# STK4500: Life Insurance and Finance

Basics on probability

# Chapter 1

# Measures

## 1.1 $\sigma$-algebras

Let $\Omega$ be a nonempty set and $\mathcal{P}(\Omega) = \{A : A \subset \Omega\}$ be the *power set of* $\Omega$, i.e., the class of all subsets of $\Omega$.

**Definition 1.1.1.** *A collection of sets $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called an* algebra *if*

*(a)* $\Omega \in \mathcal{F}$,

*(b)* $A \in \mathcal{F}$ *implies* $A^c = \Omega \setminus A \in \mathcal{F}$,

*(c)* $A, B \in \mathcal{F}$ *implies* $A \cup B \in \mathcal{F}$.

Hence, an algebra of subsets of a given set $\Omega$ is a class of sets containing $\Omega$ that is closed under complementation and pairwise (and hence finite) unions. It is easy to see that one can equivalently exchange property $(c)$ by

$$(c)' \; A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}.$$

**Definition 1.1.2.** *A class $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called a $\sigma$-algebra if it an algebra and satisfies in addition*

$$(d) \; A_n \in \mathcal{F} \text{ for } n \geqslant 1 \Rightarrow \bigcup_{n \geqslant 1} A_n \in \mathcal{F}.$$

Thus, a $\sigma$-algebra is a class of subsets of $\Omega$ that contains $\Omega$ and is closed under complementation and countable unions

The concept of $\sigma$-algebra is motivated by the concept of *event* in an experiment. Then a $\sigma$-algebra of events gives intuition to all possible events one can observe in a given experiment.

**Example 1.1.3.** *We consider tossing a coin twice. Then $\Omega = \{hh, ht, th, tt\}$ is our sample space, describing the outcomes of our experiment. The $\sigma$-algebra on $\Omega$ for this experiment is a set of 16 sets. namely,*

$$\mathcal{F} = \Big\{\emptyset, \{hh\}, \{ht\}, \{th\}, \{tt\}, \{hh, ht\}, \{hh, th\}, \{hh, tt\}, \{ht, th\}, \{hh, tt\},$$
$$\{th, tt\}, \{hh, ht, th\}, \{hh, ht, tt\}, \{hh, th, tt\}, \{ht, th, tt\}, \Omega\Big\}$$

**Example 1.1.4.** *If $\Omega = \mathbb{R}$ then an example of $\sigma$-algebra on $\Omega$ is given by $\mathcal{F} = \mathcal{B}(\mathbb{R})$ where $\mathcal{B}(\mathbb{R})$ denotes the Borel $\sigma$-algebra of $\mathbb{R}$, which is given by all sets $(a,b), [a,b], (a,b]$ and $[a,b)$ for $a, b \in \mathbb{R}$, $a < b$, $\emptyset$ and $\mathbb{R}$.*

If $\mathcal{A}$ is a set of subsets of $\Omega$, then $\mathcal{A}$ is in general not a $\sigma$-algebra. But we can define the smallest $\sigma$-algebra containing $\mathcal{A}$.

**Definition 1.1.5.** *If $\mathcal{A}$ is a class of subsets of $\Omega$, the the $\sigma$-algebra generated by $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$, is defined as*

$$\sigma(\mathcal{A}) = \bigcap_{\mathcal{F} \in \mathcal{I}(\mathcal{A})} \mathcal{F},$$

*where $\mathcal{I}(\mathcal{A}) = \{\mathcal{F} : \mathcal{A} \subset \mathcal{F}$ and $\mathcal{F}$ is a $\sigma$-algebra on $\Omega\}$ is the collection of all $\sigma$-algebras containing $\mathcal{A}$.*

Note that since the power $\mathcal{P}(\mathcal{A})$ is itself a $\sigma$-algebra containing $\mathcal{A}$, the collection $\mathcal{I}(\mathcal{A})$ is not empty and hence, the above intersection is well defined.

**Example 1.1.6.** *If $\Omega = \{a,b,c\}$ then $\mathcal{A}_1 = \{\{a\},\{c\}\}$ and $\mathcal{A}_2 = \{\{a,b\}\}$ are both a collection of subsets of $\Omega$ but not a $\sigma$-algebra.*

*The $\sigma$-algebras of $\Omega$ are:*

$$\begin{aligned}
\mathcal{F}_1 &= \{\emptyset, \Omega\} \\
\mathcal{F}_2 &= \{\mathcal{F}_1, \{a\}, \{b\}, \{c\}, \{a,b\}, \{b,c\}, \{a,c\}\} \\
\mathcal{F}_3 &= \{\mathcal{F}_1, \{a\}, \{b,c\}\} \\
\mathcal{F}_4 &= \{\mathcal{F}_1, \{b\}, \{a,c\}\} \\
\mathcal{F}_5 &= \{\mathcal{F}_1, \{c\}, \{a,b\}\}.
\end{aligned}$$

*Then*

$$\mathcal{I}(\mathcal{A}_1) = \{\mathcal{F}_2\}, \quad \mathcal{I}(\mathcal{A}_2) = \{\mathcal{F}_2, \mathcal{F}_5\}.$$

*Hence,*

$$\sigma(\mathcal{A}_1) = \mathcal{F}_2 = \mathcal{P}(\Omega), \quad \sigma(\mathcal{A}_2) = \mathcal{F}_2 \cap \mathcal{F}_5 = \mathcal{F}_5.$$

*Usually on a finite set one can construct the corresponding minimal $\sigma$-algebra by adding complements and then adding unions and complements of the obtained sets.*

## 1.2 Metric spaces and the Borel $\sigma$-algebra

A *metric space* is a pair $(M,d)$ where $M$ is a nonempty set and $d$ is a function $d : M \times M \to [0,\infty)$ satisfying

(i) $d(x,y) = d(y,x)$ for all $x, y \in M$,

(ii) $d(x,y) = 0$ if, and only if $x = y$ for all $x, y \in M$,

(iii) $d(x,y) \leqslant d(x,z) + d(z,y)$ for all $x, y, z \in M$ (triangle inequality).

The function $d$ is called a metric on $M$ or a more commonly, a distance.

Any Euclidean space $\mathbb{R}^n$, for an integer $n \geqslant 1$, is a metric space under any of the following metrics:

(a) For $1 \leqslant p < \infty$, $d_p(x,y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$,

(b) $d_\infty(x,y) = \max_{i=1,\dots,n} |x_i - y_i|$,

(c) $0 < p < 1$, $d_p(x,y) = \sum_{i=1}^n |x_i - y_i|^p$.

Given an element $x \in M$, we define open neighbourhoods of $x$ of the form $B(x,\varepsilon) = \{y \in M : d(x,y) < \varepsilon\}$. These are often called balls. We say that a set $U \subset M$ is open if for all $x \in U$ there is $\varepsilon > 0$ such that $B(x,\varepsilon) \subset U$. In this way, a metric space is also a topological space where the topology $\tau$ is the collection of all open sets as defined above. This gives rise to the following definition.

**Definition 1.2.1.** *The Borel $\sigma$-algebra on a topological space $M$ (in particular, on a metric space or an Euclidean space) is defined as the $\sigma$-algebra generated by the collection of open sets in $M$.*

**Example 1.2.2.** *Let $\mathcal{B}(\mathbb{R}^n)$ denote the Borel $\sigma$-algebra on $\mathbb{R}^n$. Then,*

$$\mathcal{B}(\mathbb{R}^n) = \sigma\left(\{U : U \text{ is an open subset of } \mathbb{R}^n\}\right)$$

*is also generated by each of the following sets*

$$\begin{aligned}
U_1 &= \{(a_1,b_1) \times \cdots \times (a_n,b_n) : -\infty \leqslant a_i < b - i \leqslant \infty, 1 \leqslant i \leqslant n\}, \\
U_2 &= \{(-\infty,x_1) \times \cdots \times (-\infty,x_n) : x_1,\dots,x_n \in \mathbb{R}\}, \\
U_3 &= \{(a_1,b_1) \times \cdots \times (a_n,b_n) : a_i,b_i \in \mathbb{Q}, a_i < b_i, 1 \leqslant i \leqslant n\}, \\
U_4 &= \{(-\infty,x_1) \times \cdots \times (-\infty,x_n) : x_1,\dots,x_n \in \mathbb{Q}\}.
\end{aligned}$$

## 1.3 Measures

A *set function* is an extended real valued function defined on a class of subsets of a set $\Omega$. Measures are nonnegative set functions acting on $\sigma$-algebras on $\Omega$. Intuitively speaking, such function, measures the content of a subset of $\Omega$. A measure has to satisfy certain natural conditions.

**Definition 1.3.1.** *Let $\Omega$ be a nonempty set and $\mathcal{F}$ be a $\sigma$-algebra on $\Omega$. Then a set function $\mu$ on $\mathcal{F}$ is called a* measure *if*

*(i) Nonnegativity: $\mu(A) \in [0,\infty]$ for all $A \in \mathcal{F}$,*

*(ii) Null empty set: $\mu(\emptyset) = 0$,*

*(iii) Countable additivity ($\sigma$-additivity): for any countable collection $\{A_n\}_{n\geqslant 1}$ of pairwise disjoints sets in $\mathcal{F}$,*

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

.

**Definition 1.3.2.** *A measure $\mu$ is called* finite *or* infinite *according to whether $\mu(\Omega) < \infty$ or $\mu(\Omega) = \infty$. A finite measure with $\mu(\Omega) = 1$ is called a* probability measure. *A measure $\mu$ is called $\sigma$-finite if there exists a countable collection $\{A_n\}_{n\geqslant 1}$ of not necessarily disjoints sets in $\mathcal{F}$ such that*

$$\bigcup_{n\geqslant 1} A_n = \Omega, \quad \mu(A_n) < \infty \text{ for all } n \geqslant 1.$$

**Example 1.3.3.** *Here are some examples:*

- *(The counting measure) Let $\Omega$ be a nonempty set and $\mathcal{F} = \mathcal{P}(\Omega)$ be the set of all subsets of $\Omega$. Define $\mu(A) = |A|$, $A \in \mathbb{F}$ where $|A|$ denotes the number of elements in $A$. It is easy to check that $\mu$ is a measure. Note that $\mu$ is finite if and only if $\Omega$ is finite and $\sigma$-finite if $\Omega$ is countably infinity.*

- *(Discrete probability measures) Let $\omega_1, \omega_2, \cdots \in \Omega$ and $p_1, p_2, \cdots \in [0,1]$ such that $\sum_{i=1}^{\infty} p_i = 1$. Define for any $A \subset \Omega$:*

$$P(A) = \sum_{i=1}^{\infty} p_i I_A(\omega_i),$$

*where $I_A(\cdot)$ denotes the indicator function of a set $A$. For any disjoint collection of sets $A_1, A_2, \cdots \in \mathbb{P}(\Omega)$,*

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{j=1}^{\infty} p_j I_{\cup_{i=1}^{\infty} A_i}(\omega_j) \\ &= \sum_{j=1}^{\infty} p_j \left(\sum_{i=1}^{\infty} I_{A_i}(\omega_j)\right) \\ &= \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} p_j I_{A_i}(\omega_j)\right) \\ &= \sum_{i=1}^{\infty} P(A_i), \end{aligned}$$

*where interchanging the order of summation is allowed since summands are nonnegative (Tonelli). Furthermore*

$$P(\Omega) = \sum_{i=1}^{\infty} p_i I_{\Omega}(\omega_i) = \sum_{i=1}^{\infty} p_i = 1.$$

*All this shows that $P$ is a probability measure on $\mathcal{P}(\Omega)$.*

- *(Lebesgue-Stieltjes measures on $\mathbb{R}$) A large class of measures on the Borel $\sigma$-algebra $\mathcal{B}(\mathbb{R})$ of subsets of $\mathbb{R}$, known as Lebesgue-Stieltjes measures, arise from non decreasing right continuous functions $F : \mathbb{R} \to \mathbb{R}$. For such a given $F$ the corresponding measure $\mu_F$ satisfies $\mu_F((a,b]) = F(b) - F(a)$ for all $-\infty < a < b < \infty$, then one can construct $\mu_F$ on any Borel sets via extensions theorems (too technical). Note that if $A_n = (-n, n)$,*

*$n = 1, 2, \ldots$ then $\mathbb{R} = cup_{n\geqslant 1} A_n$ and $\mu_F(A_n) < \infty$ for every $n \geqslant 1$ and thus $\mu_F$ are $\sigma$-finite (this is good to define integration theory).*

*Observe that if in addition $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$ then $\mu_F$ is called a law which, in particular,*

$$\mu_F((-\infty, x]) = F(x)$$

*defines a distribution. Integration of the Lebesgue-Stieltjes type is in particular used for defining expectations of random variables.*

*If $F(x) = x$ then $\mu_F$ is simply called the Lebesgue measure.*

**Proposition 1.3.4.** *Let $\mu$ be a measure on a $\sigma$-algebra $\mathcal{F}$,*

*(i) (Monotonicity) $\mu(A) \leqslant \mu(B)$ if $A \subseteq B$,*

*(ii) ($\sigma$-subadditivity) For any countable collection of events $\{A_n\}_{n\geqslant 1}$ not necessarily disjoints,*

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

*(iii) (Inclusion-exclusion formula) If $A_1, \ldots, A_n$ $1 \leqslant n < \infty$ and $\mu(A_n) < \infty$ for all $n$ then*

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_n) - \sum_{i\leqslant i<j<n} \mu(A_i \cap A_j) + \cdots + (-1)^{n-1} \mu\left(\bigcap_{i=1}^{n} A_i\right)$$

*(iv) (Monotone continuity from above and below) Let $\{A_n\}_{n\geqslant 1}$ be a sequence of sets in $\mathcal{F}$ such that $A_{n+1} \subseteq A_n$ for all $n \geqslant 1$. Also, $\mu(A_{n_0}) < \infty$ for some $n_0$. Then, the intersection of sets is measurable and*

$$\mu\left(\bigcap_{n\geqslant 1} A_n\right) = \lim_{n\to\infty} \mu(A_n).$$

*Similarly, let $\{A_n\}_{n\geqslant 1}$ be a sequence of sets in $\mathcal{F}$ such that $A_n \subseteq A_{n+1}$ for all $n \geqslant 1$ then the union of sets is measurable and*

$$\mu\left(\bigcup_{n\geqslant 1} A_n\right) = \lim_{n\to\infty} \mu(A_n).$$

# Chapter 2

# Probability spaces

## 2.1 Kolmogorov's axiomatic probability model

Probability theory provides a mathematical model for random phenomena, i.e., those involving uncertainty. First, one identifies the set $\Omega$ of possible outcomes of a (random) experiment associated with the phenomenon. This set $\Omega$ is called *sample space*, and an individual element $\omega$ of $\Omega$ is called a *sample point*. Even though the outcome is not predictable ahead of time, one is interested in the chances of some particular statement to be valid for the resulting outcome. The set of $\omega$'s for which a given statement is valid is called an *event*. Thus, an event is a subset of $\Omega$. One then identifies a class $\mathcal{F}$ of events, i.e., a class $\mathcal{F}$ of subsets of $\Omega$ (not necessarily $\mathcal{P}(\Omega)$), and then a set function $P$ on $\mathcal{F}$ such that for $A$ in $\mathcal{F}$, $P(A)$ represents the "chance" of the event $A$ happening. Thus, it is reasonable to impose the following conditions on $\mathcal{F}$ and $P$:

(i) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.

(ii) $A_1, A_2 \in \mathcal{F} \Rightarrow A_1 \cup A_2 \in \mathcal{F}$.

(iii) for all $A \in \mathcal{F}$, $0 \leqslant P(A) \leqslant 1$, $P(\emptyset) = 0$ and $P(\Omega) = 1$.

(iv) $A_1, A_2 \in \mathcal{F}$, $A_1 \cap A_2 = \emptyset$ implies $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

(v) $A_n \in \mathcal{F}$, $A_n \subset A_{n+1}$ for all $n = 1, 2, \ldots$ then $\cup_{n \geqslant 1} A_n \in \mathcal{F}$ and $\lim_n P(A_n) = P(\cup_n A_n)$.

Conditions $(i) - (v)$ imply that $(\Omega, \mathcal{F}, P)$ is a *measure space*, i.e., $\mathcal{F}$ is a $\sigma$-algebra and $P$ is a measure on $\mathcal{F}$ with $P(\Omega) = 1$. That is, $(\Omega, \mathcal{F}, P)$ is a *probability space*. Here are some examples.

**Example 2.1.1** (Finite sample spaces)**.** *Let $\Omega = \{\omega_1, \ldots, \omega_k\}$ for some integer $k \geqslant 1$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $P(A) = \sum_{i=1}^k p_i I_A(\omega_i)$ where $\{p_i\}_{i=1}^k$ are such that $0 \leqslant p_i \leqslant 1$ and $\sum_{i=1}^k p_i = 1$. This is a probability model for random experiments with finitely many possible outcomes.*

*An important application of this probability model is finite population sampling. Let $\{U_1, \ldots, U_N\}$ be a finite populations of $N$ units or objects. These could be individuals in a city, counties in a state, etc. In a typical sample survey procedure, one chooses a subset of size $n$, $1 \leqslant n \leqslant N$, from this population. Let $\Omega$ denote the collection of all possible subsets of size $n$. Here $k = \binom{N}{n}$,*

*each $\omega_i$ is a sample of size $n$ and $p_i$ is the selection probability of $\omega_i$. The assignment of $\{p_i\}_{i=1}^k$ is determined by a given* sampling scheme. *For example, in simple random sampling without replacement, $p_i = \frac{1}{k}$ for $i = 1, \ldots, k$.*

**Example 2.1.2** (Countably infinite sample spaces)**.** *Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be a countable set, $\mathcal{F} = \mathcal{P}(\Omega)$, and $P(A) = \sum_{i=1}^{\infty} p_i I_A(\omega_i)$ where $\{p_i\}_{i=1}^{\infty}$ are such that $0 \leqslant p_i \leqslant 1$ and $\sum_{i=1}^{\infty} p_i = 1$. It is easy to verify that $(\Omega, \mathcal{F}, P)$ is a probability space. This is a probability model for random experiments with countably infinite number of outcomes. For example, the experiment of tossing a coin until a "head" is produced is such a probability space.*

**Example 2.1.3** (Uncountable sample spaces)**.** *Here are some examples of uncountable sample spaces:*

*(a) (Random variables). Let $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, $P = \mu_F$, the Lebesgue-Stieltjes measure corresponding to a cdf $F$, i.e., corresponding to a function $F : \mathbb{R} \to \mathbb{R}$ that is nondecreasing, right-continuous and satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$. This serves as a model for a single random variable $X$.*

*(b) (Random vectors). Let $\Omega = \mathbb{R}^k$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^k)$, $P = \mu_F$, the Lebesgue-Stieltjes measure corresponding to a (multidimensional) cdf $F$ on $\mathbb{R}^k$.*

*(c) (Random sequences). Let $\Omega = \mathbb{R}^\infty$ the set of all sequences $\{x_n\}_{n\geqslant 1}$ of real numbers. Let $\mathcal{C}$ be the class of all finite dimensional sets of the form $A\times\mathbb{R}\times\cdots$, where $A \in \mathcal{B}(\mathbb{R}^k)$ for some $1 \leqslant k < \infty$, let $\mu_k$ be a probability measure on $\mathcal{B}(\mathbb{R}^k)$ such that $\mu_{k+1}(A \times \mathbb{R}) = \mu_k(A)$ for all $A \in \mathcal{B}(\mathbb{R}^k)$ and $k$. This will be the model for a sequence $\{X_n\}_{n\geqslant 1}$ of random variables such that for*

## 2.2 Random variables and random vectors

Random variables and vector are the central objects of this course.

**Definition 2.2.1** (Random variable)**.** *Let $(\Omega, ,F, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ be a measurable set function, i.e. $X^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{B}(\mathbb{R})$. Then, $X$ is called a* random variable *on $(\Omega, ,F, P)$.*

Recall that, $X$ is measurable if, and only if for all $x \in \mathbb{R}$, $\{\omega : X(\omega) \leqslant x\} \in \mathcal{F}$.

**Definition 2.2.2** (Law/ probability distribution of a random variable)**.** *Let $X$ be a random variable on $(\Omega, ,F, P)$. Let*

$$P_X(A) = P(X^{-1}(A)), \quad A \in \mathcal{F}.$$

*Then, the probability measure $P_X$ is called the* probability distribution *of the* law *of $X$.*

**Definition 2.2.3** (Cumulative distribution function)**.** *Let $X$ be a random variable on $(\Omega, ,F, P)$. Let*

$$F_X(x) = P(\{\omega : X(\omega) \leqslant x\}), \quad x \in \mathbb{R}.$$

*Then, $F_X$ is called the* cumulative distribution function *(cdf) of $X$.*

Observe that $F_X$ is a function and $P_X$ is a set function. Also, if$A_x = (-\infty, x]$ for $x \in \mathbb{R}$ then $P_X(A_x) = F_X(x)$.

(add generalization to vectors from page 192).

It is clear that the distribution of $X = (X_1, \ldots, X_k)$ determines the marginal distribution $P_{X_i}$ of $X_i$ for all $i = 1, \ldots, k$. However, the marginal distributions $\{P_{X_i} : i = 1, \ldots, k\}$ do not uniquely determine the joint distribution $P_X$, without additional conditions, such as independence.

**Definition 2.2.4** (Expected value)**.** *Let $X$ be a random variable on $(\Omega, ,F, P)$. The* expected value *ofX, denoted by $EX$ or $E[X]$, is defined as*

$$E[X] = \int_\Omega X(\omega) P(d\omega),$$

*provided the integral is well defined. That is, at least one of the two quantities $\int X^+ dP$ and $\int X^- dP$ is finite.*

**Remark 2.2.5.**

- *If $h$ is a Borel measurable function then*

$$E[h(X)] = \int_\Omega h(X(\omega)) P(d\omega).$$

- *The case $h = \mathbf{1}_A$ for some Borel set $A \in \mathbb{R}$ then*

$$E[\mathbf{1}_A] = \int_\Omega \mathbf{1}_A(\omega) P(d\omega) = P(A)$$

- *If $X$ has distribution function $F_X$ then*

$$E[h(X)] = \int_{\mathbb{R}} h(x) F_X(dx).$$

  *In particular, if $F_X(x) = P(X \leqslant x)$ is absolutely continuous and $f_X$ is its absolutely continuous derivative then*

$$E[h(X)] = \int_{\mathbb{R}} h(x) f_X(x) dx.$$

**Proposition 2.2.6** (Change of variables formula)**.** *Let $X$ be an absolutely continuous random variable with density function $f_X$ defined over the support $[a, b]$. Let $Y = h(X)$ be an invertible function of $X$ with inverse denoted by $h^{-1}$, i.e. $X = h^{-1}(Y)$. Then the probability density function of $Y$ is given by*

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right|,$$

*defined over the support $[u(a), u(b)]$.*

**Definition 2.2.7.** *For any positive integer $n$, the nth moment $\mu_n$ of a random variable $X$ is defined by*

$$\mu_n = E[X^n],$$

*provided the expectation is well defined.*

**Definition 2.2.8.** *The variance of a random variable $X$ is defined as $Var[X] = E[(X-E[X])^2]$, provided $E[X^2] < \infty$. The variance is a measure of distance or deviation to the mean.*

**Definition 2.2.9** (Moment generating function (mgf))**.** *The* moment generating function (mgf) *of a random variable $X$ is defined by*

$$M_X(t) = E[e^{tX}], \quad t \in \mathbb{R}.$$

Since $e^{tX}$ is always almost surely nonnegative, $M_X$ is always well defined but could be infinity.

**Proposition 2.2.10.** *Let $X$ be a nonnegative random variable and $t \geqslant 0$. Then*

$$M_X(t) = E[e^{tX}] = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mu_n.$$

**Proposition 2.2.11.** *Let $X$ be a random variable and let $M_X(t)$ be finite for all $|t| < \varepsilon$ for some $\varepsilon > 0$. Then*

*(i)* $E|X|^n < \infty$ *for all* $n \geqslant 1$,

*(ii)* $M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mu_n$ *for all* $|t| < \varepsilon$,

*(iii)* $M_X$ *is infinitely differentiable on* $(-\varepsilon, \varepsilon)$ *and for* $k \in \mathbb{N}$ *the* $k$*-th derivative of* $M_X$ *is*

$$M_X^{(k)}(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mu_{n+k} = E[e^{tX}X^k], \quad |t| < \varepsilon.$$

*In particular,*

$$M_X^{(k)}(0) = \mu_k = E[X^k].$$

**Theorem 2.2.12.** *If two random variables have the same moment generating function, then they have the same distribution. That is*

$$M_X(t) = M_Y(t) \Rightarrow X \sim Y.$$

**Remark 2.2.13.** *If $M_X(t)$ is finite for $|t| < \varepsilon$ for some $\varepsilon > 0$, then all the moments $\{\mu_n\}_{n\geqslant 0}$ of $X$ are determined and also its probability distribution. However, in general, the sequence $\{\mu_n\}_{n\geqslant 0}$ of moments of $X$ need not determine the distribution of $X$ uniquely.*

**Definition 2.2.14** (Characteristic function)**.** *Let $X$ be a random variable. The* characteristic function *(cf) of $X$ is defined as*

$$\varphi_X(t) = E[e^{\boldsymbol{i}tX}], \quad t \in \mathbb{R}.$$

**Theorem 2.2.15.** *Two random variables (not necessarily on the same probability space) have the same law if, and only if they have the same characteristic function.*

## 2.3 Some elementary inequalities

**Proposition 2.3.1** (Markov's inequality). *Let $X$ be a random variable on $(\Omega, \mathcal{F}, P)$. Then for any $\phi : [0,\infty) \to [0,\infty)$ nondecreasing and any $t > 0$ with $\phi(t) > 0$,*

$$P(|X| > t) \leqslant \frac{E[\phi(|X|)]}{\phi(t)}.$$

*In particular,*

$$P(|X| > t) \leqslant \frac{E[|X|^k]}{t^k},$$

$$P(|X| > t) \leqslant \frac{E[e^{\alpha|X|}]}{e^{\alpha t}},$$

*for every $\alpha > 0$, and hence,*

$$P(|X| > t) \leqslant \inf_{\alpha>0} \frac{E[e^{\alpha|X|}]}{e^{\alpha t}},$$

**Proposition 2.3.2** (Chebyshev's inequality). *Let $X$ be a random variable with $EX^2 < \infty$, $EX = \mu$, $VarX = \sigma^2$. Then for any $k > 0$,*

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}.$$

**Proposition 2.3.3** (Jensen's inequality). *Let $X$ be a random variable with $P(a < X < b) = 1$ for $-\infty \leqslant a < b \leqslant \infty$. Let $\phi : (a,b) \to \mathbb{R}$ be convex on $(a,b)$. Then*

$$\phi(EX) \leqslant E\phi(X)$$

*provided $E|X| < \infty$ and $E|\phi(X)| < \infty$.*

**Proposition 2.3.4** (Hölder's inequality). *Let $X$ and $Y$ be random variables on $(\Omega, \mathcal{F}, P)$ with $E|X|^p < \infty$, $E|Y|^q < \infty$, $1 < p < \infty$, $1 < q < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$E|XY| \leqslant (E|X|^p)^{1/p}(E|Y|^q)^{1/q},$$

*with equality if, and only if $P(a|X|^p = b|Y|^q) = 1$ for some $0 \leqslant a, b < \infty$. In particular, if $p = q = 2$ we have*

$$E|XY| \leqslant \sqrt{EX^2}\sqrt{EY^2},$$

*or also,*

$$Cov(X,Y) \leqslant \sqrt{VarX}\sqrt{VarY}.$$

**Proposition 2.3.5** (Minkowski's inequality). *Let $X$ and $Y$ be random variables on $(\Omega, \mathcal{F}, P)$ with $E|X|^p < \infty$, $E|Y|^p < \infty$, $1 \leqslant p < \infty$. Then*

$$(E[|X+Y|^p])^{1/p} \leqslant (E|X|^p)^{1/p} + (E|Y|^p)^{1/p}.$$

## 2.4 Convergence of random variables

When measuring a physical quantity such as the mass of an object, it is commonly believed that the average of several measurements is more reliable than a single one. Similarly, in applications of statistical inference when estimating a population mean $\mu$, a random sample $\{X_1, \ldots, X_n\}$ of size $n$ is drawn from the population, and the sample average or empirical mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is used as an estimator for the parameter $\mu$. This is based on the idea that as $n$ gets large, $\overline{X}_n$ will get closed to $\mu$ in some suitable sense.

In what follows we will review the different ideas of convergence one can give to a sequence of random variables and how they are interrelated.

**Definition 2.4.1** (Pointwise or sure convergence). *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$. The sequence $\{X_n\}_{n\geqslant 1}$ is said to* converge surely *or* pointwise *to a random variable $X$ if*

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \textit{for all} \quad \omega \in \Omega.$$

**Definition 2.4.2** (Convergence almost sure). *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$. The sequence $\{X_n\}_{n\geqslant 1}$ is said to* converge almost surely ($P$-a.s.) *or* with probability one *to a random variable $X$ if there exists a set $N \in \mathcal{F}$ such that $P(N) = 0$ and*

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \textit{for all} \quad \omega \in \Omega \setminus N.$$

*This is often written as $X_n \to X$ a.s. or $X_n \to X$ $P$-a.s.*

**Definition 2.4.3** (Convergence in probability). *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$. The sequence $\{X_n\}_{n\geqslant 1}$ is said to* converge in probability *to a random variable $X$ if for every $\varepsilon > 0$*

$$\lim_{n\to\infty} P(|X_n - X| \geqslant \varepsilon) = 0.$$

*This is often written as $X_n \xrightarrow{P} X$*

**Definition 2.4.4** (Convergence in distribution). *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$. The sequence $\{X_n\}_{n\geqslant 1}$ is said to* converge in distribution *or* in law *to a random variable $X$ if*

$$\lim_{n\to\infty} P(X_n < x) = P(X < x)$$

*for every $x \in \mathbb{R}$ at which $F(x) := P(X < x)$ is continuous.*

*This is often written as $X_n \xrightarrow{d} X$ or $X_n \xrightarrow{\mathcal{L}} X$. In fact, the random variables do not need to share a common probability space.*

**Definition 2.4.5** (Convergence in $p$th mean). *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$ and $p \geqslant 1$ a real number. The sequence $\{X_n\}_{n\geqslant 1}$ is said to* converge in the $p$th mean *or* in the $L^p$-norm *to a random variable $X$ if the $p$th moments $E[|X_n|^p]$ and $E[|X|^p]$ exist and*

$$\lim_{n\to\infty} E[|X_n - X|^p] = 0.$$

*This is often written as $X_n \xrightarrow{L^p} X$.*

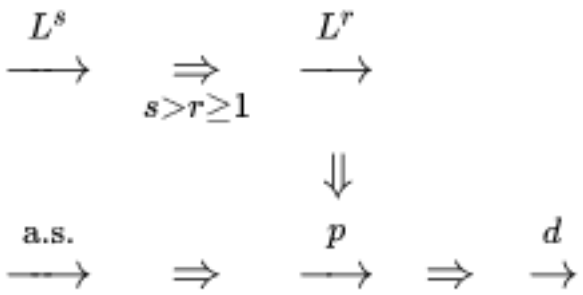


FIGURE 2.1: Interrelation of different types of convergence

**Theorem 2.4.6** (Lévy continuity criterion). *A sequence of random random variables $\{X_n\}_{n=1}^{\infty}$ (not necessarily sharing the same probability space) convergences in law to some random variable $X$ if, and only if*

$$\varphi_{X_n}(t) \xrightarrow{n\to\infty} \varphi_X(t),$$

*for every $t \in \mathbb{R}$ and $\varphi_X(t)$ is continuous at $t = 0$.*

**Example 2.4.7.** *Show that a sequence of Bernoulli random variables $X_n$ with probability $1/n$ converges to 0 in probability, in distribution and in any mean $p \geqslant 1$.*

**Example 2.4.8.** *A sequence of i.i.d. $\{X_n\}_{n=1}^{\infty}$ with law $N(0, n)$ for each $n \geqslant 1$ does not convergence in any of the above mentioned senses.*

**Example 2.4.9.** *Convergence in distribution of $\{X_n\}_{n=1}^{\infty}$ to $X$ means that the law of $X_n$ gets arbitrarily close to the law of $X$. Whilst, convergence in probability of $X_n$ to $X$ means that the random variable $X_n$ gets very close to the random variable $X$. For example, let $X \sim N(0, 1)$ and define $Y = -X$ . Then $X$ and $Y$ have the same law, so if $X_n$ convergence in law to $X$ so does it to $Y$ as well, but if $X_n$ convergences in probability to $X$, then it can not converge in probability to $Y$ since $X$ and $Y$ are different elements in $L^0(\Omega)$.*

### 2.4.1 Almost sure convergence vs convergence in probability

From a statistical point of view (and hence practical) there is no difference between the two notions. The difference is of philosophical nature. Hence, for a statistician, whether an estimator is consistent or strongly consistent is not really relevant,so proving that an estimator is (weakly) consistent is enough for sampling and statistical purposes.

Nevertheless, the difference between the two types of convergence is important in mathematics and mathematical statistics, as each type allows one for the use of different techniques that may lead to new theories. Hence, theoretically, they are important.

In any case, let us look at an example that may throw some light on how one can interpret the difference with an example. Imagine we use a device such that the probability of it failing is less than before. Convergence in probability says that the chance (probability) of it failing goes to 0 as the number of trials increases towards infinity. So, after using the device a large number of times, you can be very confident of it working correctly, it still might fail, it's just very unlikely.

Convergence almost surely is a bit stronger. It says that the total number of failures is *finite*. That is, if you count the number of failures as the number of usages goes to infinity, you will get a finite number. The impact of this is as follows: As you use the device more and more, you will, after some finite number of usages, exhaust all failures. From then on the device will work perfectly.

On the other hand, you do not actually know when you have exhausted all failures, so from a purely practical point of view, there is not much difference between the two modes of convergence.

However, it has importance from a philosophical point of view, for example, in the fact that the strong law of large numbers exists, as opposed to just the weak law. Because now, a scientific experiment to obtain, say, the speed of light, is justified in taking averages. At least in theory, after obtaining enough data, you can get arbitrarily close to the true speed of light. There will not be any failures (however improbable) in the averaging process.

**Example 2.4.10** (Lawof large numbers)**.** *Choose some $\varepsilon > 0$. Collect $n$ estimates $X_1, \ldots, X_n$ of the speed of light (or some other quantity) that has some* true *value $\mu$. We compute the average*

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

*As we obtain more data ($n$ increases) we can compute $S_n$, $n = 1, 2, \ldots$. The weak law says (under some assumptions on $X_i$) that the probability*

$$P(|S_n - \mu| > \varepsilon) \to 0$$

*as $n$ goes to infinity. Whereas the strong law says that the number of times $|S_n - \mu| > \varepsilon$ is finite with probability one. That is, if we define $I(|S_n - \mu| > \varepsilon)$ that returns one when $|S_n - \mu| > \varepsilon$ and zero otherwise, then*

$$\sum_{n=1}^{\infty} I(|S_n - \mu| > \varepsilon)$$

*convergences. This gives us considerable confidence in the value of $S_n$, because it guarantees (i.e. with probability one) the existence of some $n_0$ such that $|S_n - \mu| < \varepsilon$ for all $n \geqslant n_0$ (i.e. the average never* fails *for $n \geqslant n_0$). Note that the weak law gives no such guarantee.*

## 2.5 Borel-Cantelli lemmas

In this section we will review some results on classes of independent events which are important in proving laws of large numbers.

**Definition 2.5.1.** *Let $(\Omega, \mathcal{F})$ be a measure space and $\{A_n\}_{n \geqslant 1} \subset \mathcal{F}$ be a sequence of sets. Then*

$$\limsup_{n \to \infty} A_n = \bigcap_{k=1}^{\infty} \bigcup_{n \geqslant k} A_k,$$

$$\liminf_{n \to \infty} A_n = \bigcup_{k=1}^{\infty} \bigcap_{n \geqslant k} A_k,$$

**Proposition 2.5.2.** *Both* $\limsup_{n\to\infty} A_n$ *and* $\liminf_{n\to\infty} A_n$ *belong to* $\mathcal{F}$.

In probability theory, $\limsup_{n\to\infty} A_n$ is referred to as the event that "$A_n$ happens infinitely often" and $\liminf_{n\to\infty} A_n$ as the event that "all but a finitely many $A_n$'s happen".

**Example 2.5.3.** *Let* $\Omega = \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ *and* $P = \lambda$ *the Lebesgue measure. Consider the family of sets*

$$A_n = \begin{cases} [0, 1/n], & n \text{ odd}, \\ [1 - 1/n, 1], & n \text{ even}. \end{cases}$$

*Then* $\limsup_n A_n = \{0, 1\}$ *and* $\liminf_n \emptyset$.

The following result on the probabilities of $\limsup_n A_n$ and $\liminf_n A_n$ is very useful in probability theory.

**Theorem 2.5.4** (First Borel-Cantelli lemma)**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and* $\{A_n\}_{n\geqslant 1}$ *be a sequence of events in* $\mathcal{F}$. *If*

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

*then*

$$P\left(\limsup_{n\to\infty} A_n\right) = 0.$$

A partial converse of this result is often referred to as second Borel-Cantelli lemma, under the additional assumption of pairwise independent events.

**Theorem 2.5.5** (Second Borel-Cantelli lemma)**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and* $\{A_n\}_{n\geqslant 1}$ *be a sequence of* pairwise independent *events in* $\mathcal{F}$. *If*

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

*then*

$$P\left(\limsup_{n\to\infty} A_n\right) = 1.$$

**Remark 2.5.6.** *This result is also called a* zero-one *law as it states that for pairwise independent events, the probability that* $A_n$ *happens infinitely often is either 1 or 0. A very popular example is that of the infinite typing monkeys.*

**Example 2.5.7.** *A monkey pushes keys of a typewriter at random. Assume this typewriter has 50 keys and that the chances of a given key are equal (uniformly distributed). Consider the word* banana *which has 6 letters. The probability of typing* banana *in a row is* $(1/50)^6 = (15\ 625\ 000\ 000)^{-1}$. *Imagine we have infinitely many strings (monkeys typing) and we let* $A_n$ *be the event that the first 6 characters of string n is the word* banana. *Obviously,* $\{E_n\}_{n=1}^{\infty}$ *is a sequence of mutually independent events with*

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

*Hence, by the second Borel-Cantelli lemma we can conclude that*

$$P\left(\limsup_{n\to\infty} A_n\right) = 1.$$

*This means that the probability that the events $E_n$ occur infinitely often is one. This translates to saying that the event that a monkey types the word* banana *during this infinite string is one and, actually, this will happen an infinite number of times. One can generalize this to any given text, such as the complete works of William Shakespeare. Of course, the notion of infinite is crucial, and the time it would get for the monkey to type such a long text in a row is inconceivable.*

# Chapter 3

# Central limit theorem and strong law of large numbers

**Theorem 3.0.1** (Weak law of large numbers). *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with $\mu = E[X_1] = \cdots = E[X_n] = \cdots$. Then*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu,$$

*that is for every $\varepsilon > 0$*

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right).$$

Interpreting this result, the weak law states that for any nonzero margin specified, no matter how small, with a sufficiently large sample there will be a very high probability that the average of the observations will be close to the expected value; that is, within the margin.

**Theorem 3.0.2** (Strong law of large numbers). *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with $\mu = E[X_1] = \cdots = E[X_n] = \cdots$. Then*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu,$$

*that is*

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

What this means is that the probability that, as the number of trials n goes to infinity, the average of the observations converges to the expected value, is equal to one.

The assumptions of both laws are quite general (e.g. independent variables, no finite variance, etc.) but the distinction between the two is important since convergence in probability doesnot imply almost sure convergence, and there are assumption under which one does not have almost sure convergence but one has convergence in probability.

**Theorem 3.0.3** (Lindeberg-Lévy CLT). *Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$, both finite. Then*

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} N(0,1).$$

This theorem only ensures convergence of the laws of $X_n$, we have not specified any probability space. This result is very useful to construct confidence intervals and build hypothesis tests when the distributions of $X_n$ are not entirely known and the samples are large enough. This result concerns only means, there are other CLT-type of results concerning other statistics, for instance the maximum of a given sample of random variables (extreme value theory) which has its interest in insurance.