligned}$$

Letting $n \to \infty$ and using the fact that μ^* is σ-subadditive, we have

$$\begin{aligned}\mu^*(E) &\geq \sum_{i=1}^{\infty} \mu^*(E \cap A_i) + \mu^*(E \cap A^c) \\ &\geq \mu^*(E \cap A) + \mu^*(E \cap A^c)\end{aligned}$$

Therefore, $A \in \mathcal{M}(\mu^*)$ and this completes the proof. ■

1.5. The Approximation and Extension Theorems

Theorem 1.24 (Approximation Theorem for Measures)**.** Let $\mathcal{A} \subseteq 2^{\Omega}$ be a semiring and let μ be a measure on $\sigma(\mathcal{A})$ that is σ-finite on $\mathcal{A}$. For any $A \in \sigma(\mathcal{A})$ with $\mu(\mathcal{A}) < \infty$ and any $\varepsilon > 0$, there exists $n \in \mathbb{N}$ and mutually disjoint sets $A_1, A_2, \ldots, A_n \in \mathcal{A}$ such that $\mu\left(A \triangle \bigcup_{i=1}^{n} A_n\right) < \varepsilon$.

Proof. Consider the outer measure μ^* as defined in lemma 1.19. Note that by lemma 1.19 and lemma 1.18, μ and μ^* are equal on $\sigma(\mathcal{A})$. By the definition of μ^*, for any $A \in \mathcal{A}$, there exists a covering $B_1, B_2, \ldots \in \mathcal{A}$ of A such that

$$\mu(A) \geq \sum_{i=1}^{\infty} \mu(B_i) - \varepsilon/2.$$

Since $\mu(A) < \infty$, there exists some $n \in \mathbb{N}$ such that $\sum_{i=n+1}^{\infty} \mu(B_i) < \varepsilon/2$. Now, let $D = \bigcup_{i=1}^{n} B_i$ and $E = \bigcup_{i=n+1}^{\infty} B_i$. We have

$$\begin{aligned}A \triangle D &= (D \setminus A) \cup (A \setminus D) \\ &\subseteq (D \setminus A) \cup (A \setminus (D \cup E)) \cup E \\ &\subseteq (A \triangle (D \cup E)) \cup E.\end{aligned}$$

This together with the fact that $A \subseteq \bigcup_{i=1}^{\infty} B_i$ implies that

$$\begin{aligned}\mu(A \triangle D) &\leq \mu(A \triangle (D \cup E)) + \mu(E) \\ &\leq \mu\left(\bigcup_{i=1}^{\infty} B_i\right) - \mu(A) + \frac{\varepsilon}{2} \\ &\leq \varepsilon.\end{aligned}$$

Now define $A_1 = B_1$ and for each $i \geq 2$, $A_i = B_i \setminus \bigcup_{j=1}^{i=1} B_j$. By definition, $A_1, A_2 \ldots$ are mutually disjoint. This proves the result.

■

The following theorem allows us to "extend" measures from a semiring to the σ-algebra generated by it. This allows us to define measures over an entire σ-algebra by defining its values over just a semiring that generates it.

Theorem 1.25 (Measure Extension Theorem)**.** Let $\mathcal{A}$ be a semiring and let $\mu : \mathcal{A} \to [0, \infty]$ be an additive, σ-subadditive and σ-finite set function with $\mu(\varnothing) = 0$. Then there is a unique σ-finite measure $\tilde{\mu} : \sigma(\mathcal{A}) \to [0, \infty]$ such that $\tilde{\mu}(A) = \mu(A)$ for all $A \in \mathcal{A}$.

Proof. Since $\mathcal{A}$ is a π-system, if such a $\tilde{\mu}$ exists, it is uniquely defined due to lemma 1.18.
We shall explicitly construct a function that satisfies the given conditions. In order to do so, define as in lemma 1.19

$$\mu^*(A) = \inf\left\{\sum_{F\in\mathcal{F}} \mu(F) : \mathcal{F} \in \mathcal{U}(A)\right\} \text{ for any } A \subseteq \Omega.$$

By lemma 1.19, μ^* is an outer measure and $\mu^*(A) = \mu(A)$ for any $A \in \mathcal{A}$.

We first claim that $\mathcal{A} \subseteq \mathcal{M}(\mu^*)$.
To prove this, let $A \in \mathcal{A}$ and $E \subseteq \Omega$ with $\mu^*(E) < \infty$. Fix some $\varepsilon > 0$. Then by the definition of μ^*, there exists a sequence $E_1, E_2, \ldots \in \mathcal{A}$ such that

$$E \subseteq \bigcup_{i=1}^{\infty} E_i \text{ and } \sum_{i=1}^{\infty} \mu(E_i) \leq \mu^*(E) + \varepsilon.$$

For each n, define $B_n = E_n \cap A$. Since $\mathcal{A}$ is a semiring, there exists for each n some $m_n \in \mathbb{N}$ and mutually disjoint sets $C_{n,1}, C_{n,2}, \ldots, C_{n,m_n}$ such that

$$E_n \setminus A = E_n \setminus B_n = \biguplus_{i=1}^{m_n} C_{n,i}$$

Then we have that

$$\begin{aligned} E \cap A &\subseteq \bigcup_{n=1}^{\infty} B_n, \\ E \cap A^c &\subseteq \bigcup_{n=1}^{\infty} \biguplus_{i=1}^{m_n} C_{n,i}, \text{ and} \\ E_n &= B_n \uplus \biguplus_{i=1}^{m_n} C_{n,i}. \end{aligned}$$

This implies that

$$\begin{aligned} \mu^*(E \cap A) + \mu^*(E \cap A^c) &\leq \sum_{n=1}^{\infty} \mu(B_n) + \sum_{n=1}^{\infty}\sum_{i=1}^{m_n} \mu(C_{n,i}) \quad \text{(since } \mu \text{ is } \sigma\text{-subadditive)} \\ &= \sum_{n=1}^{\infty} \left(\mu(B_n) + \sum_{i=1}^{m_n} \mu(C_{n,i})\right) \\ &= \sum_{n=1}^{\infty} \mu(E_n) \quad \text{(since } \mu \text{ is additive)} \\ &\leq \mu^*(E) + \varepsilon. \end{aligned}$$

lemma 1.20 implies that $A \in \mathcal{M}(\mu^*)$, that is, $\mathcal{A} \subseteq \mathcal{M}(\mu^*)$. This in turn in implies that $\sigma(\mathcal{A}) \subseteq \mathcal{M}(\mu^*)$. Define the required function by $\tilde{\mu} : \sigma(\mathcal{A}) \to [0, \infty]$, $A \mapsto \mu^*(A)$. By lemma 1.22, $\tilde{\mu}$ is σ-additive. Since μ is σ-finite, $\tilde{\mu}$ is σ-finite as well. This proves the result. ∎

1.6. Important Examples of Measures

Now that we have the Measure Extension Theorem, we may introduce the Lebesgue-Stieltjes measure, a very useful measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which is given as follows.

Definition 1.22 (Lebesgue-Stieltjes Measure)**.** Let $F : \mathbb{R} \to \mathbb{R}$ be monotone increasing and right continuous. The measure μ_F on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$\mu_F((a, b]) = F(b) - F(a) \text{ for all } a, b \in \mathbb{R} \text{ such that } a < b$$

is called the *Lebesgue-Stieltjes measure* with distribution function F.

The Lebesgue-Stieltjes measure is well-defined due to the Measure Extension Theorem theorem 1.25.

To see this more clearly, let $\mathcal{A} = \{(a, b] : a, b \in \mathbb{R} \text{ and } a \leq b\}$. It may be checked that $\mathcal{A}$ is a semiring. Further, $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$. Now, define the function $\tilde{\mu}_F : \mathcal{A} \to [0, \infty)$ by $(a, b] \mapsto F(b) - F(a)$. Clearly $\tilde{\mu}_F(\varnothing) = 0$ and the function is additive. It remains to check that $\tilde{\mu}_F$ is σ-subadditive.

Let $(a, b], (a_1, b_1], (a_2, b_2], \ldots \in \mathcal{A}$ such that $(a, b] \subseteq \bigcup_{i=1}^{\infty}(a_i, b_i]$. Fix some $\varepsilon > 0$ and choose $a_\varepsilon \in (a, b)$ such that

$$F(a_\varepsilon) - F(a) < \varepsilon/2 \implies \tilde{\mu}_F((a, b]) - \tilde{\mu}_F((a_\varepsilon, b]) < \varepsilon/2.$$

It is possible to choose such an ε due to the right continuity of F. Also, for any $k \in \mathbb{N}$, choose $b_{k,\varepsilon}$ such that

$$F(b_{k,\varepsilon}) - F(b_k) < \varepsilon 2^{-k-1} \implies \tilde{\mu}_F((a_k, b_{k,\varepsilon}]) - \tilde{\mu}_F((a_k, b_k]) < \varepsilon 2^{-k-1}.$$

We now have

$$[a_\varepsilon, b] \subseteq (a, b] \subseteq \bigcup_{i=1}^{\infty}(a_k, b_k] \subseteq \bigcup_{k=1}^{\infty}(a_k, b_{k,\varepsilon}]$$

Due to the compactness of $[a_\varepsilon, b]$, there then exists some $k_0 \in \mathbb{N}$ such that

$$(a_\varepsilon, b] \subseteq \bigcup_{k=1}^{k_0}(a_k, b_{k,\varepsilon}].$$

This implies that

$$\begin{aligned}
\tilde{\mu}_F((a, b]) &\leq \frac{\varepsilon}{2} + \tilde{\mu}_F((a, b]) \\
&\leq \frac{\varepsilon}{2} + \sum_{k=1}^{k_0} \tilde{\mu}_F((a_k, b_{k,\varepsilon}]) \\
&\leq \frac{\varepsilon}{2} + \sum_{k=1}^{k_0} \left(\tilde{\mu}_F((a_k, b_k]) + \varepsilon 2^{-k-1}\right) \\
&\leq \varepsilon + \sum_{k=1}^{\infty} \tilde{\mu}_F((a_k, b_k])
\end{aligned}$$

As this is true for any choice of ε, $\tilde{\mu}_F$ is σ-subadditive.

Then the extension of $\tilde{\mu}_F$ uniquely to a σ-finite measure is guaranteed by theorem 1.25. This measure is known as the Lebesgue-Stieltjes measure.

The measure that results when the function F is equal to the identity function is referred to the *Lebesgue measure* on $\mathbb{R}^1$. Similar to this, we can define the Lebesgue measure in general as follows.

Definition 1.23 (Lebesgue Measure)**.** There exists a unique measure λ^n on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that for all $a, b \in \mathbb{R}^n$ with $a < b$,

$$\lambda^n((a, b]) = \prod_{i=1}^{n}(b_i - a_i).$$

λ^n is called the *Lebesgue measure* on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ or the *Lebesgue-Borel measure*.

Let E be a finite nonempty set and $\Omega = E^{\mathbb{N}}$. If $\omega_1, \omega_2, \ldots, \omega_n \in E$, we define the following.

$$[\omega_1, \omega_2, \ldots, \omega_n] = \{\omega' \in \Omega : \omega'_i = \omega_i \text{ for } i \in [n]\}.$$

This represents the set of all sequences whose first n elements are $\omega_1, \omega_2, \ldots, \omega_n$.

Theorem 1.26 (Finite Products of Measures)**.** Let $n \in \mathbb{N}$ and $\mu_1, \mu_2, \ldots, \mu_n$ be Lebesgue-Stieltjes measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then there exists a unique σ-finite measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that for all $a, b \in \mathbb{R}^n$ with $a < b$,

$$\mu((a, b]) = \prod_{i=1}^{n} \mu_i((a_i, b_i])$$

We call μ the *product measure of* $\mu_1, \mu_2, \ldots, \mu_n$ and denote it by $\bigotimes_{i=1}^{n} \mu_i$.

The proof of the above is similar to that of theorem 1.25. We choose intervals $(a, b_\varepsilon]$ and so on such that $\mu((a, b_\varepsilon]) < \mu((a, b]) + \varepsilon$. Such b_ε exists due to the right continuity of each of the F_is corresponding to each of the μ_is.

§2. Introduction to Probability

2.1. Basic Definitions

Definition 2.1 (Probability Space)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. If $\mu(\Omega) = 1$, then $(\Omega, \mathcal{A}, \mu)$ is called a *probability space* and μ is called a *probability measure*.

In the above definition, Ω is called the *sample space*, $\mathcal{A}$ is called the *event space* (and its elements are called *events*), and μ is called the *probability function*.

While the above definition may appear to be completely unrelated to the intuitive notion of probability we have, the following example should hopefully make the meanings clear.

Consider a coin toss. The sample space Ω has two elements, H (for heads) and T (for tails). The event space $\mathcal{A}$ then has four elements: $\varnothing$, $\{H\}$, $\{T\}$, and $\{H, T\}$. Each of these events have associated probabilities $0, \frac{1}{2}, \frac{1}{2}$, and 1 respectively. Note that $\{H, T\}$ represents the event that either a heads or a tails occurs.

The requirement of the event space to be a σ-algebra is quite natural as well. The event Ω corresponds to saying that *something* happens, which must occur with certainty. Closedness under complements means that if we have an event A, we should have another event that corresponds to A not occurring. Finally, σ-$\cup$-closedness corresponds to the occurrence of at least one of the events we are taking the union of.

- *Uniform distribution.* Let Ω be a finite nonempty set. Consider the function $\mu : 2^\Omega \to [0, 1]$ given by

  $$\mu(A) = \frac{|A|}{|\Omega|} \text{ for each } A \subseteq \Omega.$$

  This defines a probability measure on 2^Ω. This function μ is called the *uniform distribution on* Ω and is denoted $\mathcal{U}_\Omega$. As the reader might expect, it represents the case where any element of Ω is equally likely to occur. The resulting probability space $(\Omega, 2^\Omega, \mathcal{U}_\Omega)$ is called a *Laplace space*.

- *Dirac measure.* Let $\omega \in \Omega$ and $\delta_\omega(A) = \mathbb{1}(\{\omega\})$. Then δ_ω is a probability measure on any σ-algebra $\mathcal{A} \subseteq 2^\Omega$. δ_ω is called the *Dirac measure* for the point ω.

  The Dirac measure is useful in constructing discrete probability distributions.

Let Ω be a countable non-empty set and $\mathcal{A} = 2^\Omega$. Further let $(p_\omega)_{\omega\in\Omega}$ be non-negative numbers. The map given by $A \mapsto \mu(A) = \sum_{\omega\in A} p_\omega$ defines a σ-finite measure on 2^Ω. $p = (p_\omega)_{\omega\in\Omega}$ is called the *weight function* of μ. p_ω is called the *weight of* μ *at point* ω.
In the case where $\sum_{\omega\in\Omega} p_\omega = 1$, μ is a probability measure. Then the vector $(p_\omega)_{\omega\in\Omega}$ is called a *probability vector*.

Given any probability space, we typically use the Pr or $\mathbf{P}$ symbol to denote the universal object of a probability measure, and the probabilities $\Pr[\cdot]$ or $\mathbf{P}[\cdot]$ are always written in (square) brackets.

Definition 2.2 (Probability Distribution Function)**.** A right continuous monotonically increasing function $F : \mathbb{R} \to [0, 1]$ such that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$ is called a *(proper) probability distribution function*, often abbreviated as *p.d.f.* If we instead have $\lim_{x\to\infty} F(x) \leq 1$, F is called a *(possibly) defective p.d.f.* If μ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then the function F_μ given by $x \mapsto \mu((-\infty, x])$ is called the *distribution function* of μ.

A probability measure is uniquely determined by its distribution function. (Why?)

Definition 2.3 (Random Variable)**.** Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, $(\Omega', \mathcal{A}')$ a measurable space, and $X : \Omega \to \Omega'$ be measurable. Then X is called a *random variable* with values in $(\Omega', \mathcal{A}')$. If $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then X is called a *real random variable*. For $A' \in \mathcal{A}'$, we often denote

$$\mathbf{P}[X^{-1}(A')] \text{ as } \mathbf{P}[X \in A'] \text{ and } X^{-1}(A') \text{ as } \{X \in A'\}.$$

In particular, we let $\{X \geq 0\} = X^{-1}([0, \infty))$ and define $\{X \leq b\}$ and other terms similarly.

As we shall primarily deal with real random variables in our study of probability, we often drop the "real" and refer to them as just random variables.

Definition 2.4. Let X be a random variable with underlying probability space $(\Omega, \mathcal{A}, \mathbf{P})$.

(i) The probability measure $\mathbf{P}_X = \mathbf{P} \circ X^{-1}$ is called the *distribution* of X.

(ii) For a real random variable X, the map F_X given by $x \mapsto \mathbf{P}[X \leq x]$ is called the *distribution function* of P_X (or X). If $\mu = \mathbf{P}_X$, we write $X \sim \mu$ and say that X has distribution μ.

(iii) A family $(X_i)_{i \in I}$ of random variables is called *identically distributed* if $\mathbf{P}_{X_i} = \mathbf{P}_{X_j}$ for all $i, j \in I$. We write $X \overset{\mathcal{D}}{=} Y$ if $\mathbf{P}_X = \mathbf{P}_Y$ ($\mathcal{D}$ for *distribution*).

The distribution of a random variable essentially gives us a probability corresponding to each element of Ω'. Two random variables being identically distributed means that they are essentially the same, in the sense that a die labelled from 1 through 6 is the same as one labelled from a through f.

Theorem 2.1. For any p.d.f. F, there exists a real random variable X with $F_X = F$.

Proof. We shall explicitly construct a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and random variable $X : \Omega \to \mathbb{R}$ such that $F_X = F$. One choice that might come to mind is to take $(\Omega, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $X : \mathbb{R} \to \mathbb{R}$ as the identity function, and $\mathbf{P}$ as the Lebesgue-Stieltjes measure with distribution function F.

While this choice of ours works, let us attempt to construct another more "standard" choice that is perhaps more enlightening. Let $\Omega = (0,1), \mathcal{A} = \mathcal{B}(\mathbb{R})|_\Omega$ and $\mathbf{P}$ be the Lebesgue measure on $(\Omega, \mathcal{A})$. This is standard in the sense that given any F, we construct a random variable over the same probability space. Define the left continuous inverse of F as

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\} \text{ for } t \in (0,1).$$

Note that $F^{-1}(t) \leq x$ if and only if $F(x) \geq t$. In particular,

$$\{t : F^{-1}(t) \leq x\} = (0, F(x)] \cap (0,1)$$

and so $F^{-1} : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable. Thus

$$\mathbf{P}\left[\{t : F^{-1}(t) \leq x\}\right] = F(x).$$

This implies that F^{-1} is the random variable we wish to construct. ■

Note that the above implies that there is a bijection between probability distribution functions and distribution functions corresponding to random variables.

Definition 2.5. If a distribution $F : \mathbb{R}^n \to [0,1]$ is of the form

$$F(x) = \int_{-\infty}^{x_1} \mathrm{d}t_1 \int_{-\infty}^{x_2} \mathrm{d}t_2 \cdots \int_{-\infty}^{x_n} \mathrm{d}t_n \, f(t_1, t_2, \ldots, t_n) \text{ for } x \in \mathbb{R}^n$$

for some integrable function $f : \mathbb{R}^n \to [0, \infty)$, then f is called the *density of the distribution*.

It is often easier to describe continuous probability distributions either in terms of their density or the corresponding probability distribution function. For example, if a random variable corresponds to picking a number uniformly randomly from $[0,1]$, then its density function is uniformly equal to 1.

2.2. Important Examples of Random Variables

We now give several important examples of random variables that we shall encounter several times in our study of probability.

1. *Bernoulli Distribution.*

   Let $p \in [0,1]$ and $\mathbf{P}[X=1] = p$, $P[X=0] = 1-p$. Then $\mathbf{P}_X$ is called the *Bernoulli distribution with parameter* p and is denoted Ber_p. More formally,

   $$\text{Ber}_p = (1-p)\delta_0 + p\delta_1.$$

   Its distribution function is

   $$F_X(x) = \begin{cases} 0, & x < 0 \\ 1-p, & x \in [0,1) \\ 1, & x \geq 1 \end{cases}$$

   Note that the above can be likened to the outcome of a weighted coin, with heads and tails corresponding to 0 and 1.

   The distribution $\mathbf{P}_Y$ of $Y = 2X - 1$ is called the *Rademacher distribution with parameter* p. More formally,

   $$\text{Rad}_p = (1-p)\delta_{-1} + p\delta_1.$$

   $\text{Rad}_{1/2}$ is simply called the Rademacher distribution.

2. *Binomial Distribution.*

   Let $p \in [0,1]$ and $n \in \mathbb{N}$. Let $X : \Omega \to [n]_0$ be such that for each valid k,

   $$\mathbf{P}[X=k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

   Then $\mathbf{P}_X$ is called the *binomial distribution with parameters* n *and* p and is denoted $b_{n,p}$. More formally,

   $$b_{n,p} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k.$$

3. *Geometric Distribution.*

   Let $p \in (0,1]$ and $X : \Omega \to \mathbb{N}_0$ be such that for each $n \in \mathbb{N}_0$,

   $$\mathbf{P}[X=n] = p(1-p)^n.$$

   Then $\mathbf{P}_X$ is called the *geometric distribution with parameter* p and is denoted γ_p or $b^-_{1,p}$. More formally,

   $$\gamma_p = \sum_{n=0}^{\infty} p(1-p)^n \delta_n.$$

   The geometric distribution γ_p represents the waiting time for a success in a series of independent random experiments, each of which succeeds with a probability p.

4. *Negative Binomial Distribution.*

   Let $r > 0$ and $p \in (0,1]$. We denote by

   $$b^-_{r,p} = \sum_{k=0}^{\infty} \binom{-r}{k} (-1)^k p^r (1-p)^k \delta_k$$

   the *negative binomial distribution* or *Pascal distribution* with parameters r and p. Note that r need not be an integer. The negative binomial distribution $b^-_{r,p}$ represents the waiting time for the rth success in a series of independent random experiments, each of which succeeds with a probability p. Based on this intuition, we expect there to be some relation between the geometric distribution and the negative binomial distribution. We shall explain this relation later in the notes.

5. *Poisson Distribution.*

   Let $\lambda \in [0, \infty)$ and $X : \Omega \to \mathbb{N}_0$ be such that for each $n \in \mathbb{N}_0$,

   $$P[X = n] = e^{-\lambda} \frac{\lambda^n}{n!}.$$

   Then $\mathbf{P}_X = \text{Poi}_\lambda$ is called the *Poisson distribution with parameter* λ.

6. *Hypergeometric Distribution.*

   Consider a basket with $B \in \mathbb{N}$ black balls and $W \in \mathbb{N}$ white balls. If we draw $n \in \mathbb{N}$ balls from the basket, some simple combinatorics shows that the probability of drawing (exactly) $b \in [n]_0$ black balls is given by the *hypergeometric distribution with parameters* B, W, n:

   $$\text{Hyp}_{B,W;n}(\{b\}) = \frac{\binom{B}{b}\binom{W}{n-b}}{\binom{B+W}{n}}.$$

   In general, if we have k colors with B_i balls of colour i for each i, the probability of drawing exactly b_i balls of colour i for each i is given by the *generalised hypergeometric distribution*:

   $$\text{Hyp}_{B_1,B_2,\ldots,B_k;n}(\{(b_1, b_2, \ldots, b_k)\}) = \frac{\binom{B_1}{b_1}\binom{B_2}{b_2}\cdots\binom{B_k}{b_k}}{\binom{B_1+B_2+\cdots+B_k}{n}}$$

   where $n = b_1 + b_2 + \cdots + b_k$.

7. *Gaussian Normal Distribution.*

   Let $\mu \in \mathbb{R}, \sigma^2 > 0$. Let X be a real random variable such that for $x \in \mathbb{R}$,

   $$\mathbf{P}[X \leq x] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \mathrm{d}t$$

   Then $\mathbf{P}_X$ is called the *Gaussian normal distribution* (or just *normal distribution*) *with parameters* μ *and* σ^2 and is denoted $\mathcal{N}_{\mu,\sigma^2}$. In particular, $\mathcal{N}_{0,1}$ is the standard normal distribution.

8. *Exponential Distribution.*

   Let $\theta > 0$ and X be a nonnegative random variable such that for each $x \geq 0$,

   $$\mathbf{P}[X \leq x] = \mathbf{P}\left[X \in [0, x]\right] = \int_0^x \theta e^{-\theta t} \, \mathrm{d}t.$$

   Then $\mathbf{P}_X$ is called the *exponential distribution with parameter* θ and is denoted $\exp_\theta$.

2.3. The Product Measure

Let E be a finite set and $\Omega = E^{\mathbb{N}}$. Let $(p_e)_{e \in E}$ be a probability vector. Define

$$\mathcal{A} = \{[\omega_1, \ldots, \omega_n] : \omega_1, \ldots, \omega_n \text{ and } n \in \mathbb{N}\}$$

and a content μ on $\mathcal{A}$ by

$$\mu([\omega_1, \omega_2, \ldots, \omega_n]) = \prod_{i=1}^{n} p_{\omega_i}$$

We wish to extend μ to a measure on $\sigma(\mathcal{A})$. Similar to how we proved the existence of the Lebesgue-Stieltjes measure definition 1.22, we use a compactness argument to show that μ is σ-subadditive.
Let $A, A_1, A_2, \ldots \in \mathcal{A}$ such that $A \subseteq \bigcup_{i=1}^{\infty} A_i$. We claim that there exists $n \in \mathbb{N}$ such that $A \subseteq \bigcup_{i=1}^{n} A_i$.
For each $n \in \mathbb{N}$, let $B_n = A \setminus \bigcup_{i=1}^{n} A_i$. We assume that $B_n \neq \varnothing$ for all $n \in \mathbb{N}$ and prove the required by contradiction.

Due to the pigeonhole principle, there exists some $\omega_1 \in E$ such that $[\omega_1] \cap B_n \neq \varnothing$ for infinitely many $n \in \mathbb{N}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, we have that

$$[\omega_1] \cap B_n \neq \varnothing \text{ for all } n \in \mathbb{N}.$$

Similarly, there exist $\omega_2, \omega_3, \ldots \in E$ such that

$$[\omega_1, \ldots, \omega_k] \cap B_n \neq \varnothing \text{ for all } k, n \in \mathbb{N}.$$

Each B_n is a disjoint union of sets $C_{n,1}, \ldots, C_{n,m_n} \in \mathcal{A}$. Thus for each $n \in \mathbb{N}$, there is some $i_n \in [m_n]$ such that

$$[\omega_1, \omega_2, \ldots, \omega_k] \cap C_{n,i_n} \neq \varnothing \text{ for infinitely many } k \in \mathbb{N}.$$

As $[\omega_1] \supseteq [\omega_1, \omega_2] \supseteq \cdots$, this implies that

$$[\omega_1, \omega_2, \ldots, \omega_k] \cap C_{n,i_n} \neq \varnothing \text{ for all } k \in \mathbb{N}$$

As $C_{n,i_n} \in \mathcal{A}$, for fixed n and large k ($k \geq m_n$), we have

$$[\omega_1, \omega_2, \ldots, \omega_k] \subseteq C_{n,i_n}.$$

This implies that $\omega = (\omega_1, \omega_2, \ldots) \in C_{n,i_n} \subseteq B_n$. This in turn implies that $\bigcap_{i=0}^{\infty} B_i \neq \varnothing$, which yields a contradiction. Therefore, $A \subseteq \bigcup_{i=1}^{n} A_n$ for some $n \in \mathbb{N}$. Since μ is known to be (finite) subadditive, we have

$$\mu(A) \leq \sum_{i=1}^{n} \mu(A_i) \leq \sum_{i=1}^{\infty} \mu(A_i),$$

which is the required result.

Definition 2.6 (Product Measure)**.** Let E be a finite nonempty set and $\Omega = E^{\mathbb{N}}$. Let $(p_e)_{e \in E}$ be a probability vector. There then is a unique probability measure μ on $\sigma(\mathcal{A}) = \mathcal{B}(\Omega)$ (where $\mathcal{A}$ is defined as above) such that

$$\mu([\omega_1, \omega_2, \ldots, \omega_n]) = \prod_{i=1}^{n} p_{\omega_i} \text{ for all } \omega_i \in E \text{ and } n \in \mathbb{N}.$$

μ is called the *product measure* or *Bernoulli measure* on Ω with weights $(p_e)_{e \in E}$ and is denoted by $\left(\sum_{e \in E} p_e \delta_e\right)^{\otimes \mathbb{N}}$. The σ-algebra $\sigma(\mathcal{A})$ is called the *product σ-algebra on Ω* and is denoted by $(2^E)^{\otimes \mathbb{N}}$.

We explain the above more intuitively in the following section.

§3. Independence

3.1. Independent Events

In the following, let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and the sets $A \in \mathcal{A}$ be events.

Definition 3.1. Two events A and B are said to be *independent* if

$$\Pr[A \cap B] = \Pr[A] \Pr[B].$$

For example, if we roll a die twice, the event of rolling a 6 the first time is independent of the event of rolling a 6 the second time.
Here, $\Omega = [6]^2, \mathcal{A} = 2^\Omega$ and the probability distribution is $\mathbf{P} = \mathcal{U}_\Omega$. Our claim may be verified as follows. Towards showing that the outcome of the first roll is independent of that of the second, let $\tilde{A}, \tilde{B} \subseteq \Omega, A = \tilde{A} \times \Omega$ and $B = \Omega \times \tilde{B}$. We must show that $\Pr[A] \Pr[B] = \Pr[A \cap B]$. This is obvious as follows:

$$\begin{aligned}
\Pr[A] &= \frac{|A|}{36} = \frac{|\tilde{A}|}{6} \\
\Pr[B] &= \frac{|B|}{36} = \frac{|\tilde{B}|}{6} \\
\Pr[A \cap B] &= \frac{|A \cap B|}{36} = \frac{|\tilde{A}||\tilde{B}|}{36} = \Pr[A] \Pr[B].
\end{aligned}$$

While in the above example it is intuitively clear that the two events must be independent, we can have less obvious examples as well. For example, the event that the sum of the two rolls is odd and the event that the first roll gives at most a three are independent. We leave it to the reader to verify this claim.

We extend this definition of two independent events to any number of independent events as follows.

Definition 3.2 (Independence of Events)**.** Let I be an index set and $(A_i)_{i \in I}$ be a family of events. The family $(A_i)_{i \in I}$ is called *independent* if for any finite subset $J \subseteq I$, the following holds:

$$\Pr\left[\bigcap_{j \in J} A_j\right] = \prod_{j \in J} \Pr[A_j].$$

Note that pairwise independence does not guarantee overall independence. For example, For example, consider (X, Y, Z) chosen uniformly from $\{(0,0,0), (1,1,0), (1,0,1), (0,1,1)\}$. Then X, Y, Z are pairwise independent but they are not overall independent.

Let us now return to the product measure defined in definition 2.6, which can be understood intuitively as follows. If E is a finite set of outcomes, consider the probability space comprising $\Omega = E^{\mathbb{N}}$, the σ-algebra

$$\mathcal{A} = \sigma([\omega_1, \ldots, \omega_n] : \omega_1, \ldots, \omega_n \in E \text{ and } n \in \mathbb{N})$$

and the product measure $\mathbf{P} = \left(\sum_{e \in E} p_e \delta_e\right)^{\otimes \mathbb{N}}$. This basically represents that we repeatedly conduct the experiment of choosing an outcome from E. Let $\tilde{A}_i \subseteq E$ for each $i \in \mathbb{N}$ and let A_i be the event such that $\tilde{A}_i$ occurs in the ith experiment, given by

$$A_i = \{\omega \in \Omega : \omega_i \in \tilde{A}_i\} = \biguplus_{(\omega_1, \ldots, \omega_i) \in E^{i-1} \times \tilde{A}_i} [\omega_1, \ldots, \omega_i]$$

Intuitively, the family $(A_i)_{i \in \mathbb{N}}$ should be independent, since the outcome of one of the conducted experiments does not depend on the outcomes of the other experiments.

We shall check this. Let $J \subseteq \mathbb{N}$ and For $j \in J$, let $B_j = A_j$ and $\tilde{B}_j = \tilde{A}_j$ and for $j \in [n] \setminus J$, let $B_j = \Omega$ and $\tilde{B}_j = E$. Then

$$\begin{aligned}
\Pr\left[\bigcap_{j\in J} A_j\right] &= \Pr\left[\bigcap_{j=1}^{n} B_j\right] \\
&= \Pr\left[\left\{\omega \in \Omega : \omega_j \in \tilde{B}_j \text{ for each } j \in [n]\right\}\right] \\
&= \sum_{e_1 \in \tilde{B}_1} \cdots \sum_{e_n \in \tilde{B}_n} \prod_{j=1}^{n} p_{e_j} \\
&= \prod_{j=1}^{n} \left(\sum_{e\in\tilde{B}_j} p_e\right) \\
&= \prod_{j\in J} \left(\sum_{e\in\tilde{A}_j} p_e\right)
\end{aligned}$$

In particular, as this is true for $|J| = 1$, we have for some fixed $i \in [n]$,

$$\Pr[A_i] = \left(\sum_{e\in\tilde{A}_i} p_e\right).$$

Substituting the above, we have

$$\Pr\left[\bigcap_{j\in J} A_j\right] = \prod_{j\in J} \left(\sum_{e\in\tilde{A}_j} p_e\right) = \prod_{j\in J} \Pr[A_j].$$

This proves the result. We state the result for future reference as follows.

Theorem 3.1. Let E be a finite set. Consider the probability space $(\Omega, \mathcal{A}, \mathbf{P})$ where $\Omega = E^{\mathbb{N}}$,

$$\mathcal{A} = \sigma([\omega_1, \ldots, \omega_n] : \omega_1, \ldots, \omega_n \in E \text{ and } n \in \mathbb{N})$$

and $\mathbf{P} = \left(\sum_{e\in E} p_e \delta_e\right)^{\otimes \mathbb{N}}$. For each $i \in \mathbb{N}$, let $\tilde{A}_i \subseteq E$ and A_i be the event such that $\tilde{A}_i$ occurs in the ith experiment, given by

$$A_i = \{\omega \in \Omega : \omega_i \in \tilde{A}_i\} = \biguplus_{(\omega_1,\ldots,\omega_i)\in E^{i-1}\times\tilde{A}_i} [\omega_1, \ldots, \omega_i]$$

Then the family $(A_i)_{i\in\mathbb{N}}$ is independent.

Note that if events A and B are independent, then the events A^c and B are independent as well. This can be described more precisely as follows.

Theorem 3.2. Let I be an index set and $(A_i)_{i\in I}$ be a family of events. Define $B_i^0 = A_i$ and $B_i^1 = A_i^c$ for each $i \in I$. Then the following statements are equivalent.

(a) The family $(A_i)_{i\in I}$ is independent.

(b) There is some $\alpha \in \{0,1\}^I$ such that the family $(B_i^{\alpha_i})_{i\in I}$ is independent.

(c) For all $\alpha \in \{0,1\}^I$, the family $(B_i^{\alpha_i})_{i\in I}$ is independent.

We leave the proof of the above to the reader.
Now, recall the limit superior defined in definition 1.7. The limit superior represents the event that a particular event occurs an infinite amount of times (for example, the event that we roll a 4 an infinite number of times when we roll a die a countably infinite number of times). This is formalized in the following.

Theorem 3.3 (Borel-Cantelli Lemma)**.** Let $A_1, A_2, \ldots$ be events and let $A^* = \limsup_{n\to\infty} A_n$. Then

(a) If $\sum_{n=1}^{\infty} \Pr[A_n] < \infty$, $\Pr[A^*] = 0$.

(b) If $(A_n)_{n\in\mathbb{N}}$ is independent and $\sum_{n=1}^{\infty} \Pr[A_n] = \infty$, then $\Pr[A^*] = 1$.

Proof. By theorem 1.10, $\mathbf{P}$ is upper semicontinuous, lower semicontinuous and σ-subadditive.

(a) As $\mathbf{P}$ is upper semicontinuous and σ-subadditive,

$$\Pr[A^*] = \Pr\left[\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\right]$$

$$\lim_{n\to\infty} \Pr\left[\bigcup_{m=n}^{\infty} A_m\right]$$

$$\leq \lim_{n\to\infty} \sum_{m=n}^{\infty} \Pr[A_m] = 0$$

The result follows.

(b) As $\mathbf{P}$ is lower semicontinuous and the family $(A_n)_{n\in\mathbb{N}}$ is independent, we have

$$\begin{aligned}\Pr[(A^*)^c] &= \Pr\left[\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} A_m^c\right] \\ &= \lim_{n\to\infty} \Pr\left[\bigcap_{m=n}^{\infty} A_m^c\right] \\ &= \lim_{n\to\infty} \prod_{m=n}^{\infty} (1 - \Pr[A_m])\end{aligned}$$

Now for any $n \in \mathbb{N}$, as $\log(1-x) \leq -x$

$$\begin{aligned}\prod_{m=n}^{\infty} (1 - \Pr[A_m]) &= \exp\left(\sum_{m=n}^{\infty} \log(1 - \Pr[A_m])\right) \\ &\leq \exp\left(-\sum_{m=n}^{\infty} \Pr[A_m]\right) = 0\end{aligned}$$

The result follows.

■

A saying the reader might have come across is that if a monkey is left with a typewriter for an infinite amount of time, it will eventually type the complete works of Shakespeare. This is in fact a consequence of the Borel-Cantelli Lemma, and is often referred to as the Infinite Monkey Theorem!

We now extend the definition of independence of a family of events as follows.

Definition 3.3 (Independence of classes of events)**.** Let I be an index set and $\mathcal{E}_i \subseteq \mathcal{A}$ for all $i \in I$. The family $(\mathcal{E}_i)_{i\in I}$ is called *independent* if for any finite $J \subseteq I$ and any choice of $j \in J$ and $E_j \in \mathcal{E}_j$, the family $(E_j)_{j\in J}$ is independent.

For example, if we roll a die an infinite number of times, for each $i \in \mathbb{N}$, consider the class of events given by $\mathcal{E}_i = \{\{\omega \in \Omega : \omega_i \in A\} : A \subseteq [6]\}$ where $\Omega = [6]^{\mathbb{N}}$. Then the family $(\mathcal{E}_i)_{i\in I}$ is independent.

Theorem 3.4. Let I be an index set and for each $i \in I$, let $\mathcal{E}_i \subseteq \mathcal{A}$. Then

(a) Let I be finite. If $\Omega \in \mathcal{E}_i$ for each i, then $(\mathcal{E}_i)_{i\in I}$ is independent if and only if $(E_i)_{i\in I}$ is independent for any choice of $E_i \in \mathcal{E}_i, i \in I$.

(b) If $(\mathcal{E}_i \cup \{\varnothing\})$ is $\cap$-closed for each i, then $(\mathcal{E}_i)_{i\in I}$ is independent if and only if $(\sigma(\mathcal{E}_i))_{i\in I}$ is independent.

Proof.

(a) The forward implication is obvious from the definition. To prove the backward implication, for $J \subseteq I$ and $j \in I \setminus J$, choose $E_j = \Omega$.

(b) The backward implication is obvious. Let us now prove the forward implication.

First, we claim that for any $J \subseteq J' \subseteq I$ where J is finite,

$$\Pr\left[\bigcap_{i\in J'} E_i\right] = \prod_{i\in J'} \Pr[E_i]$$

for any choice of $E_i \in \sigma(\mathcal{E}_i)$ if $i \in J$ and $E_i \in \mathcal{E}_i$ if $i \in J' \setminus J$.

We shall prove the above claim by induction on $|J|$. If $|J| = 0$, then the claim is true as $(\mathcal{E}_i)_{i\in I}$ is independent. Now assume that the claim is true for all $J \subseteq I$ with $|J| = n$ and all finite $J' \supseteq J$. Fix such a J. Let $j \in I \setminus J$. Define $\tilde{J} = J \cup \{j\}$ and choose some $J' \supseteq \tilde{J}$. We shall show that the claim is true if we replace J with $\tilde{J}$, thus proving the inductive step.

Fix $E_i \in \sigma(\mathcal{E}_i)$ for each $i \in J$ and $E_i \in \mathcal{E}_i$ for each $i \in J' \setminus \tilde{J}$. Consider measures μ, ν on $(\Omega, \mathcal{A})$ such that

$$\mu : E_j \mapsto \Pr\left[\bigcap_{i\in J'} E_i\right]$$
$$\nu : E_j \mapsto \prod_{i\in J'} \Pr[E_i]$$

By the induction hypothesis, $\mu(E_j) = \nu(E_j)$ for all $E_j \in \mathcal{E}_j \cup \{\varnothing, \Omega\}$. As $\mathcal{E}_j \cup \{\varnothing\}$ is $\cap$-closed, lemma 1.18 implies that $\mu(E_j) = \nu(E_j)$ for all $E_j \in \sigma(\mathcal{E}_j)$.

This proves our claim. Setting $J = J'$ yields the required result.

■

3.2. Independence of Random Variables

Let I be an index set and $(\Omega, \mathcal{A})$ be measurable space. For each $i \in I$, let $(\Omega_i, \mathcal{A}_i)$ be a measurable space and $X_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i)$ be a random variable with generated σ-algebra $\sigma(X_i)$.

Definition 3.4. The family $(X_i)_{i\in I}$ of random variables is said to be independent if the family $(\sigma(X_i))_{i\in I}$ of generated σ-algebras is independent.

We say that a family $(X_i)_{i\in I}$ of random variables is said to be i.i.d. (independent and identically distributed) if the family $(X_i)_{i\in I}$ is independent and $\mathbf{P}_{X_i} = \mathbf{P}_{X_j}$ for any $i, j \in I$.

The meaning of independence of random variables might be more clear from the following restructuring of the definition. A family $(X_i)_{i\in I}$ of random variables is independent if and only if for any finite set $J \subseteq I$ and any choice of $A_j \in \mathcal{A}_j, j \in J$, we have

$$\Pr\left[\bigcap_{j\in J} \{X_j \in A_j\}\right] = \prod_{j\in J} \Pr[X_j \in A_j].$$

For example, let us flip a coin four times. Let X be the random variable be the number of heads that show up in the first two tosses and Y be the random variable given by the number of tails that show up in the next two tosses. Then X and Y are independent.

For each $i \in I$, let $(\Omega_i', \mathcal{A}_i')$ be another measurable space and assume that $f_i : (\Omega, \mathcal{A}) \to (\Omega_i', \mathcal{A}_i')$ is a measurable map. If $(X_i)_{i\in I}$ is independent, then $(f_i \circ X_i)_{i\in I}$ is independent as well. This is a consequence of the fact that $f_i \circ X_i$ is $\sigma(X_i) - \mathcal{A}_i'$-measurable.

Theorem 3.5. For each $i \in I$, let $\mathcal{E}_i$ be a π-system that generates $\mathcal{A}$. If $(X^{-1}(\mathcal{E}_i))$ is independent, then $(X_i)_{i\in I}$ is independent.

Proof. By theorem 1.12, $X^{-1}(\mathcal{E}_i)$ is a π-system that generates $X^{-1}(\mathcal{A}_i) = \sigma(X_i)$. The result follows from theorem 3.4. ■

Theorem 3.6. Let E be a finite set $(p_e)_{e\in E}$ be a probability vector on E. There then exists a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and an independent family $(X_n)_{n\in\mathbb{N}}$ of E-valued random variables on $(\Omega, \mathcal{A}, \mathbf{P})$ such that $\Pr[X_n = e] = p_e$ for each $e \in E$.

Proof. We shall prove this by constructing the required probability space $(\Omega, \mathcal{A}, \mathbf{P})$. Let $\Omega = E^{\mathbb{N}}$ and $\mathcal{A} = \sigma\left(\{[\omega_1, \omega_2, \ldots, \omega_k] : \omega_i \in E \text{ for each } i \in [k] \text{ and } k \in \mathbb{N}\}\right)$. Let $\mathbf{P} = \left(\sum_{e\in E} p_e\delta_e\right)^{\otimes\mathbb{N}}$ be the product measure. For each $n \in \mathbb{N}$, define the random variable $X_n : \Omega \to E$ by $\omega \mapsto \omega_n$, where ω_n represents the nth coordinate of ω.
As a consequence of theorem 3.1, $(X_j)_{j\in\mathbb{N}}$ is independent. and the result is proved. ■

Definition 3.5. Let I be an index set and for each $i \in I$, let X_i be a random variable. For any $J \subseteq I$, let $F_{(X_j)_{j\in J}} : \mathbb{R}^J \to [0, 1]$ be given by

$$x \mapsto \Pr[X_j \leq x_j \text{ for each } j \in J] = \Pr\left[\bigcap_{j\in J} X_j \leq x_j\right].$$

This function is called the *joint distribution function of* $(X_j)_{j\in J}$ and is denoted F_J. The probability measure $\mathbf{P}_{(X_j)_{j\in J}}$ on $\mathbb{R}^J$ is called the *joint distribution of* $(X_j)_{j\in J}$.

Theorem 3.7. A family $(X_i)_{i\in I}$ of random variables is independent if and only if for every finite $J \subseteq I$ and $x \in \mathbb{R}^J$,

$$F_J(x) = \prod_{j\in J} F_{\{j\}}(x_j).$$

Proof. The forward implication is obvious.
The class of sets $\{(-\infty, a] : a \in \mathbb{R}\}$ is a $\cap$-closed generator of $\mathcal{B}(\mathbb{R})$. The given condition is equivalent to saying that the events $\{X_j \in (-\infty, x_j]\}$ are independent. By theorem 3.5, the backward implication is proved. ■

Corollary 3.8. If in addition to the conditions of the previous theorem, each F_J has a continuous density function $f_J = f_{(X_j)_{j\in J}}$, then the family $(X_i)_{i\in I}$ is independent if and only if for any finite $J \subseteq I$ and $x \in \mathbb{R}^J$,

$$f_J(x) = \prod_{j\in J} f_{\{j\}}(x_j).$$

3.3. The Convolution

Definition 3.6. Let μ and ν be probability measures on $(\mathbb{Z}, 2^{\mathbb{Z}})$. The *convolution* $(\mu * \nu)$ is defined as the probability measure on $(\mathbb{Z}, 2^{\mathbb{Z}})$ given by

$$(\mu * \nu)(\{n\}) = \sum_{m=-\infty}^{\infty} \mu(\{m\})\nu(\{n - m\}).$$

We define the nth convolution power by $\mu^{*1} = \mu$ and $\mu^{*n} = \mu^{*(n-1)} * \mu$.

Theorem 3.9. If X and Y are independent $\mathbb{Z}$-valued random variables, then $\mathbf{P}_{X+Y} = \mathbf{P}_X * \mathbf{P}_Y$.

Proof. For any $n \in \mathbb{Z}$, we have

$$\begin{aligned}\mathbf{P}_{X+Y}[\{n\}] &= \Pr[X+Y=n] \\ &= \Pr\left[\biguplus_{m=-\infty}^{\infty} \{X=m\} \cap \{Y=n-m\}\right] \\ &= \sum_{m=-\infty}^{\infty} \Pr[X=m]\Pr[Y=n-m] \\ &= \sum_{m=-\infty}^{\infty} \mathbf{P}_X[\{m\}]\mathbf{P}_Y[\{n-m\}] = (\mathbf{P}_X * \mathbf{P}_Y)[\{n\}]\end{aligned}$$

■

Given the above theorem, we can generalise the convolution as follows.

Definition 3.7 (Convolution of Probability Measures)**.** Let X and Y be independent random variables on $\mathbb{R}^n$ such that $\mu = \mathbf{P}_X$ and $\nu = \mathbf{P}_Y$. The *convolution* $(\mu * \nu)$ is defined as $\mathbf{P}_{X+Y}$.

We define μ^{*k} for $k \in \mathbb{N}$ recursively similarly to the first case with $\mu^{*0} = \delta_0$.

For example, let $\lambda, \mu \in [0, \infty)$. Consider independent random variables X, Y such that $X \sim \text{Poi}_\mu$ and $Y \sim \text{Poi}_\lambda$. Then for $n \in \mathbb{N}_0$

$$\begin{aligned}\Pr[X+Y=n] &= e^{-\mu}e^{-\lambda}\sum_{m=0}^{n}\frac{\mu^m}{m!}\frac{\lambda^m}{(n-m)!} \\ &= e^{-(\mu+\lambda)}\frac{(\mu+\lambda)^n}{n!}.\end{aligned}$$

Thus $\text{Poi}_\lambda * \text{Poi}_\mu = \text{Poi}_{\lambda+\mu}$.

3.4. Kolmogorov's 0-1 Law

The Borel-Cantelli Lemma was an example of a so-called $0-1$ law. In this subsection, we study another such 0-1 law.

Definition 3.8 (Tail σ-algebra)**.** Let I be a countably infinite index set and $(\mathcal{A}_i)_{i\in I}$ be a family of σ-algebras. Then

$$\mathcal{T}((\mathcal{A}_i)_{i\in I}) = \bigcap_{\substack{J\subseteq I \\ |J|<\infty}} \sigma\left(\bigcup_{j\in I\setminus J}\mathcal{A}_j\right)$$

is called the *tail σ-algebra* of $(\mathcal{A}_i)_{i\in I}$. If $(A_i)_{i\in I}$ is a family of events, we define

$$\mathcal{T}((A_i)_{i\in I}) = \mathcal{T}((\{\varnothing, A_i, A_i^c, \Omega\})_{i\in I}).$$

If $(X_i)_{i\in I}$ is a family of random variables, we define

$$\mathcal{T}((X_i)_{i\in I}) = \mathcal{T}((\sigma(X_i))_{i\in I}).$$

If the meaning is clear from context, we represent the tail σ-algebra as just $\mathcal{T}$.
Intuitively, the above means that we consider those events that are independent of the values of any finite subfamily of $(X_i)_{i\in I}$. This shall be made clearer as follows.

Theorem 3.10. Let $J_1, J_2, \ldots$ be finite sets with $J_n \uparrow I$. Then

$$\mathcal{T}((\mathcal{A}_i)_{i\in I}) = \bigcap_{n=1}^{\infty}\sigma\left(\bigcup_{m\in I\setminus J_n}\mathcal{A}_m\right).$$

If $I = \mathbb{N}$, then this says that

$$\mathcal{T}((\mathcal{A}_i)_{i\in\mathbb{N}}) = \bigcap_{n=1}^{\infty} \sigma\left(\bigcup_{m=n}^{\infty} \mathcal{A}_m\right).$$

Proof. It is obvious from the definition that $\mathcal{T}((\mathcal{A}_i)_{i\in I})$ is a subset of the expression on the right.
Let $J_n \uparrow I$ and $J \subseteq I$ be a finite set. There exists some $n_0 \in \mathbb{N}$ such that $J \subseteq J_{n_0}$. We then have

$$\begin{aligned}\bigcap_{n=1}^{\infty} \sigma\left(\bigcup_{m\in I\setminus J_n} \mathcal{A}_m\right) &\subseteq \bigcap_{n=1}^{N} \sigma\left(\bigcup_{m\in I\setminus J_n} \mathcal{A}_m\right)\\ &= \sigma\left(\bigcup_{m\in I\setminus J_N} \mathcal{A}_m\right)\\ &\subseteq \sigma\left(\bigcup_{m\in I\setminus J} \mathcal{A}_m\right)\end{aligned}$$

Noting that the expression on the left does not depend on J and taking the intersection over all J implies the reverse inclusion and the result follows. ■

If we interpret $I = \mathbb{N}$ as a set of times, the above theorem essentially says that any event in $\mathcal{T}$ is independent of the first finitely many time points.

Now, it is not immediately clear whether $\mathcal{T}$ even contains any nontrivial events (events other than $\varnothing$ and Ω).

For starters, if $A_1, A_2, \ldots$ are events, then $A^* = \limsup_{n\to\infty} A_n$ and $A_* = \liminf_{n\to\infty} A_n$ are both in $\mathcal{T}((A_i)_{i\in\mathbb{N}})$.
To see this for A_*, define $B_n = \bigcap_{m=n}^{\infty} A_m$ for $n \in \mathbb{N}$. We then have that $B_n \uparrow A_*$ and $B_n \in \sigma(\bigcup_{m=n_0}^{\infty} A_m)$ for any $n \geq n_0$. This implies that $A_* \in \sigma(\bigcup_{m=n_0}^{\infty} A_m)$ for any $n_0 \in \mathbb{N}$ and thus, $A_* \in \mathcal{T}$.
For A^*, we must show that for any $m \in \mathbb{N}$, $\limsup_{n\to\infty} A_n \in \sigma(\bigcup_{n=m}^{\infty} \sigma(A_n))$. This is true as $\limsup_{n\to\infty} A_n = \limsup_{n\to\infty} A_{n+m}$. Since each A_{n+m} is in $\sigma(\bigcup_{n=m}^{\infty} \sigma(A_n))$, the statement is true.

Let $(X_n)_{n\in\mathbb{N}}$ be real random variables. Then the Cesàro limits

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i$$

are $\mathcal{T}((X_n)_{n\in\mathbb{N}})$-measurable.

Theorem 3.11 (Kolmogorov's 0-1 Law)**.** Let I be a countable infinite index set and $((\mathcal{A}_i)_{i\in I})$ be an independent family of σ-algebras. Then the tail σ-algebra is **P**-trivial, that is,

$$\Pr[A] \in \{0, 1\} \text{ for any } A \in \mathcal{T}((\mathcal{A}_i)_{i\in I}).$$

Proof. Assume w.l.o.g. that $I = \mathbb{N}$. For each $i \in \mathbb{N}$, let

$$\mathcal{F}_n = \left\{\bigcap_{i=1}^{n} A_k : A_j \in \mathcal{A}_j \text{ for each } j \in [n]\right\}.$$

Let $\mathcal{F} = \bigcup_{i=1}^{\infty} \mathcal{F}_i$. Note that $\mathcal{F}$ is a semiring.
Further, for any $m \in \mathbb{N}$ and $A_m \in \mathcal{A}_m$, we have $A_m \in \mathcal{F}$. This implies that $\sigma\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) \subseteq \sigma(\mathcal{F})$. We also have

$$\mathcal{F}_m \subseteq \sigma\left(\bigcup_{i=1}^{m} \mathcal{A}_i\right) \subseteq \sigma\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right).$$

This implies that $\mathcal{F} = \sigma\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right)$.

Let $A \in \mathcal{T}((\mathcal{A}_n)_{n\in\mathbb{N}})$ and $\varepsilon > 0$. By theorem 1.24, there exists $n_0 \in \mathbb{N}$ and mutually disjoint sets $F_1, F_2, \ldots, F_{n_0}$ such that

$$\Pr\left[A \triangle \bigcup_{i=1}^{n_0} F_i\right] < \varepsilon.$$

Let $F = \bigcup_{i=1}^{n_0} F_i$. There must be some $n \in \mathbb{N}$ such that $F_1, \ldots, F_{n_0} \in \mathcal{F}_n$. This implies that $F \in \sigma(\bigcup_{i=1}^{n} \mathcal{A}_i)$. By the definition of the tail σ-algebra, $A \in \sigma(\bigcup_{i=n+1}^{\infty} \mathcal{A}_i)$ so A must be independent of F. Therefore,

$$\begin{aligned}\varepsilon &> \Pr[A \setminus F] \\ &= \Pr[A](1 - \Pr[F]) \\ &\geq \Pr[A](1 - \Pr[A] - \varepsilon).\end{aligned}$$

As this is true for any $\varepsilon > 0$, $0 = \Pr[A](1 - \Pr[A])$ and the result is proved. ■

Corollary 3.12. Let $(A_n)_{n\in\mathbb{N}}$ be an independent family of events. Then

$$\Pr\left[\limsup_{n\to\infty} A_n\right] \text{ and } \Pr\left[\liminf_{n\to\infty} A_n\right] \text{ are in } \{0, 1\}.$$

The above can be inferred from the fact that the lim sup and lim inf lie in the tail σ-algebra. It also follows from the Borel-Cantelli Lemma.

Corollary 3.13. Let $(X_n)_{n\in\mathbb{N}}$ be an independent family of $\overline{\mathbb{R}}$-valued random variables. Then $X_* = \liminf_{n\to\infty} X_n$ and $X^* = \limsup_{n\to\infty} X_n$ are almost surely constant, that is, there exist $x_*, x^* \in \overline{\mathbb{R}}$ such that $\Pr[X_* = x_*] = 1$ and $\Pr[X^* = x^*] = 1$.

The above follows from the fact that for any $x \in \overline{\mathbb{R}}$, $\{X_* < x\} \in \mathcal{T}((X_n)_{n\in\mathbb{N}})$ and $\{X^* > x\} \in \mathcal{T}((X_n)_{n\in\mathbb{N}})$.

§4. Generating Functions

It is a common theme in mathematics to determine relations between objects that are of interest and objects that are easy to compute with. In probability theory, probability generating functions, Laplace transforms and characteristic functions fall in the former category while the mean, median and variance of random variables fall in the latter.

4.1. Definitions and Basics

Definition 4.1 (Probability Generating Function)**.** Let X be an $\mathbb{N}_0$-valued random variable. The *probability generating function* (abbreviated *pgf*) of $\mathbf{P}_X$ (or X) is the map $\psi_{\mathbf{P}_X} = \psi_X : [0,1] \to [0,1]$ is given by (where $0^0 = 1$)

$$z \mapsto \sum_{n=0}^{\infty} \Pr[X = n] z^n.$$

From the properties of a power series, we have the following result, which we do not prove.

Theorem 4.1.

(a) ψ_X is continuous on $[0,1]$ and infinitely often continuously differentiable on (0,1). For $n \in \mathbb{N}$,

$$\lim_{z \to 1^-} \psi_X^{(n)}(z) = \sum_{k=n}^{\infty} \Pr[X = k] \cdot k(k-1)\cdots(k-n+1)$$

where both sides can equal ∞.

(b) $\mathbf{P}_X$ is uniquely determined by ψ_X.

(c) For any $r \in (0,1)$, ψ_X is uniquely determined by countably many values $\psi_X(x_i)$ where $x_i \in [0,r]$ for each $i \in \mathbb{N}$. If the series given in ψ_X converges for some $z > 1$, then this statement is also true for any $r \in (0,z)$ and

$$\lim_{z \to 1^-} \psi_X^{(n)}(z) = \psi_X^{(n)}(1) < \infty \text{ for } n \in \mathbb{N}.$$

Here, ψ_X is uniquely determined by $\psi_X^{(n)}(1), n \in \mathbb{N}$.

While this definition of a pgf may seem quite arbitrary, it is useful when we want to add random variables, as is evident from the following theorem.

Theorem 4.2. Let $X_1, X_2, \ldots, X_n$ be independent $\mathbb{N}_0$-valued random variables. Then

$$\psi_{X_1+X_2+\ldots+X_n} = \prod_{i=1}^{n} \psi_{X_i}.$$

Proof. We shall prove the claim for $n = 2$ and the result will follow inductively. For any $z \in [0,1]$

$$\begin{aligned}
\psi_{X_1}(z) \cdot \psi_{X_2}(z) &= \left(\sum_{n=0}^{\infty} \Pr[X_1 = n] z^n\right)\left(\sum_{n=0}^{\infty} \Pr[X_2 = n] z^n\right) \\
&= \sum_{n=0}^{\infty} z^n \left(\sum_{m=0}^{n} \Pr[X_1 = m] \Pr[X_2 = n-m]\right) \\
&= \sum_{n=0}^{\infty} z^n \left(\sum_{m=0}^{n} \Pr\left[\{X_1 = m\} \cap \{X_2 = n-m\}\right]\right) \\
&= \sum_{n=0}^{\infty} z^n \Pr[X_1 + X_2 = n] \\
&= \psi_{X_1+X_2}.
\end{aligned}$$

■

For example,

- For some $m, n \in \mathbb{N}$ and $p \in [0,1]$, let $X \sim b_{n,p}$ and $Y \sim b_{m,p}$. Then

$$\psi_X(z) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^i z^i = (pz + 1 - p)^n.$$

  If X and Y are independent, then

$$\psi_{X+Y}(z) = \psi_X(z) \cdot \psi_Y(z) = (pz + 1 - p)^{m+n}.$$

  Therefore, $X + Y \sim b_{m+n,p}$, which is not immediately apparent otherwise. (Note that this also implies that $b_{m,p} * b_{n,p} = b_{m+n,p}$)

- Let $p \in (0,1]$ and $X_1, \ldots, X_n \sim \gamma_p$ be independent random variables. Define $Y = X_1 + \cdots + X_n$. For any $z \in [0,1]$,

$$\psi_{X_1}(z) = \sum_{i=0}^{\infty} p(z(1-p))^i = \frac{p}{1 - z(1-p)}.$$

  Then

$$\begin{aligned}
\psi_Y(z) &= \frac{p^n}{(1 - z(1-p))^n} \\
&= \sum_{i=0}^{\infty} p^n \binom{-n}{i} (-1)^i (1-p)^i z^i \\
&= \sum_{i=0}^{\infty} b^-_{n,p}(\{i\}) z^i.
\end{aligned}$$

  Therefore, $\gamma_p^{*n} = b^-_{n,p}$. Note that this matches with the intuition we introduced while defining the geometric and negative binomial distributions. The waiting time for the nth success is equivalent to waiting for a single success n times, that is, $X_1 + \cdots + X_n$.

- We can also show that for $\lambda, \mu \in [0, \infty)$,

$$\psi_{\mathrm{Poi}_\lambda}(z) = e^{\lambda(z-1)}.$$

  This implies that $\mathrm{Poi}_\lambda * \mathrm{Poi}_\mu = \mathrm{Poi}_{\lambda+\mu}$ (Recall that we had also proved this earlier by manually calculating the convolution).

- For $r, s \in (0, \infty)$ and $p \in (0,1]$, we have $b^-_{r,p} * b^-_{s,p} = b^-_{r+s,p}$. We leave it to the reader to prove this.

4.2. The Poisson Approximation

Theorem 4.3. Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{N}_0, 2^{\mathbb{N}_0})$ with generating functions $\psi, \psi_1, \psi_2, \ldots$. Then the following are equivalent.

(a) $\lim_{n\to\infty} \mu_n(\{k\}) = \mu(\{k\})$ for all $k \in \mathbb{N}_0$.

(b) $\lim_{n\to\infty} \mu_n(A) = \mu(A)$ for all $A \subseteq \mathbb{N}_0$.

(c) $\lim_{n\to\infty} \psi_n(z) = \psi(z)$ for all $z \in [0,1]$.

(d) $\lim_{n\to\infty} \psi_n(z) = \psi(z)$ for all $z \in [0, \eta)$ for some $\eta \in (0,1)$.

If any of the above is true, we write $\lim_{n\to\infty} \mu_n = \mu$ and say that $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ.

Proof.

- (a) $\implies$ (b).
  Fix $\varepsilon > 0$. Choose some $N \in \mathbb{N}$ such that
  $$\mu(\{N+1, N+2, \ldots\}) < \frac{\varepsilon}{4}$$
  and sufficiently large $n_0 \in \mathbb{N}$ such that
  $$\sum_{k=0}^{N} |\mu_n(\{k\}) - \mu(\{k\})| < \frac{\varepsilon}{4} \text{ for all } n \geq n_0.$$
  Note that for any $n \geq n_0$, $\mu_n(\{N+1, N+2, \ldots\}) < \frac{\varepsilon}{2}$. Therefore, for any $A \subseteq \mathbb{N}$ and $n \geq n_0$,
  $$\begin{aligned} |\mu_n(A) - \mu(A)| &\leq \mu_n(\{N+1, N+2, \ldots\}) + \mu(\{N+1, N+2, \ldots\}) + \sum_{k \in A \cap \{0,1,\ldots,N\}} |\mu_n(\{k\}) - \mu(\{k\})| \\ &\leq \varepsilon. \end{aligned}$$
  This proves the result.
- (b) $\implies$ (a).
  This is trivial.
- (a) $\iff$ (c) $\iff$ (d)
  This follows from theorem 4.1.

■

When we are dealing with the binomial distribution for large n, it is usually inconvenient to calculate the terms. However, in the case where n is very large and p is very small such that $np = \lambda$ is of reasonable magnitude, then we can approximate $b_{n,p}$ by Poi_λ. This is made rigorous as follows.

Let $(p_{n,k})_{n,k \in \mathbb{N}}$ such that $p_{n,k} \in [0,1]$. Let
$$\lambda = \lim_{n \to \infty} \sum_{k=1}^{\infty} p_{n,k} \in (0, \infty) \text{ and } \lim_{n \to \infty} \sum_{k=1}^{\infty} p_{n,k}^2 = 0.$$
An example of such a family would be $p_{n,k} = \lambda/n$ for $k \leq n$ and $p_{n,k} = 0$ otherwise.
For each $n \in \mathbb{N}$, let $(X_{n,k})_{k \in \mathbb{N}}$ be a family of independent random variables such that $X_{n,k} \sim \text{Ber}_{p_{n,k}}$. For $n, k \in \mathbb{N}$, define
$$S^n = \sum_{l=1}^{\infty} X_{n,l} \text{ and } S_k^n = \sum_{l=1}^{k} X_{n,l}.$$

Theorem 4.4 (Poisson Approximation)**.** With the above notation, the distributions $(\mathbf{P}_{S^n})_{n \in \mathbb{N}}$ converge weakly to Poi_λ.

Proof. For any $z \in [0,1]$,
$$\begin{aligned} \psi_{S^n}(z) &= \prod_{l=1}^{\infty} (p_{n,l} z + 1 - p_{n,l}) \\ &= \exp\left(\sum_{l=1}^{\infty} \log(p_{n,l} z + 1 - p_{n,l}) \right). \end{aligned}$$
Now, as $|\log(1-x) - x| \leq x^2$ for $|x| < \frac{1}{2}$, we have
$$\left| \left(\sum_{l=1}^{\infty} \log(p_{n,l} z + 1 - p_{n,l}) \right) - \left((z-1) \sum_{l=1}^{\infty} p_{n,l} \right) \right| \leq \sum_{l=1}^{\infty} p_{n,l}^2.$$

Taking the limit as $n \to \infty$, we have

$$\lim_{n\to\infty} \psi_{S^n}(z) = \exp\left((z-1)\sum_{l=1}^{\infty} p_{n,l}\right) = e^{\lambda(z-1)}.$$

As $\psi_{\mathrm{Poi}_\lambda} = e^{\lambda(z-1)}$, this completes the proof. ■

For example, let $p_{n,k}$ be λ/n if $k \leq n$ and 0 otherwise. Then by the Poisson approximation, $\lim_{n\to\infty} b_{n,\frac{\lambda}{n}} = \mathrm{Poi}_\lambda$.

4.3. Branching Processes

Let $T, X_1, X_2, \ldots$ be $\mathbb{N}_0$-valued random variables and $S = \sum_{i=1}^{T} X_i$. Note that S is measurable (and thus a random variable) as

$$\{S = k\} = \bigcup_{i=0}^{\infty} \{T = i\} \cap \{X_1 + \cdots + X_n = k\}.$$

Theorem 4.5. With the above notation, if $X_1, X_2, \ldots$ are all identically distributed, then $\psi_S = \psi_T \circ \psi_{X_1}$.

Proof. We have

$$\begin{aligned}
\psi_S(z) &= \sum_{k=0}^{\infty} \Pr[S = k] z^k \\
&= \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \Pr[T = i] \Pr[X_1 + \cdots + X_n = k] z^k \\
&= \sum_{i=0}^{\infty} \Pr[T = i] (\psi_{X_1}(z))^n \\
&= (\psi_T \circ \psi_{X_1})(z).
\end{aligned}$$

■

Let $p_0, p_1, \ldots \in [0,1]$ such that $\sum_{i=0}^{\infty} p_i = 1$. Let $(X_{n,i})_{n,i\in\mathbb{N}_0}$ be an independent family of random variables such that $\Pr[X_{n,i} = k] = p_k$ for any $k, n, i \in \mathbb{N}$.

Let $Z_0 = 1$ and $Z_n = \sum_{i=1}^{Z_{n-1}} X_{n,i}$ for each $n \in \mathbb{N}$. This can be interpreted as the number of members in each generation of a family where the number of offspring each person has is random and given by $X_{n,i}$.

Definition 4.2. $(Z_n)_{n\in\mathbb{N}_0}$ is called a *Galton-Watson process* or *branching process* with offspring distribution $(p_k)_{k\in\mathbb{N}_0}$

Branching processes are easier to study with the assistance of generator functions. For $z \in [0,1]$, let

$$\psi(z) = \sum_{k=0}^{\infty} p_k z^k.$$

Recursively define

$$\psi_1 = \psi \text{ and } \psi_{n+1} = \psi_n \text{ for all } n \in \mathbb{N}.$$

Theorem 4.6. $\psi_{Z_n} = \psi_n$ for all $n \in \mathbb{N}$.

Proof. We have $\psi_{Z_1} = \psi_1$ (by definition). By theorem 4.5, $\psi_{Z_{n+1}} = \psi \circ \psi_{Z_n}$. The result follows inductively. ■

Now, let us study the probability that the family dies out.
Denote by $q_n = \Pr[Z_n = 0]$ the probability that Z is extinct by time n. Clearly, q_n is monotone increasing in n. In the limiting case, we have the following.

Definition 4.3. Let $(Z_n)_{n\in\mathbb{N}_0}$ be a branching process. We define its extinction probability by

$$q = \lim_{n\to\infty} \Pr[Z_n = 0].$$

A natural question to ask is: under what condition does the family definitely die out, that is, $q = 1$? We clearly have $q \geq p_0$ since q_n is monotone increasing and $q = \lim_{n\to\infty} q_n$. If $p_0 = 0$, then Z_n is monotone increasing in n and as a result, $q = 0$ as well.

Theorem 4.7 (Extinction Probability of a Branching Process)**.** Let $(Z_n)_{n\in\mathbb{N}_0}$ be a branching process with offspring distribution $(p_k)_{k\in\mathbb{N}_0}$ such that $p_1 \neq 1$. Then

(a) $\{r \in [0,1] : \psi(r) = r\} = \{q, 1\}$.

(b) The following holds:

$$q < 1 \iff \lim_{z\to 1} \psi'(z) > 1 \iff \sum_{k=1}^{\infty} kp_k > 1.$$

Proof.

(a) Let $F = \{r \in [0,1] : \psi(r) = r\}$. Clearly, $1 \in F$ as $\psi(1) = 1$. Note that $q_n = \psi_n(0) = \psi(q_{n-1})$ for all $n \in \mathbb{N}$. As ψ is continuous,

$$\psi(q) = \psi\left(\lim_{n\to\infty} q_n\right) = \lim_{n\to\infty} \psi(q_n) = \lim_{n\to\infty} q_{n+1} = q.$$

Thus $q \in F$.

We next claim that $q = \min F$. Let $r \in F$. Then $r \geq 0 = q_0$. If $r \geq q_n$ for some $n \in \mathbb{N}$, then as ψ is monotone increasing, $r = \psi(r) \geq \psi(q_n) = q_{n+1}$.

Inductively, $r \geq q_n$ for every $n \in \mathbb{N}$. The claim follows. We complete the remainder of the proof in the second part.

(b) The second equivalence follows by theorem 4.1. For the first equivalence, consider the following two cases.

- $\lim_{z\to 1} \psi'(z) \leq 1$. Since ψ is strictly convex, it follows that $\psi(z) > z$ for all $z \in [0,1)$ and so $F = \{1\}$. By the first part of the proof, $q = 1$.
- $\lim_{z\to 1} \psi'(z) > 1$. Since ψ is strictly convex and $\psi(0) \geq 0$, there is some unique $r \in (0,1)$ such that $\psi(r) = r$. Then $F = \{r, 1\}$ and by the first part, $q = \min F = r < 1$.

This completes the proof. ■

§5. The Integral

In the following, we assume $(\Omega, \mathcal{A}, \mu)$ to be a measure space. We denote by $\mathcal{E}$ the vector space of simple functions on $(\Omega, \mathcal{A})$ and by $\mathbb{E}^+ = \{f \in E : f \geq 0\}$ the cone of nonnegative simple functions. If

$$f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}$$

for some $n \in \mathbb{N}$, mutually disjoint sets $A_1, \ldots, A_n \in \mathcal{A}$ and $\alpha_1, \ldots, \alpha_n \in (0, \infty)$ then this representation is said to be a *normal representation of* f.

5.1. Set Up to Define the Integral

Lemma 5.1. Let $f = \sum_{i=1}^{m} \alpha_i \mathbb{1}_{A_i}$ and $f = \sum_{j=1}^{n} \beta_j \mathbb{1}_{B_j}$ be two normal representations of $f \in \mathbb{E}^+$. Then

$$\sum_{i=1}^{m} \alpha_i \mu(A_i) = \sum_{j=1}^{n} \beta_j \mu(B_j)$$

Proof. Clearly, if $\alpha_i \neq 0$ for some i, then $A_i \subseteq \bigcup_{j=1}^{n} B_j$. A similar result holds for B_j. Thus,

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

If $\mu(A_i \cap B_j) \neq 0$, then $f(\omega) = \alpha_i = \beta_j$ for any $\omega \in A_i \cap B_j$. Therefore,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \mu(A_i \cap B_j) = \sum_{j=1}^{n} \sum_{i=1}^{m} \beta_j \mu(A_i \cap B_j) = \sum_{j=1}^{n} \beta_j \mu(B_j).$$

■

Definition 5.1. Define $I : \mathbb{E}^+ \to [0, \infty]$ by

$$I(f) = \sum_{i=1}^{m} \alpha_i \mu(A_i)$$

if f has the normal representation $f = \sum_{i=1}^{m} \alpha_i \mathbb{1}_{A_i}$.

The above definition makes sense due to the previous lemma.

Lemma 5.2. Let $f, g \in \mathbb{E}^+$ and $\alpha \geq 0$. Then

(a) $I(\alpha f) = \alpha I(f)$,

(b) $I(f + g) = I(f) + I(g)$, and

(c) If $f \leq g$, then $I(f) \leq I(g)$.

We leave the proof of this theorem as an exercise to the reader.

Definition 5.2 (Integral)**.** If $f : \Omega \to [0, \infty]$ is measurable, then we define the *integral* of f with respect to μ by

$$\int f \, \mathrm{d}\mu = \sup \left\{ I(g) : g \in \mathbb{E}^+, g \leq f \right\}.$$

Note that by lemma 5.2(iii), $I(f) = \int f \, \mathrm{d}\mu$ for any $f \in \mathbb{E}^+$. That is, the integral is an extension of I from $\mathbb{E}^+$ to the set of non-negative measurable functions. We expand this to measurable functions in general in the next subsection.

Let $f, g : \Omega \to \overline{\mathbb{R}}$. Similar to how we write $f \leq g$ if $f(\omega) \leq g(\omega)$ for all $\omega \in \Omega$, we write $f \leq g$ almost everywhere if there exists some set $N \in \mathcal{A}$ such that $\mu(N) = 0$ and $f(\omega) \leq g(\omega)$ for all $\omega \in \Omega \setminus N$.

Theorem 5.3. Let $f, g, f_1, f_2, \ldots$ be measurable maps $\Omega \to [0, \infty]$. Then

(a) If $f \leq g$, then $\int f \, d\mu \leq \int g \, d\mu$.

(b) If $f_n \uparrow f$, then $\int f_n \, d\mu \uparrow \int f \, d\mu$.

(c) If $\alpha, \beta \in [0, \infty]$, then

$$\int (\alpha f + \beta g) \, d\mu = \alpha \int f \, d\mu + \beta \int g \, d\mu$$

where we take $\infty \cdot 0 = 0$.

Proof.

(a) This is obvious from the definition of the integral.

(b) By the definition of the integral,

$$\lim_{n\to\infty} \int f_n \, d\mu = \sup_{n\in\mathbb{N}} \int f_n \, d\mu \leq \int f \, d\mu$$

We must now show that $\int f \, d\mu \leq \sup_{n\in\mathbb{N}} \int f_n \, d\mu$. Let $g \in \mathbb{E}^+$ with $g \leq f$. It is enough to show that $\sup_{n\in\mathbb{N}} \int f_n \, d\mu \geq \int g \, d\mu$. Fix some $t \in (0, 1)$. For each $n \in \mathbb{N}$, define

$$A_n = \{\omega \in \Omega : f_n(x) \geq tg(x)\}.$$

Note that each A_n is measurable (Why?) and that $A_i \subset A_{i+1}$ for each $i \in \mathbb{N}$.

We first claim that $\bigcup_{i=1}^{\infty} A_i = \Omega$. We prove this as follows. For any $\omega \in \Omega$,

- If $f(\omega) \leq tg(\omega)$, then $f(\omega) = 0$ and $\omega \in A_n$ for every $n \in \mathbb{N}$.
- If $f(\omega) > tg(\omega)$, then there exists some $n \in \mathbb{N}$ such that $f_n(\omega) > tg(\omega)$. It follows that $\omega \in A_n$.

Therefore, $\Omega \subseteq \bigcup_{i=1}^{\infty} A_i$. Since the reverse inclusion is obviously true, we have $\bigcup_{i=1}^{n} A_i = \Omega$.

Now,

$$\int f_n \, d\mu \geq \int tg \mathbb{1}_{A_n} \, d\mu$$

Taking the limit as $n \to \infty$ and $t \to 1$, we have

$$\lim_{n\to\infty} \int f_n \, d\mu \geq \int g \, d\mu.$$

This completes the proof.

(c) By theorem 1.17, there exist sequences $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ in $\mathbb{E}^+$ such that $f_n \uparrow f$ and $g_n \uparrow g$. Then by lemma 5.2 and (ii),

$$\begin{aligned}\int (\alpha f + \beta g) \, d\mu &= \lim_{n\to\infty} \int (\alpha f + \beta g) \, d\mu \\ &= \alpha \lim_{n\to\infty} \int f \, d\mu + \beta \lim_{n\to\infty} \int g \, d\mu = \alpha \int f \, d\mu + \beta \int g \, d\mu.\end{aligned}$$

■

It is important to note that (b) in the above only works if the sequence of functions is non-decreasing. We do *not* have a similar result for a non-increasing sequence. To see this, consider $f_1, f_2, \ldots : \mathcal{B}(\mathbb{R}) \to \overline{\mathbb{R}}$ given by $f_n = \mathbb{1}_{[n,\infty)}$ for each $n \in \mathbb{N}$.

5.2. The Integral and some Properties

We have now introduced enough to define the integral for measurable functions in general.

Definition 5.3 (Integrals of Measurable Functions)**.** Let $f : \Omega \to \overline{\mathbb{R}}$ be measurable. We call f *μ-integrable* if $\int |f| \, \mathrm{d}\mu < \infty$ and write

$$\mathcal{L}^1(\mu) = \mathcal{L}^1(\Omega, \mathcal{A}, \mu) = \left\{ f : \Omega \to \overline{\mathbb{R}} : f \text{ is } \mu\text{-integrable} \right\}.$$

For $f \in \mathcal{L}^1(\mu)$, we define the integral of f with respect to μ by

$$\int f(\omega)\mu(\mathrm{d}\omega) = \int f \, \mathrm{d}\mu = \int f^+ \, \mathrm{d}\mu - \int f^- \, \mathrm{d}\mu.$$

For $A \in \mathcal{A}$, we define

$$\int_A f \, \mathrm{d}\mu = \int (f \mathbb{1}_A) \, \mathrm{d}\mu.$$

Elements of $\mathcal{L}^1(\mathbf{P})$ are also sometimes called *absolutely integrable* functions.

Theorem 5.4. Let $f : \Omega \to [0, \infty]$ be measurable.

(a) $f = 0$ almost everywhere if and only if $\int f \, \mathrm{d}\mu = 0$.

(b) If $\int f \, \mathrm{d}\mu < \infty$, then $f < \infty$ almost everywhere.

Proof.

(a) Let us first prove the forward implication. Let $P = \{\omega \in \Omega : f(\omega) > 0\}$. Then $f \leq \infty \cdot \mathbb{1}_P$. As $n\mathbb{1}_N \uparrow \infty \mathbb{1}_N$, by theorem 5.3,

$$0 \leq \int f \, \mathrm{d}\mu \leq \lim_{n \to \infty} \int n \mathbb{1}_N \, \mathrm{d}\mu = 0.$$

For the backward implication, let $A_n = \{\omega \in \Omega : f(\omega) \geq 1/n\}$. Then $A_n \uparrow P$ and for any $n \in \mathbb{N}$,

$$0 = \int f \, \mathrm{d}\mu \geq \int \frac{1}{n} \mathbb{1}_{A_n} = \frac{\mu(A_n)}{n}.$$

This implies that $\mu(A_n) = 0$ for any $n \in \mathbb{N}$ and therefore, $\mu(P) = 0$.

(b) Let $A = \{\omega \in \Omega : f(\omega) = \infty\}$. Then for any $n \in \mathbb{N}$,

$$\mu(A) = \int \mathbb{1}_A \, \mathrm{d}\mu \leq \frac{1}{n} \int f \mathbb{1}_{f \geq n} \leq \frac{1}{n} \int f \, \mathrm{d}\mu.$$

Taking the limit as $n \to \infty$, we get $\mu(A) = 0$.

■

We now expand some of the properties that we proved earlier for non-negative measurable functions to measurable functions in general.

Theorem 5.5. Let $f, g \in \mathcal{L}^1(\mu)$.

(a) (Monotonicity). If $f \leq g$ almost everywhere, then $\int f \, \mathrm{d}\mu \leq \int g \, \mathrm{d}\mu$. In particular, if $f = g$ almost everywhere, then $\int f \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu$.

(b) (Triangle Inequality). $|\int f \, \mathrm{d}\mu| \leq \int |f| \, \mathrm{d}\mu$.

(c) (Linearity). If $\alpha, \beta \in \mathbb{R}$, then $\alpha f + \beta g \in \mathcal{L}^1(\mu)$ and

$$\int \alpha f + \beta g \, \mathrm{d}\mu = \alpha \int f \, \mathrm{d}\mu + \beta \int g \, \mathrm{d}\mu.$$

Proof.

(a) Since $f \leq g$ almost everywhere, $f^+ \leq g^+$ and $f^- \geq g^-$ almost everywhere. It is enough to show that $\int f^+ \, d\mu \leq \int g^+ \, d\mu$ and $\int f^- \, d\mu \geq \int g^- \, d\mu$. Let us prove the former. The latter can similarly be shown.

Let $h = g^+ - f^+$. As $h^- = 0$ almost everywhere, by theorem 5.4, $\int h \, d\mu = \int h^+ \, d\mu \geq 0$. The result follows.

(b) We have

$$\begin{aligned}\left|\int f \, d\mu\right| &= \left|\int f^+ \, d\mu - \int f^- \, d\mu\right| \\ &\leq \int (f^+ + f^-) \, d\mu \\ &= \int |f| \, d\mu.\end{aligned}$$

(c) To show linearity, it suffices to show that for any $\alpha \in [0, \infty)$,

- $\int (f+g) \, d\mu = \int f \, d\mu + \int g \, d\mu$,
- $\int \alpha f \, d\mu = \alpha \int f \, d\mu$, and
- $\int (-f) \, d\mu = -\int f \, d\mu$.

These are easily shown by splitting each function f into f^+ and f^- and using theorem 5.3 wherever necessary.

■

Definition 5.4. Let $f : \Omega \to [0, \infty)$ be measurable. Define the measure ν by

$$\nu(A) = \int (\mathbb{1}_A f) \, d\mu \text{ for } A \in \mathcal{A}.$$

Then ν is said to have *density* f with respect to μ. We also denote ν by $f\mu$.

In showing that ν is a measure using theorem 1.10, finite additivity follows from finite additivity of the integral and lower semicontinuity follows from theorem 5.9.

Theorem 5.6. Let $f : \Omega \to [0, \infty)$ and $g : \Omega \to \overline{\mathbb{R}}$ be measurable. Then $g \in \mathcal{L}^1(f\mu)$ if and only if $(gf) \in \mathcal{L}^1(\mu)$. In this case,

$$\int g \, d(f\mu) = \int (gf) \, d\mu.$$

We omit the proof of the above. It may be shown by first assuming g to be an indicator function, then extending this to simple functions, non-negative measurable functions and measurable functions.

Definition 5.5. Let $f : \Omega \to \overline{\mathbb{R}}$ be measurable and $p \in [1, \infty)$. Define

$$\|f\|_p = \left(\int |f|^p \, d\mu\right)^{1/p} \text{ and}$$

$$\|f\|_\infty = \inf\{k \geq 0 : \mu(\{|f| \geq k\}) = 0\}.$$

Further, for any $p \in [0, \infty]$, we define the vector space

$$\mathcal{L}^p(\mu) = \{f : \Omega \to \overline{\mathbb{R}} : f \text{ is measurable and } \|f\|_p < \infty\}.$$

Theorem 5.7. The map $\|\cdot\|_1$ is a seminorm on $\mathcal{L}^1(\mu)$, that is, for any $f, g \in \mathcal{L}^1(\mu)$ and $\alpha \in \mathbb{R}$,

$$\begin{aligned}\|\alpha f\|_1 &= |\alpha| \cdot \|f\|_1 \\ \|f+g\|_1 &\leq \|f\|_1 + \|g\|_1 \\ \|f\|_1 &\geq 0 \text{ with equality if and only if } f = 0 \text{ almost everywhere.}\end{aligned}$$

Proof. The first and third (in)equalities follow from theorem 5.5(c) and theorem 5.4(a) respectively. The second inequality follows from the fact that $|f+g| \leq |f| + |g|$. We leave the details of the proof to the reader. ■

In fact, $\|\cdot\|_p$ is a seminorm on $\mathcal{L}^p(\mu)$ for any $p \in [1, \infty]$. The proofs of the first and third (in)equalities are similarly straightforward. The proof of the second however requires Minkowski's inequality.

Theorem 5.8. Let $\mu(\Omega) < \infty$ and $1 \leq p' \leq p \leq \infty$. Then $\mathcal{L}^p(\mu) \subseteq \mathcal{L}^{p'}(\mu)$ and further, the canonical inclusion $\mathcal{L}^p(\mu) \hookrightarrow \mathcal{L}^{p'}(\mu)$ given by $f \mapsto f$ is continuous.

Proof. Let us first take the case where $p = \infty$. For any $f \in \mathcal{L}^\infty(\mu)$, since $|f| \leq \|f\|_\infty$ almost everywhere,

$$\int |f|^{p'} \, \mathrm{d}\mu \leq \int \|f\|_\infty^{p'} \, \mathrm{d}\mu = \mu(\Omega) \, \|f\|_\infty^{p'} < \infty.$$

It follows that for any $f, g \in \mathcal{L}^\infty(\mu)$,

$$\|f - g\|_{p'} \leq (\mu(\Omega))^{1/p'} \, \|f - g\|_\infty$$

and so the inclusion map is continuous.
Let us next take the case where p is finite. Then for any $f \in \mathcal{L}^p(\mu)$ since $|f|^{p'} \leq 1 + |f|^p$,

$$\int |f|^{p'} \, \mathrm{d}\mu \leq \int 1 + |f|^p \, \mathrm{d}\mu = \mu(\Omega) + \int |f|^p \, \mathrm{d}\mu < \infty.$$

Now, for any $f, g \in \mathcal{L}^p(\mu)$, let $c = \|f - g\|_p$. Then

$$\begin{aligned} |f-g|^{p'} &= |f-g|^{p'} \mathbb{1}_{|f-g| \leq c} + |f-g|^{p'} \mathbb{1}_{|f-g| > c} \\ &\leq c^{p'} + c^{p'-p} |f-g|^p \end{aligned}$$

This implies that

$$\begin{aligned} \|f-g\|_{p'} &\leq c \left(\int 1 + c^{-p} |f-g|^p \, \mathrm{d}\mu \right)^{1/p'} \\ &= \|f-g\|_p \, (1 + \mu(\Omega))^{1/p'}. \end{aligned}$$

This completes the proof. ■

5.3. Monotone Convergence and Fatou's Lemma

Under what conditions can we exchange the limit and the integral? We answered this question in part in theorem 5.3(b). Over the course of this subsection, we attempt to answer this.

Theorem 5.9 (Monotone Convergence, Beppo-Levi Theorem)**.** Let $f_1, f_2, \ldots \in \mathcal{L}^1(\mu)$ and $f : \Omega \to \overline{\mathbb{R}}$ be measurable. Assume that $f_n \uparrow f$ almost everywhere. Then

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

where both sides can equal ∞.

Proof. Let N be a set such that $\mu(N) = 0$ and $f_n(\omega) \uparrow f(\omega)$ for all $\omega \in N^c$. For each $n \in \mathbb{N}$, define

$$f_n' = (f_n - f_1)\mathbb{1}_{N^c} \text{ and } f' = (f - f_1)\mathbb{1}_{N^c}.$$

Note that $f_n' \uparrow f'$ and each of these functions are non-negative. By theorem 5.3(b), $\int f_n' \, d\mu \uparrow \int f' \, d\mu$. Then by theorem 5.5(a),

$$\begin{aligned}\lim_{n\to\infty} \int f_n \, d\mu &= \lim_{n\to\infty} \left(\int f_n' \, d\mu + \int f_1 \, d\mu \right) \\ &= \int f' \, d\mu + \int f_1 \, d\mu \\ &= \int f \, d\mu.\end{aligned}$$

■

Theorem 5.10 (Fatou's Lemma)**.** Let $f \in \mathcal{L}^1(\mu)$ and let $f_1, f_2, \ldots$ be measurable with $f_n \geq f$ almost everywhere for all $n \in \mathbb{N}$. Then

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

Proof. Considering $f_n - f$ for each $n \in \mathbb{N}$, we may assume each f_n to be non-negative almost everywhere. For each $m \in \mathbb{N}$, consider $g_m = \inf_{n \geq m} f_n$. Note that $g_m \uparrow \liminf_{n\to\infty} f_n$. Thus, using theorem 5.5(a) and theorem 5.9,

$$\int \liminf_{n\to\infty} f_n = \int \lim_{m\to\infty} g_m \, d\mu = \lim_{m\to\infty} \int g_m \, d\mu \leq \liminf_{n\to\infty} \int f_n \, d\mu.$$

■

In the above theorem, we require an integrable f for the statement to hold. This f is called a "minorant".

An example to show this is the following, known as Peterburg's game.
Consider a gamble in a casino where you double your bet with probability $p \leq \frac{1}{2}$ and lose it with probability $1-p$. We gamble over and over. This can be modelled by the probability space $(\Omega, \mathcal{A}, \mathbf{P})$ where $\Omega = \{-1,1\}^{\mathbb{N}}$, $\mathcal{A} = (2^{\{-1,1\}})^{\otimes \mathbb{N}}$, and $\mathbf{P} = ((1-p)\delta_{-1} + p\delta_1)^{\mathbb{N}}$. Let us denote by $D_n : \Omega \to \{-1,1\}$ the output of the nth game
If the player bets b_i dollars on the ith game, then his total profit after the nth game is $S_n \sum_{i=1}^n b_i D_i$. Now, let the gambler assume the following strategy. He bets 1 dollar on the first game. If he wins, he stops playing ($H_n = 0$ for all $n \geq 2$). If he loses, he doubles the bet in the subsequent round.That is,

$$H_n = \begin{cases} 0, & \text{if } D_i = 1 \text{ for some } i < n \\ 1, & \text{otherwise.} \end{cases}$$

The probability of no win until the nth game is $(1-p)^n$.
Therefore, $\mathbf{P}[S_n = 1 - 2^n] = (1-p)^n$ and $\mathbf{P}[S_n = 1] = 1 - (1-p)^n$. The average expected gain is

$$\int S_n \, d\mathbf{P} = (1-2^n)(1-p)^n + (1-(1-p)^n) = 1 - (2(1-p))^n \leq 0.$$

Define

$$S = \begin{cases} -\infty, & \text{if } -1 = D_1 = D_2 = \cdots \\ 1, & \text{otherwise.} \end{cases}$$

Then $\lim_{n\to\infty} S_n = S$ almost surely. However, $\lim_{n\to\infty} \int S_n \, d\mathbf{P} < \lim_{n\to\infty} \int S \, d\mathbf{P}$ since $S = 1$ almost surely.
By Fatou's Lemma, this is only possible if there is no integrable minorant f for $(S_n)_{n\in\mathbb{N}}$.
Indeed, letting $\tilde{S} = \inf\{S_n : n \in \mathbb{N}\}$,

$$\mathbf{P}[\tilde{S} = 1 - 2^{n-1}] = \mathbf{P}[D_1 = D_2 = \cdots = D_{n-1} = -1 \text{ and } D_n = 1] = p(1-p)^{n-1}.$$

Therefore,

$$\int \tilde{S} \, d\mathbf{P} = \sum_{n=1}^{\infty} (1-2^{n-1}) p (1-p)^{n-1} = -\infty.$$

Also, in Fatou's Lemma, we cannot replace the lim inf with a lim sup, which can be seen from the following example. Let X be drawn uniformly randomly from $[0,1]$, and let X_n be the nth binary digit of X (we needn't worry about the case where there is ambiguity in the binary expression as this almost surely does not occur). We have $\mathbf{E}[X_n] = 1/2$ for all n. Then $\limsup_{n\to\infty} X_n$ is almost surely 1, so we cannot replace lim inf with lim sup in Fatou's lemma.

5.4. Miscellaneous

It may be shown that if $f : I \to \mathbb{R}$ is Riemann integrable on $I = [a,b]$, then f is Lebesgue integrable on I with integral

$$\int_I f \,\mathrm{d}\lambda = \int_a^b f(x)\,\mathrm{d}x$$

The converse is not true, that is, a function that is Lebesgue integrable need not be Riemann integrable. An example of this is $\mathbb{1}_{\mathbb{Q}}$.

Theorem 5.11. Let $f : \Omega \to \mathbb{R}$ be measurable and $f \geq 0$ almost everywhere. Then

$$\sum_{n=1}^{\infty} \mu(\{f \geq n\}) \leq \int f \,\mathrm{d}\mu \leq \sum_{n=0}^{\infty} \mu(\{f > n\}).$$

Proof. Define $f_1 = \lfloor f \rfloor$ and $f_2 = \lceil f \rceil$. Clearly, $f_1 \leq f \leq f_2$ and so, $\int f_1 \,\mathrm{d}\mu \leq \int f \,\mathrm{d}\mu \leq \int f_2 \,\mathrm{d}\mu$. We then have

$$\begin{aligned}
\int f_1 \,\mathrm{d}\mu &= \sum_{k=1}^{\infty} k\mu(\{f_1 = k\}) \\
&= \sum_{k=1}^{\infty} \sum_{n=1}^{k} \mu(\{f_1 = k\}) \\
&= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} \mu(\{f_1 = k\}) \\
&= \sum_{n=1}^{\infty} \mu(\{f_1 \geq n\}) \\
&= \sum_{n=1}^{\infty} \mu(\{f \geq n\}).
\end{aligned}$$

Similarly, for the second inequality,

$$\begin{aligned}
\int f_2 \,\mathrm{d}\mu &= \sum_{n=1}^{\infty} \mu(\{f_2 \geq n\}) \\
&= \sum_{n=1}^{\infty} \mu(\{f > n-1\}).
\end{aligned}$$

This proves the result. ∎

Theorem 5.12. Let $f : \Omega \to \mathbb{R}$ be measurable and $f \geq 0$ almost everywhere. Then

$$\int f \,\mathrm{d}\mu = \int_0^{\infty} \mu(\{f \geq t\})\,\mathrm{d}t.$$

Proof. Define g by $g(t) = \mu(\{f \geq t\})$. We may assume that $g(t) < \infty$ for all $t > 0$.
For $\varepsilon > 0$ and $k \in \mathbb{N}$, let $g_\varepsilon = \min(g, g(\varepsilon))$, $f_\varepsilon = f\mathbb{1}_{\{f \geq \varepsilon\}}$, and $f_{\varepsilon,k} = 2^k f_\varepsilon$. and

$$\alpha_{\varepsilon,k} = 2^{-k} \sum_{n=1}^{\infty} \mu(\{f_\varepsilon \geq 2^{-k} n\}).$$

Note that $\alpha_{\varepsilon,k} \xrightarrow{k\to\infty} \int_0^\infty g_\varepsilon \, \mathrm{d}\mu$ (Why?). Using theorem 5.11 on $f_{\varepsilon,k}$,

$$\begin{aligned}
\alpha_{\varepsilon,k} &= 2^{-k} \sum_{n=1}^{\infty} \mu(\{f_{\varepsilon,k} \geq n\}) \leq \int f_\varepsilon \, \mathrm{d}\mu \\
&\leq 2^{-k} \sum_{n=0}^{\infty} \mu(\{f_{\varepsilon,k} > n\}) \\
&= 2^{-k} \sum_{n=0}^{\infty} \mu(\{f_\varepsilon > n2^{-k}\}) \leq \alpha_{\varepsilon,k} + 2^{-k} g(\varepsilon).
\end{aligned}$$

Taking the limit as $k \to \infty$, equality must hold everywhere so we get

$$\int_0^\infty g_\varepsilon \, \mathrm{d}\mu = \int f_\varepsilon \, \mathrm{d}\mu.$$

Taking the limit as $\varepsilon \downarrow 0$ completes the proof. ■

§6. Moments

When describing random variables, quantities like the expectation, median and variance are often used. They describe the behaviour of the variable on average, and how much they deviate from the average.

6.1. Parameters of Random Variables

In the following, let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space.

Definition 6.1. Let X be a real random variable.

(i) If $X \in \mathcal{L}^1(\mathbf{P})$, then X is called integrable and we call

$$\mathbf{E}[X] = \int X \, \mathrm{d}\mathbf{P}$$

the *expectation* or *mean* of X. If $\mathbf{E}[X] = 0$, we call X *centered*. The expectation of X is given by the same expression even if only X^+ or X^- is integrable.

(ii) Let $n \in \mathbb{N}$ and $X \in \mathcal{L}^1(\mathbf{P})$. Then for any $k \in [n]$,

$$m_k = \mathbf{E}[X^k] \text{ and } M_k = \mathbf{E}[|X|^k]$$

are called the *kth moments* and *kth absolute moments* of X respectively.

(iii) If $X \in \mathcal{L}^2(\mathbf{P})$, X is called *square integrable* and

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

is called the *variance* of X. The number $\sigma = \sqrt{\mathbf{Var}[X]}$ is called the *standard deviation* of X (This makes sense as we prove in theorem 6.4(a) that $\mathbf{Var}[X] \geq 0$). We sometimes write $\mathbf{Var}[X] = \infty$ if $\mathbf{E}[X^2] = \infty$.

(iv) If $X, Y \in \mathcal{L}^2(\mathbf{P})$, we define the *covariance of X and Y* by

$$\mathbf{Cov}[X, Y] = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])].$$

X and Y are called *uncorrelated* if $\mathbf{Cov}[X, Y] = 0$ and *correlated* otherwise.

We now give some basic properties of expectation. All of them follow from the corresponding properties of the integral.

The mean represents the average value of the random variable, the variance gives a measure of the deviation or dispersion of the random variable from the mean, and the covariance gives a measure of how related two random variables are.

Theorem 6.1 (Properties of Expectation)**.** Let X, Y, X_n, Z_n $(n \in \mathbb{N})$ be real integrable random variables on $(\Omega, \mathcal{A}, \mathbf{P})$. Then

(a) If $\mathbf{P}_X = \mathbf{P}_Y$, then $\mathbf{E}[X] = \mathbf{E}[Y]$.

(b) For any $c \in \mathbb{R}$, $cX \in \mathcal{L}^1(\mathbf{P})$ and $X + Y \in \mathcal{L}^1(\mathbf{P})$. Further,

$$\mathbf{E}[cX] = c\mathbf{E}[X] \text{ and } \mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y].$$

(c) If $X \geq 0$ almost surely, then $\mathbf{E}[X] = 0$ if and only if $X = 0$ almost surely.

(d) If $X \leq Y$ almost surely, then $\mathbf{E}[X] \leq \mathbf{E}[Y]$. Further, $\mathbf{E}[X] = \mathbf{E}[Y]$ if and only if $X = Y$ almost surely.

(e) $|\mathbf{E}[X]| \leq \mathbf{E}[|X|]$.

(f) If $X_n \geq 0$ almost surely for each $n \in \mathbb{N}$, then $\mathbf{E}[\sum_{i=0}^{\infty} X_i] = \sum_{i=0}^{\infty} \mathbf{E}[X_i]$.

(g) If $Z_n \uparrow Z$ for some Z, then $\mathbf{E}[Z] = \lim_{n\to\infty} Z_n$.

Note that as a consequence of part (b) of the above,

$$\mathbf{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

Setting $X = Y$, we have $\mathbf{Cov}[X, X] = \mathbf{Var}[X]$.

In the above, we did not use anything other than the properties of the integral itself. When we involve probability, namely independence, we get the following result.

Theorem 6.2 (Independent Variables are Uncorrelated)**.** Let $X, Y \in \mathcal{L}^1(\mathbf{P})$ be independent. Then $XY \in \mathcal{L}^1(\mathbf{P})$ and further, $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. Note that this implies $\mathbf{Cov}[X, Y] = 0$.

Proof. Let us first take the case where X and Y each take finitely many values. Then $Z = XY$ takes finitely many values as well and so, $Z \in \mathcal{L}^1(\mathbf{P})$. We now have

$$\begin{aligned}
\mathbf{E}[Z] &= \sum_{z\in\mathbb{R}\setminus\{0\}} z \Pr[Z = z] \\
&= \sum_{x\in\mathbb{R}\setminus\{0\}} \sum_{y\in\mathbb{R}\setminus\{0\}} xy \Pr[X = x, Y = y] \\
&= \left(\sum_{x\in\mathbb{R}\setminus\{0\}} x \Pr[X = x]\right)\left(\sum_{y\in\mathbb{R}\setminus\{0\}} y \Pr[Y = y]\right) = \mathbf{E}[X]\mathbf{E}[Y].
\end{aligned}$$

Now, take the case where X and Y are each non-negative. (They may take an infinite number of values) Using theorem 1.17(a), let $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ be sequences of simple functions such that $X_n \uparrow X$ and $Y_n \uparrow Y$. By theorem 5.9,

$$\begin{aligned}
\mathbf{E}[XY] &= \lim_{n\to\infty} \mathbf{E}[X_n Y_n] \\
&= \lim_{n\to\infty} \mathbf{E}[X_n]\mathbf{E}[Y_n] \\
&= \lim_{n\to\infty} \mathbf{E}[X_n] \lim_{n\to\infty} \mathbf{E}[Y_n] = \mathbf{E}[X]\mathbf{E}[Y].
\end{aligned}$$

The general result where X and Y need not be non-negative is easily proved by splitting X into $X^+ - X^-$ and Y as $Y^+ - Y^-$. ■

The converse however, is not true. That is, uncorrelated variables need not be independent. For example, let X take 0 and 1 with probability $1/2$ each and Y take -1 and 1 with probability $1/2$ each. Then XY and X are uncorrelated but not independent.

Theorem 6.3 (Wald's Identity)**.** Let $T, X_1, X_2, \ldots \in \mathcal{L}^1(\mathbf{P})$ such that $X_1, X_2, \ldots$ are identically distributed and let T take values only in $\mathbb{N}_0$. Define $S_T = \sum_{i=1}^{T} X_i$. Then $S_T \in \mathcal{L}^1(\mathbf{P})$ and $\mathbf{E}[S_T] = \mathbf{E}[T]\mathbf{E}[X_1]$.

Proof. For each $n \in \mathbb{N}$, let $S_n = \sum_{i=1}^{n} X_n$. Then we can write S_T as

$$S_T = \sum_{n=0}^{\infty} \mathbb{1}_{\{T=n\}} S_n.$$

We may prove that $S_T \in \mathcal{L}^1(\mathbf{P})$ by performing the following calculation using $|S_T|$ instead of S_T.

Since $\mathbb{1}_{\{T=n\}}$ and S_n are independent for each $n \in \mathbb{N}$, we now have

$$\begin{aligned}\mathbf{E}[S_T] &= \mathbf{E}\left[\sum_{n=0}^{\infty} \mathbb{1}_{\{T=n\}} S_n\right] \\ &= \sum_{n=0}^{\infty} \mathbf{E}[\mathbb{1}_{\{T=n\}}]\mathbf{E}[S_n] \\ &= \sum_{n=0}^{\infty} \Pr[T=n]\mathbf{E}\left[\sum_{i=0}^{n} X_i\right] \\ &= \sum_{n=0}^{\infty} n \Pr[T=n]\mathbf{E}[X_1]) \\ &= \mathbf{E}[T]\mathbf{E}[X_1].\end{aligned}$$

■

Let us now discuss some properties of the variance.

Theorem 6.4 (Properties of Variance)**.** Let $X \in \mathcal{L}^2(\mathbf{P})$. Then

(a) $\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] \geq 0$,

(b) $\mathbf{Var}[X] = 0$ if and only if $X = \mathbf{E}[X]$ almost surely, and

(c) the map $f : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto \mathbf{E}[(X-x)^2]$ attains its minimum at $x_0 = \mathbf{E}[X]$ taking value $f(x_0) = \mathbf{Var}[X]$.

Proof.

(a) This is clear from the fact that $\mathbf{Cov}[X, X] = \mathbf{Var}[X]$.

(b) This follows as $\mathbf{E}[(X - \mathbf{E}[X])^2] = 0$ if and only if $(X - \mathbf{E}[X])^2 = 0$ almost surely.

(c) Expanding the expression of f as $\mathbf{E}[X^2] + x^2 - 2x\mathbf{E}[X]$, we see that $f(x) = \mathbf{Var}[X] + (x - \mathbf{E}[X])^2$. The result is clear.

■

Theorem 6.5. Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n \in \mathcal{L}^2(\mathbf{P})$ and $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n, d, e \in \mathbb{R}$. Then

$$\mathbf{Cov}\left[d + \sum_{i=1}^{m} \alpha_i X_i, e + \sum_{j=1}^{n} \beta_j Y_j\right] = \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \alpha_i \beta_j \mathbf{Cov}[X_i, Y_j].$$

Proof. By theorem 6.1,

$$\begin{aligned}\mathbf{Cov}\left[d + \sum_{i=1}^{m} \alpha_i X_i, e + \sum_{j=1}^{n} \beta_j Y_j\right] &= \mathbf{E}\left[\left(\sum_{i=1}^{m} \alpha_i (X_i - \mathbf{E}[X_i])\right)\left(\sum_{j=1}^{m} \beta_i (Y_j - \mathbf{E}[Y_j])\right)\right] \\ &= \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \alpha_i \beta_j \mathbf{E}[(X_i - \mathbf{E}[X_i])(Y_j - \mathbf{E}[Y_j])] \\ &= \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \alpha_i \beta_j \mathbf{Cov}[X_i, Y_j].\end{aligned}$$

■

The above can be stated more concisely as:

The map $\mathbf{Cov} : \mathcal{L}^2(\mathbf{P}) \times \mathcal{L}^2(\mathbf{P}) \to \mathbb{R}$ is a positive semidefinite symmetric bilinear form and $\mathbf{Cov}[X, Y] = 0$ if Y is almost surely constant.

Breaking up the individual terms,

- $\mathbf{Cov}[X, Y] = 0$ if Y is almost surely constant means that the d and e in the equation do not contribute to the covariance.
- "Symmetric bilinear form" means that $\mathbf{Cov}[X, Y] = \mathbf{Cov}[Y, X]$, $\mathbf{Cov}[X_1 + X_2, Y] = \mathbf{Cov}[X_1, Y] + \mathbf{Cov}[X_2, Y]$, and $\mathbf{Cov}[\alpha X, Y] = \alpha\,\mathbf{Cov}[X, Y]$ for all $X, Y, X_1, X_2 \in \mathcal{L}^2(\mathbf{P})$ and $\alpha \in \mathbb{R}$.
- "Positive semidefinite" means that $\mathbf{Cov}[X, X] \geq 0$ for all $X \in \mathcal{L}^2(\mathbf{P})$. This is because $\mathbf{Cov}[X, X] = \mathbf{Var}[X] \geq 0$.

Corollary 6.6. For any $X, X_1, \ldots, X_m \in \mathcal{L}^2(\mathbf{P})$ and $\alpha \in \mathbb{R}$, $\mathbf{Var}[\alpha X] = \alpha^2\,\mathbf{Var}[X]$ and

$$\mathbf{Var}\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \mathbf{Var}[X_i] + \sum_{\substack{1 \leq i,j \leq m \\ i \neq j}} \mathbf{Cov}[X_i, X_j].$$

In particular, for uncorrelated $X_1, \ldots, X_m$,

$$\mathbf{Var}\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \mathbf{Var}[X_i].$$

The above is known as the *Bienaymé formula*.

Theorem 6.7 (Cauchy-Schwarz Inequality)**.** Let $X, Y \in \mathcal{L}^2(\mathbf{P})$. Then

$$(\mathbf{Cov}[X, Y])^2 \leq \mathbf{Var}[X]\,\mathbf{Var}[Y].$$

Proof. If $\mathbf{Var}[Y] = 0$, the statement is trivial. If $\mathbf{Var}[Y] \neq 0$,

$$\begin{aligned}
0 &\leq \mathbf{Var}\left[X - \frac{\mathbf{Cov}[X, Y]}{\mathbf{Var}[Y]} Y\right] \mathbf{Var}[Y] \\
&= \left(\mathbf{Var}[X] + \left(\frac{\mathbf{Cov}[X, Y]}{\mathbf{Var}[Y]}\right)^2 \mathbf{Var}[Y] - 2\frac{\mathbf{Cov}[X, Y]}{\mathbf{Var}[Y]}\,\mathbf{Cov}[X, Y]\right) \mathbf{Var}[Y] \\
&= \mathbf{Var}[X]\,\mathbf{Var}[Y] - \mathbf{Cov}[X, Y]^2.
\end{aligned}$$

■

Our choice of $\frac{\mathbf{Cov}[X,Y]}{\mathbf{Var}[Y]}$ in the above is not arbitrary. The term

$$\rho = \frac{\mathbf{Cov}[X, Y]}{\sqrt{\mathbf{Var}[X]\,\mathbf{Var}[Y]}}$$

gives a measure of the correlation of X and Y, and is sometimes called the *correlation coefficient*. It shows how linearly dependent $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ are.
The Cauchy-Schwarz Inequality applied to $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ implies that $-1 \leq \rho \leq 1$. Now, let us try to estimate $Y - \mathbf{E}[Y]$ by a linear combination $a(X - \mathbf{E}[X]) + d$. We shall try to minimize

$$\begin{aligned}
\mathbf{E}[(Y - \mathbf{E}[Y] - c(X - \mathbf{E}[X]) - d)^2] &= \mathbf{Var}[Y] + c^2\,\mathbf{Var}[X] - 2c\,\mathbf{Cov}[X, Y] + d^2 \\
&= \mathbf{Var}[Y] + c^2\,\mathbf{Var}[X] - 2c\rho\sqrt{\mathbf{Var}[X]\,\mathbf{Var}[Y]} + d^2.
\end{aligned}$$

We must clearly take $d = 0$. The minimum of the resulting expression is attained at $c = \rho\sqrt{\frac{\mathbf{Var}[Y]}{\mathbf{Var}[X]}}$, and the minimum expectation is $\mathbf{Var}[Y](1 - \rho^2)$. The closer ρ is to 1, the more linearly dependent $Y - \mathbf{E}[Y]$ and $X - \mathbf{E}[X]$ are. Here,

by linearly dependent, we mean that there exist a_1, a_2 not both 0 such that $a_1(X - \mathbf{E}[X]) + a_2(Y - \mathbf{E}[Y]) = 0$ almost surely.
Further, $|\rho| = 1$ if and only if $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ are linearly dependent.

It follows that equality holds in the Cauchy-Schwarz inequality if and only if there exist $a, b, c \in \mathbb{R}$ not all 0 such that $aX + bY + c = 0$ almost surely.

Theorem 6.8 (Blackwell-Girshick Equation)**.** Let $T, X_1, X_2, \ldots \in \mathcal{L}^1(\mathbf{P})$ such that $X_1, X_2, \ldots$ are identically distributed and let T take values only in $\mathbb{N}_0$. Define $S_T = \sum_{i=1}^{T} X_i$. Then

$$\mathbf{Var}[S_T] = \mathbf{E}[X_1]^2 \mathbf{Var}[T] + \mathbf{E}[T] \mathbf{Var}[X_1].$$

Proof. For each $n \in \mathbb{N}$, let $S_n = \sum_{i=1}^{n} X_n$. Then writing S_T as we did in the proof of Wald's Identity,

$$\begin{aligned}
\mathbf{E}[S_T^2] &= \mathbf{E}\left[\sum_{n=0}^{\infty} \mathbb{1}_{\{T=n\}} S_n\right] \\
&= \sum_{n=0}^{\infty} \mathbf{E}[\mathbb{1}_{\{T=n\}}] \cdot \mathbf{E}[S_n^2] \\
&= \sum_{n=0}^{\infty} \Pr[T = n] \left(\mathbf{Var}[S_n] + \mathbf{E}[S_n]^2\right) \\
&= \sum_{n=0}^{\infty} \Pr[T = n] \left(n \mathbf{Var}[X_1] + n^2 \mathbf{E}[X_1]^2\right) \\
&= \mathbf{E}[T] \mathbf{Var}[X_1] + \mathbf{E}[T^2] \mathbf{E}[X_1]^2.
\end{aligned}$$

Now, Wald's identity implies

$$\begin{aligned}
\mathbf{Var}[S_T] &= \mathbf{E}[T] \mathbf{Var}[X_1] + \mathbf{E}[T^2] \mathbf{E}[X_1]^2 - (\mathbf{E}[T] \mathbf{E}[X_1])^2 \\
&= \mathbf{E}[T] \mathbf{Var}[X_1] + \mathbf{Var}[T] \mathbf{E}[X_1]^2.
\end{aligned}$$

■

We now state the expectations and variances for some of the probability distributions that we mentioned earlier in section 2.2.

1. Let $p \in [0, 1]$ and $X \sim \mathrm{Ber}_p$. Then $\mathbf{E}[X] = p$ and $\mathbf{Var}[X] = p(1 - p)$.
2. Let $p \in [0, 1]$, $n \in \mathbb{N}$ and $X \sim b_{n,p}$. Then $\mathbf{E}[X] = np$ and $\mathbf{Var}[X] = np(1 - p)$. This can easily be shown by noting that X is equal to the sum of n i.i.d. random variables, each of which has distribution Ber_p.
3. Let $\mu \in \mathbb{R}$, $\sigma_2 > 0$ and $X \sim \mathcal{N}_{\mu,\sigma^2}$. Then $\mathbf{E}[X] = \mu$ and $\mathbf{Var}[X] = \sigma^2$.
4. Let $\theta > 0$ and $X \sim \exp_\theta$. Then $\mathbf{E}[X] = \theta^{-1}$ and $\mathbf{Var}[X] = \theta^{-2}$.

6.2. The Weak Law of Large Numbers

We earlier mentioned that variance gives a measure of how much a random variable deviates from the expectation. This is made rigorous by the Chebyshev inequality.

Theorem 6.9 (Markov Inequality)**.** Let X be a real random variable and $f : [0, \infty) \to [0, \infty)$ be monotone increasing. Then for any $\varepsilon > 0$ with $f(\varepsilon) > 0$,

$$\Pr[|X| \geq \varepsilon] \leq \frac{\mathbf{E}[f(|X|)]}{f(\varepsilon)}.$$

Proof. We have

$$\begin{aligned}\mathbf{E}[f(|X|)] &\geq \mathbf{E}[f(|X|)\mathbb{1}_{f(|X|)\geq f(\varepsilon)}] \\ &\geq \mathbf{E}[f(\varepsilon)\mathbb{1}_{f(|X|)\geq f(\varepsilon)}] \\ &\geq f(\varepsilon)\Pr[|X|\geq\varepsilon].\end{aligned}$$

■

The Markov inequality forms a basis for most probability related inequalities we encounter.

Corollary 6.10 (Chebyshev Inequality)**.** Let $X \in \mathcal{L}^2(\mathbf{P})$ and $\varepsilon > 0$. Then

$$\Pr[|X - \mathbf{E}[X]| \geq \varepsilon] \leq \varepsilon^{-2}\,\mathbf{Var}[X].$$

Proof. Using the Markov inequality on $f : x \mapsto x^2$ and $X - \mathbf{E}[X]$ gives the required result. ■

This shows that the variance quantifies how much a random variable deviates from it's expectation.

Definition 6.2 (Convergence of Random Variables)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.

(i) $(X_n)_{n\in\mathbb{N}}$ is said to *converge in probability* towards a random variable X if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr[|X_n - X| > \varepsilon] = 0.$$

(ii) $(X_n)_{n\in\mathbb{N}}$ is said to *converge almost surely* towards a random variable X if

$$\Pr\left[\lim_{n\to\infty} X_n = X\right] = 1.$$

To see that these two are not the same, consider the sequence of random variables $(X_n)_{n\in\mathbb{N}}$ where X_n takes 1 with probability $1/n$ and 0 otherwise. Then $(X_n)_{n\in\mathbb{N}}$ converges in probability to the 0 random variable, but does not converge almost surely.

Given a random variable X, we expect the arithmetic mean of a large number of independent observations of X to be close to $\mathbf{E}[X]$ (assuming, of course, that its variance is finite.).

Making this property more general, we define the following.

Definition 6.3 (Laws of Large Numbers)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real random variables in $\mathcal{L}^1(\mathbf{P})$ and let $\tilde{S}_n = \sum_{i=1}^{n}(X_n - \mathbf{E}[X_n])$.

(i) We say that $(X_n)_{n\in\mathbb{N}}$ satisfies the *weak law of large numbers* if for any $\varepsilon > 0$,

$$\lim_{n\to\infty}\left(\Pr\left[\left|\frac{1}{n}\tilde{S}_n\right| < \varepsilon\right]\right) = 1.$$

(ii) We say that $(X_n)_{n\in\mathbb{N}}$ satisfies the *strong law of large numbers* if

$$\Pr\left[\lim_{n\to\infty}\left|\frac{1}{n}\tilde{S}_n\right| = 0\right] = 1.$$

The weak law is equivalent to saying that $(\tilde{S}_n/n)_{n\in\mathbb{N}}$ converges in probability to the 0 random variable and the strong law is equivalent to saying that $(\tilde{S}_n/n)_{n\in\mathbb{N}}$ converges almost surely to the 0 random variable.

The weak law says that for any nonzero (specified) margin, the average of a sufficiently large number of observations will be within the margin of the expectation with high probability.

On the other hand, the strong law says that almost surely, the sequence of *sample* means converges to the expectation.

Note that as a consequence, if $(X_n)_{n\in\mathbb{N}}$ satisfies the strong law of large numbers, it must also satisfy the weak law of large numbers.

Theorem 6.11. Let $X_1, \ldots, X_n \in \mathcal{L}^2(\mathbf{P})$ be uncorrelated random variables with $V = \sup_{n\in\mathbb{N}} \mathbf{Var}[X_n] < \infty$. Then $(X_n)_{n\in\mathbb{N}}$ satisfies the weak law of large numbers. Further,

$$\Pr\left[\left|\frac{1}{n}\tilde{S}_n\right| \geq \varepsilon\right] \leq \frac{V}{\varepsilon^2 n} \text{ for all } n \in \mathbb{N}.$$

Proof. Without loss of generality, assume $\mathbf{E}[X_i] = 0$ for all $i \in \mathbb{N}$. Then the Bienaymé formula implies that

$$\mathbf{Var}\left[\frac{1}{n}\tilde{S}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{Var}[X_i] \leq \frac{V}{n}.$$

This shows how the more random variables we add, the more it "concentrates" around the mean 0. Now, by Chebyshev's inequality,

$$\Pr\left[\left|\frac{1}{n}\tilde{S}_n\right| \geq \varepsilon\right] \leq \frac{V}{\varepsilon^2 n}.$$

Taking the limit as $n \to \infty$ gives that $(X_n)_{n\in\mathbb{N}}$ satisfies the weak law of large numbers. ■

In general, if we wish to show the existence of an object that has some properties, we give a constructive proof. Equipped with probability now, we have a non-constructive method to prove the existence of something, commonly known as the probabilistic method of proof. We do this by considering some distribution over all objects, and showing that the probability of a randomly selected object having the required properties is non-zero, thus implying the existence of the required object.
An example of this is the following.

Corollary 6.12 (Weirstrass' Approximation Theorem)**.** Let $f : [0,1] \to \mathbb{R}$ be a continuous real map. Then for $n \in \mathbb{N}$ there exist polynomials f_n of degree at most n such that

$$\|f_n(x) - f(x)\|_\infty \xrightarrow{n\to\infty} 0.$$

Proof. For $n \in \mathbb{N}$, define $f_n : [0,1] \to \mathbb{R}$, known as the Bernstein polynomial of order n, by

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

As f is continuous on the compact interval $[0,1]$, it is uniformly continuous. Fixing some $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|f(x) - f(y)| < \varepsilon \text{ for all } x, y \text{ such that } |x - y| < \delta.$$

Now, let $p \in [0,1]$, $X_1, X_2, \ldots$ be i.i.d. random variables with distribution Ber_p, and $S_n = \sum_{i=1}^{n} X_n \sim b_{n,p}$. Note that

$$\mathbf{E}[f(S_n/n)] = f(p).$$

Uniform continuity implies that

$$|f(S_n/n) - f(p)| \leq \varepsilon + 2\|f\|_\infty \mathbb{1}_{|S_n/n - p| \geq \delta}$$

Therefore,

$$\begin{aligned}
|f_n(p) - f(p)| &\leq \mathbf{E}[|f(S_n/n) - f(p)|] \\
&\leq \varepsilon + 2\|f\|_\infty \Pr\left[\left|\frac{S_n}{n} - p\right| \geq \delta\right] \\
&\leq \varepsilon + 2\|f\|_\infty \cdot \frac{p(1-p)}{n\delta^2} \\
&\leq \varepsilon + \frac{\|f\|_\infty}{2n\delta^2}.
\end{aligned}$$

The result follows. ■

Corollary 6.13. Let $n \in \mathbb{N}$ and $p_1, \ldots, p_n \in [0,1]$. Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \sim \text{Ber}_{p_i}$ for each $i \in [n]$. Define $S_n = \sum_{i=1}^n X_i$ and $m = \mathbf{E}[S_n]$. Then for any $\delta > 0$,

$$\Pr[S_n \geq (1+\delta)m] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^m$$

and

$$\Pr[S_n \leq (1-\delta)m] \leq \exp\left(-\frac{\delta^2 m}{2}\right).$$

Proof. For some $\lambda > 0$, consider $f : [0,\infty) \to [0,\infty)$ given by $f(x) = e^{\lambda x}$. By the Markov Inequality, for any $\varepsilon > 0$,

$$\begin{aligned}\Pr[S_n \geq (1+\delta)m] &\leq \frac{\mathbf{E}[e^{\lambda S_n}]}{e^{\lambda m(1+\delta)}} \\ &= e^{-\lambda m(1+\delta)} \prod_{i=1}^n \mathbf{E}[e^{\lambda X_i}]\end{aligned}$$

Now, for each $i \in [n]$,

$$\mathbf{E}[e^{\lambda X_i}] = 1 + p_i(e^\lambda - 1) \leq e^{p_i(e^\lambda - 1)}.$$

Therefore, since $m = p_1 + \cdots + p_n$,

$$\begin{aligned}\Pr[S_n \geq (1+\delta)m] &\leq e^{-\lambda m(1+\delta)} \prod_{i=1}^n e^{p(e^\lambda - 1)} \\ &= e^{-\lambda m(1+\delta)} e^{m(e^\lambda - 1)}.\end{aligned}$$

We can now optimize over λ to find the minimum of the expression on the right. Setting $\lambda = \ln(1+\delta)$, we obtain

$$\Pr[S_n \geq (1+\delta)m] \leq \left(\frac{\delta}{(1+\delta)^{1+\delta}}\right)^m.$$

The second bound can be obtained similarly as we can optimize the following inequality over λ:

$$\begin{aligned}\Pr[X \leq (1-\delta)m] &= \Pr[e^{-\lambda X} \geq e^{-\lambda m(1-\delta)}] \\ &\leq \frac{\mathbf{E}[e^{S_n}]}{e^{-\lambda m(1-\delta)}}.\end{aligned}$$

■

More generally, if X is the sum of n independent random variables $X_1, \ldots, X_n$, then

$$\Pr[X \geq a] \leq \min_{t>0} e^{-ta} \prod_i \mathbf{E}[e^{tX_i}]$$

and

$$\Pr[X \leq a] \leq \min_{t>0} e^{ta} \prod_i \mathbf{E}[e^{-tX_i}].$$

The above inequalities are known as the *Chernoff Bounds*.

6.3. The Strong Law of Large Numbers

There are many versions of the Strong Law of Large Numbers, the one we present here is that given by Etemadi. Namely, if $X_1, X_2, \ldots \in \mathcal{L}^2(\mathbf{P})$ are pairwise independent and identically distributed, then $(X_n)_{n\in\mathbb{N}}$ follows the strong law of large numbers.
Before we prove this, we start with some lemmas.
Define $\mu = \mathbf{E}[X_1]$ and $S_n = X_1 + \cdots + X_n$.

Lemma 6.14. For $n \in \mathbb{N}$, let $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$ and $T_n = Y_1 + \cdots + Y_n$. Then $(X_n)_{n\in\mathbb{N}}$ follows the strong law of large numbers if $T_n/n \xrightarrow{n\to\infty} \mu$ almost surely.

Proof. By theorem 5.11,

$$\sum_{n=1}^{\infty} \Pr[|X_n| > n] \leq \sum_{n=1}^{\infty} \Pr[|X_n| \geq n] \leq \mathbf{E}[X_1] < \infty.$$

Now due to to the Borel-Cantelli Lemma,

$$\Pr[X_n \neq Y_n \text{ for infinitely many } n] = 0.$$

There then almost surely exists some n_0 such that $X_n = Y_n$ for all $n \geq n_0$. Therefore, for all $n \geq n_0$,

$$\frac{T_n - S_n}{n} = \frac{T_{n_0} - S_{n_0}}{n} \xrightarrow{n\to\infty} 0.$$

Since $T_n/n \xrightarrow{n\to\infty} \mu$ almost surely, $S_n/n \xrightarrow{n\to\infty} \mu$ almost surely and our proof is complete. ■

Lemma 6.15. For all $x > 0$, $2x \sum_{n>x} n^{-2} \leq 4$.

Proof. For any $m \in \mathbb{N}$,

$$\sum_{n=m}^{\infty} n^{-2} \leq m^{-2} + \int_m^{\infty} t^{-2}\,\mathrm{d}t = m^{-2} + m^{-1} \leq \frac{2}{m}.$$

The result follows. ■

Lemma 6.16. With the above notation,

$$\sum_{n=1}^{\infty} \frac{\mathbf{E}[Y_n^2]}{n^2} \leq 4\mathbf{E}[|X_1|].$$

Proof. By theorem 5.12,

$$\mathbf{E}[Y_n^2] = \int_0^{\infty} \Pr(Y_n^2 \geq t)\,\mathrm{d}t.$$

Simplifying by substituting $t = x^2$,

$$\mathbf{E}[Y_n^2] = \int_0^{\infty} 2x \Pr[|Y_n| \geq x]\,\mathrm{d}x \leq \int_0^{n} 2x \Pr[|X_1| \geq x]\,\mathrm{d}x.$$

For $m \in \mathbb{N}$, define the function f_m by

$$f_m(x) = \left(\sum_{n=1}^{m} n^{-2} \mathbb{1}_{\{x<n\}}\right) 2x \Pr[|X_1| > x] \cdot .$$

Let $f_m \uparrow f$. Note that $f \leq 4 \Pr[|X_1| > x]$.
Finally, by the Monotone Convergence Theorem,

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbf{E}[Y_n^2]}{n^2} &\leq \sum_{n=1}^{\infty} n^{-2} \int_0^{n} 2x \Pr[|X_1| \geq x]\,\mathrm{d}x \\
&= \int_0^{\infty} \left(\sum_{n=1}^{\infty} n^{-2} \mathbb{1}_{\{x<n\}}\right) 2x \Pr[|X_1| > x] \\
&\leq 4 \int_0^{\infty} \Pr[|X_1| > x] = 4\mathbf{E}[|X_1|].
\end{aligned}$$

■

We may now prove the main result of this subsection.

Theorem 6.17 (Etemadi's Strong Law of Large Numbers[1])**.** Let $X_1, X_2, \ldots \in \mathcal{L}^1(\mathbf{P})$ be pairwise independent and identically distributed random variables. If $\mathbf{E}[|X_1|] < \infty$, then $(X_n)_{n\in\mathbb{N}}$ satisfies the strong law of large numbers.

Proof. Since $(X_n^+)_{n\in\mathbb{N}}$ and $(X_n^-)_{n\in\mathbb{N}}$ satisfy the conditions of the theorem, it suffices to consider only $(X_n^+)_{n\in\mathbb{N}}$, that is, $X_n \geq 0$ for all $n \in \mathbb{N}$.
Fix some $\varepsilon > 0$ and $\alpha > 1$ and let $k_n = \lfloor \alpha^n \rfloor$ for each n. Then by Chebyshev's inequality,

$$\begin{aligned}
\sum_{n=1}^{\infty} \Pr\left[\left|\frac{T_{k_n} - \mathbf{E}[T_{k_n}]}{k_n}\right| > \varepsilon\right] &\leq \varepsilon^{-2} \sum_{n=1}^{\infty} \frac{\mathbf{Var}[T_{k_n}]}{k_n^2} \\
&= \varepsilon^{-2} \sum_{n=1}^{\infty} \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathbf{Var}[Y_i] \\
&= \varepsilon^{-2} \sum_{i=1}^{\infty} \mathbf{Var}[Y_i] \sum_{n:k_n \geq m} \frac{1}{k_n^2} \\
&\leq \frac{1}{\varepsilon^2(1-\alpha^{-2})} \sum_{i=1}^{\infty} \frac{\mathbf{Var}[Y_i]}{m^2} = \frac{1}{\varepsilon^2(1-\alpha^{-2})} \sum_{i=1}^{\infty} \frac{\mathbf{E}[Y_i^2]}{m^2}
\end{aligned}$$

By lemma 6.16, this is finite. Letting $\varepsilon \downarrow 0$, the Borel-Cantelli Lemma implies that

$$\lim_{n\to\infty} \frac{T_{k_n} - \mathbf{E}[T_{k_n}]}{k_n} = 0 \text{ almost surely.}$$

By the Monotone Convergence Theorem, $\mathbf{E}[Y_n] \xrightarrow{n\to\infty} \mathbf{E}[X_1]$.
This in turn gives that $\mathbf{E}[T_{k_n}]/k_n \xrightarrow{n\to\infty} \mathbf{E}[X_1]$ and therefore,

$$\lim_{n\to\infty} \frac{T_n}{n} = \lim_{n\to\infty} \frac{T_{k_n}}{k_n} = \mathbf{E}[X_1] \text{ almost surely.}$$

Lemma 6.14 completes the proof. ∎

References

[1] N. Etemadi. An elementary proof of the strong law of large numbers. *Z. Wahrscheinlichkeitstheorie verw Gebiete*, 55 (1):119–122, 1981. doi:10.1007/BF01013465.

[2] Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, 2 edition, 2006.