Chapter 4

Probability and Measure

4.1 Introduction

In this chapter we will examine probability theory from the measure theoretic perspective. The realisation that measure theory is the foundation of probability is due to the great Russian mathematician A.N.Kolmogorov (1903-1987) who published the hugely influential "Grundbegriffe der Wahrscheinlichkeitsrechnung" (*Foundations of the Theory of Probability* in English) in 1933. Since that time, measure theory has been at the centre of all mathematically rigorous work in probability theory and has been a vital tool in enabling the theory to develop both conceptually and in applications.

We have already seen that probability is a measure, random variables are measurable functions and expectation is a Lebesgue integral. But it is not true that "probability theory" can be reduced to a subset of "measure theory". This is because there are important probabilistic concepts such as independence and conditioning that do not appear naturally from within measure theory itself, although its important to appreciate (as we will see) that they can then be given precise mathematical meaning within the measure theoretic framework.

The famous Polish mathematician Mark Kac (1914-1984) once remarked that "Probability theory is measure theory with a soul." By the end of the course, you should be able to decide for yourself if you agree with this statement.

4.2 Basic Concepts of Probability Theory

4.2.1 Probability as Measure

Let us review what we know so far. In this chapter we will work with general probability spaces of the form $(\Omega, \mathcal{F}, P)$ where the *probability* P is a finite measure on $(\Omega, \mathcal{F})$ having total mass 1. So

$$P(\Omega) = 1 \quad \text{and} \quad 0 \leq P(A) \leq 1 \text{ for all } A \in \mathcal{F}.$$

$P(A)$ is the probability that the event $A \in \mathcal{F}$ takes place. Since $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, by M(ii) we have $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$ so that

$$P(A^c) = 1 - P(A).$$

A random variable X is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If $A \in \mathcal{B}(\mathbb{R})$, it is standard to use the notation $(X \in A)$ to denote the event $X^{-1}(A) \in \mathcal{F}$. The *law* or *distribution* of X is the induced probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by $p_X(B) = P(X^{-1}(B))$ for $B \in \mathcal{B}(\mathbb{R})$. So

$$p_X(B) = P(X \in B) = P(X^{-1}(B)) = P(\{\omega \in \Omega; X(\omega) \in B\}).$$

The *expectation* of X is the Lebesgue integral:

$$\mathbb{E}(X) = \int_\Omega X(\omega) dP(\omega) = \int_{\mathbb{R}} x dp_X(x),$$

(see Problem 51) which makes sense and yields a finite quantity if and only if X is *integrable*, i.e. $\mathbb{E}(|X|) < \infty$. In this case, we write $\mu = \mathbb{E}(X)$ and call it the *mean* of X. Note that for all $A \in \mathcal{F}$

$$P(A) = \mathbb{E}(\mathbf{1}_A).$$

By the result of Problem 15, any Borel measurable function f from $\mathbb{R}$ to $\mathbb{R}$ enables us to construct a new random variable $f(X)$ for which $f(X)(\omega) = f(X(\omega))$ for all $\omega \in \Omega$. For example we may take $f(x) = x^n$ for all $n \in \mathbb{N}$. Then the nth moment $\mathbb{E}(X^n)$ will exist and be finite if $|X|^n$ is integrable. If X has a finite second moment then its *variance* $\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2)$ always exists (see Problem 52). It is common to use the notation $\sigma^2 = \mathrm{Var}(X)$. The *standard deviation* of X is $\sigma = \sqrt{\mathrm{Var}(X)}$.

Here's some useful notation. If X and Y are random variables defined on the same probability space and $A_1, A_2 \in \mathcal{F}$ it is standard to write:

$$P(X \in A_1, Y \in A_2) = P((X \in A_1) \cap (Y \in A_2)).$$

4.2.2 Continuity of Probabilities

Recall from section 1.5 that a sequence of sets (A_n) with $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$ is *increasing* if $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$ and we write $A = \bigcup_{n\in\mathbb{N}} A_n$. We similarly say that a sequence (B_n) with $B_n \in \mathcal{F}$ for all $n \in \mathbb{N}$ is *decreasing* if $B_n \supseteq B_{n+1}$ for all $n \in \mathbb{N}$ and we write $B = \bigcap_{n\in\mathbb{N}} B_n$ in this case.

Theorem 4.2.1 *[Continuity of Probabilities]*

1. *If (A_n) is increasing then $P(A) = \lim_{n\to\infty} P(A_n)$.*
2. *If (B_n) is decreasing then $P(B) = \lim_{n\to\infty} P(B_n)$.*

Proof. (1) is just Theorem 1.5.1 of Chapter 1 applied to probability measures and (2) follows from Problem 9. □

Note that Theorem 4.2.1 (2) does not hold for general measures, only for finite (and hence probability) measures.

4.2.3 The Cumulative Distribution Function

Let $X : \Omega \to \mathbb{R}$ be a random variable. Its *cumulative distribution function* or *cdf* is the mapping $F_X : \mathbb{R} \to [0, 1]$ defined for each $x \in \mathbb{R}$ by

$$F_X(x) = P(X \leq x) = p_X((-\infty, x]).$$

When X is understood we will just denote F_X by F. The next result gathers together some useful properties of the cdf. Recall that if $f : \mathbb{R} \to \mathbb{R}$, the *left limit* at x is $\lim_{y\uparrow x} f(y) = \lim_{y\to x, y<x} f(y)$, and the *right limit* at x is $\lim_{y\downarrow x} f(y) = \lim_{y\to x, y>x} f(y)$. In general, left and right limits may not exist, but they do at every point when the function f is monotonic increasing (or decreasing).

Theorem 4.2.2 *Let X be a random variable having cdf F_X.*

1. *$P(X > x) = 1 - F(x)$,*
2. *$P(x < X \leq y) = F(y) - F(x)$ for all $x < y$.*
3. *F is monotonic increasing, i.e $F(x) \leq F(y)$ for all $x < y$,*
4. *$P(X = x) = F(x) - \lim_{y\uparrow x} F(y)$,*
5. *The mapping $x \to F(x)$ is right continuous, i.e. $F(x) = \lim_{y\downarrow x} F(y)$, for all $x \in \mathbb{R}$,*

6. $\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to\infty} F(x) = 1.$

Proof. (1), (2) and (3) are easy exercises.

(4) Let (a_n) be a sequence of positive numbers that decreases to zero. Let $x \in \mathbb{R}$ be arbitrary and for each $n \in \mathbb{N}$, define $B_n = (x - a_n < X \leq x)$. Then (B_n) decreases to the event $(X = x)$ and using (2) and Theorem 4.2.1 (2),

$$P(X = x) = \lim_{n\to\infty} P(B_n) = F(x) - \lim_{n\to\infty} F(x - a_n),$$

and the result follows.

(5) Let x and (a_n) be as in (4) and for each $n \in \mathbb{N}$ define $A_n = (X > x+a_n)$. The sets (A_n) are increasing to $(X > x)$ and using (1) and Theorem 4.2.1 (1) we find that

$$1 - F(x) = \lim_{n\to\infty} P(A_n) = 1 - \lim_{n\to\infty} F(x + a_n),$$

and the result follows.

(6) is Problem 43. □

Remark. It can be shown that a function $F : \mathbb{R} \to \mathbb{R}$ is the cdf of a random variable X if and only if it satisfies (3), (5) and (6) of Theorem 4.2.2.

We say that $x \in \mathbb{R}$ is a *continuity point* of F if F is continuous there, i.e. if it is left continuous as well as right continuous. In the literature, X is called a *continuous random variable* if its cdf F_X is continuous at every point $x \in \mathbb{R}$ and is called a *discrete random variable* if F_X has jump discontinuities at a countable set of points and is constant between these jumps. Note that if x is a continuity point of F then $P(X = x) = 0$ by Theorem 4.2.2(4). We say that F_X is *absolutely continuous* if there exists an integrable function $f_X : \mathbb{R} \to \mathbb{R}$ so that $F_X(x) = \int_{-\infty}^{x} f_X(y)dy$ for all $x \in \mathbb{R}$. In this case F_X is certainly continuous. The function f_X is called the *probability density function* or *pdf* of X. Clearly $f_X \geq 0$ (a.e.) and by Theorem 4.2.2 (6) we have $\int_{-\infty}^{\infty} f_X(y)dy = 1$. We have already seen the example of the Gaussian random variable that is absolutely continuous. Other examples that you may have encountered previously include the uniform, t, gamma and beta distributions. Typical examples of discrete random variables are the binomial and Poisson distributions.

4.2.4 Independence

In this subsection we consider the meaning of independence for events, random variables and σ-algebras.

Independence means multiply. We say that two events $A_1, A_2 \in \mathcal{F}$ are *independent* if

$$P(A_1 \cap A_2) = P(A_1)P(A_2).$$

We extend this by induction to n events. But for many applications, we want to discuss independence of infinitely many events, or to be precise a sequence (A_n) of events with $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$. The definition of independence is extended from the finite case by considering all finite subsets of the sequence. Formally:

Definition 4.1 We say that the events in the sequence (A_n) are *independent* if the finite set $\{A_{i_1}, A_{i_2}, \ldots, A_{i_m}\}$ is independent for all finite subsets $\{i_1, i_2, \ldots, i_m\}$ of the natural numbers, i.e.

$$P(A_{i_1} \cap A_{i_2} \cap \cdots, A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m}).$$

We recall that two random variables X and Y are said to be independent if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$, for all $A, B \in \mathcal{B}(\mathbb{R})$. In other words the events $(X \in A)$ and $(Y \in B)$ are independent for all $A, B \in \mathcal{B}(\mathbb{R})$. Again this is easily extended to finite collections of random variables. Now suppose we are given a sequence of random variables (X_n). We say that the X_n's are independent if every finite subset $X_{i_1}, X_{i_2}, \ldots, X_{i_m}$ of random variables is independent, i.e.

$$P(X_{i_1} \in A_{i_1}, X_{i_2} \in A_{i_2}, \ldots, X_{i_m} \in A_{i_m}) = P(X_{i_1} \in A_{i_1})P(X_{i_2} \in A_{i_2}) \cdots P(X_{i_m} \in A_{i_m})$$

for all $A_{i_1}, A_{i_2}, \ldots, A_{i_m} \in \mathcal{B}(\mathbb{R})$ and for all finite $\{i_1, i_2, \ldots, i_m\} \subset \mathbb{N}$.

In the case where there are two random variables, we may consider the random vector $Z = (X, Y)$ as a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ where the Borel σ-algebra $\mathcal{B}(\mathbb{R}^2)$ is the smallest σ-algebra generated by all open intervals of the form $(a, b) \times (c, d)$. The law of Z is, as usual, $p_Z = P \circ Z^{-1}$ and the *joint law of X and Y* is precisely $p_Z(A \times B) = P(X \in A, Y \in B)$ for $A, B \in \mathcal{B}(\mathbb{R})$. Then X and Y are independent if and only if

$$p_Z(A \times B) = p_X(A)p_Y(B),$$

i.e. the joint law factorises as the product of the marginals.

Theorem 4.2.3 *If X and Y are independent integrable random variables. Then*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Proof. By the two-dimensional version of Problem 51,

$$\begin{aligned}\mathbb{E}(XY) &= \int_{\mathbb{R}^2} xyp_Z(dx,dy) \\ &= \left(\int_{\mathbb{R}} xp_X(dx)\right)\left(\int_{\mathbb{R}} yp_Y(dy)\right) \\ &= \mathbb{E}(X)\mathbb{E}(Y),\end{aligned}$$

where we have used Fubini's theorem to write the integral over $\mathbb{R}^2$ as a repeated integral. □

Lets go back to measure theory and consider a measurable space (S, Σ). We say that $\Sigma' \subseteq \Sigma$ is a *sub-σ-algebra* if it is itself a σ-algebra. For example the *trivial* σ-algebra $\{S, \emptyset\}$ is a sub-σ-algebra of any σ-algebra defined on S. If (A_n) is sequence of sets in Σ then $\sigma(A_1, A_2, \ldots)$ is defined to be the smallest sub-σ-algebra of Σ that contains A_n for all $n \in \mathbb{N}$.

Sub-σ-algebras play an important role in probability theory. For example let X be a random variable defined on $(\Omega, \mathcal{F}, P)$. Then $\sigma(X)$ is the smallest sub-σ-algebra of $\mathcal{F}$ that contains all the events $X^{-1}(A)$ for $A \in \mathcal{B}(\mathbb{R})$. For example let X describe a simple coin toss so that

$$X = \begin{cases} 0 & \text{if the coin shows tails} \\ 1 & \text{if the coin shows heads} \end{cases}$$

If $A = X^{-1}(\{0\})$ then $A^c = X^{-1}(\{1\})$ and $\sigma(X) = \{\emptyset, A, A^c, \Omega\}$.

Two sub-σ-algebras $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be independent if

$$P(A \cap B) = P(A)P(B)$$

for all $A \in \mathcal{G}_1, B \in \mathcal{G}_2$. In the next section we will need the following proposition which is here stated without proof.

Proposition 4.2.1 *Let (A_n) and (B_n) be sequences of events in $\mathcal{F}$ which are such that the combined sequence (A_n, B_n) is independent (i.e. any finite subset containing both A_n's and B_m's is independent.) Then the two sub-σ-algebras $\sigma(A_1, A_2, \ldots)$ and $\sigma(B_1, B_2, \ldots)$ are independent.*

4.3 Tail Events

4.3.1 Limsup, Liminf and Borel-Cantelli

One of our main aims in this chapter is to establish key results about asymptotic behaviour of sequences of random variables, namely the law(s) of large

numbers and the central limit theorem. This focuses our attention on so called "tail events". We'll define these in the next subsection. We'll start with two important examples. Let (A_n) be a sequence of events so that $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$. We define

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

It is clear that $\liminf_{n\to\infty} A_n, \limsup_{n\to\infty} A_n \in \mathcal{F}$ and you can show in Problem 48 that $\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n$.

The event $\limsup_{n\to\infty} A_n$ is sometimes written $\{A_n, \text{ i.o.}\}$. The i.o. stands for *infinitely often* as intuitively, it is the event where infinitely many of the A_ns occur.

The event $\liminf_{n\to\infty} A_n$ is sometimes written $\{A_n, \text{ a.a.}\}$. The a.a. stands for *almost always* as intuitively, it is the event that occurs if all of the A_ns occur except for a possible finite number.

Theorem 4.3.1

$$P\left(\liminf_{n\to\infty} A_n\right) \leq \liminf_{n\to\infty} P(A_n) \leq \limsup_{n\to\infty} P(A_n) \leq P\left(\limsup_{n\to\infty} A_n\right).$$

Proof. We only prove the first inequality on the left here. As the events $\bigcap_{k=n}^{\infty} A_k$ are increasing to $\liminf_{n\to\infty} A_n$ we can use continuity of probability (Theorem 4.2.1. (1)) to show that

$$\begin{aligned} P\left(\liminf_{n\to\infty} A_n\right) &= \lim_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k\right) \\ &= \liminf_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k\right) \\ &\leq \liminf_{n\to\infty} P(A_n), \end{aligned}$$

where the last line uses monotonicity. The last inequality on the right in the statement of the theorem is Problem 49(b). □

The next result (particularly the first part) is very important and plays a vital role in proving later results. It is called the *Borel-Cantelli lemma* after Emile Borel who we've already met and the Italian mathematician Francesco Paulo Cantelli (1875-1966). Before we present the theorem and its proof, we

give a useful inequality. By Taylor's theorem there exists $0 < \theta < 1$ so that for all $x \geq 0$,

$$e^{-x} = 1 - x + \frac{x^2}{2} e^{-\theta x},$$

and so

$$e^{-x} \geq 1 - x \tag{4.3.1}$$

Theorem 4.3.2 *[The Borel-Cantelli Lemma.] Let (A_n) be a sequence of events in $\mathcal{F}$.*

1. *If $\sum_{n=1}^{\infty} P(A_n) < \infty$ then $P\left(\limsup_{n\to\infty} A_n\right) = 0$.*
2. *If (A_n) is independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$ then $P\left(\limsup_{n\to\infty} A_n\right) = 1$.*

Proof.

1. For all $m \in \mathbb{N}$ we have $\limsup_{n\to\infty} A_n \subseteq \bigcup_{k=m}^{\infty} A_k$ and so by monotonicity and Theorem 1.5.2,

$$\begin{aligned} P\left(\limsup_{n\to\infty} A_n\right) &\leq P\left(\bigcup_{k=m}^{\infty} A_k\right) \\ &\leq \sum_{k=m}^{\infty} P(A_k) \ \to 0 \text{ as } m \to \infty. \end{aligned}$$

2. By Problem 49(a),its enough to show that $P\left(\liminf_{n\to\infty} A_n^c\right) = 0$. Now let $m \in \mathbb{N}$ be arbitrary, then for all $n > m$ by independence and then using (4.3.1)

$$\begin{aligned} P\left(\bigcap_{k=m}^{n} A_k^c\right) &= \prod_{k=m}^{n} P(A_k^c) \\ &= \prod_{k=m}^{n} (1 - P(A_k)) \\ &\leq \prod_{k=m}^{n} e^{-P(A_k)} \\ &= \exp\left(-\sum_{k=m}^{n} P(A_k)\right) \ \to 0 \text{ as } n \to \infty. \end{aligned}$$

Then by continuity of probability (Theorem 4.2.1 (b)) $P\left(\bigcap_{k=m}^{\infty} A_k^c\right) = 0$. Since this holds for all $m \in \mathbb{N}$ we have

$$P\left(\liminf_{n\to\infty} A_n^c\right) = P\left(\bigcup_{m=1}^{\infty}\bigcap_{k=m}^{\infty} A_k^c\right) \leq \sum_{m=1}^{\infty} P\left(\bigcap_{k=m}^{\infty} A_k^c\right) = 0. \qquad \square$$

Let (X_n) be a sequence of independent random variables, each of which takes values 0 or 1 with probability 1/2. The possible scenarios $(X_1(\omega), X_1(\omega), \ldots)$ where $\omega \in \Omega$ are called *Bernoulli sequences*. So each Bernoulli sequence is a sequence of 0s and 1s.

Example Show that the pattern 1001 occurs infinitely often (with probability one) in a Bernoulli sequence.

Solution. Let E_n be the event that 1001 occurs starting at the nth point in the sequence. Then $P(E_n) = 1/16$. So $\sum_{n=1}^{\infty} P(E_n) = \infty$. Now E_n and E_{n+1} are not independent. But E_n and E_{n+4} are independent for all $n \in \mathbb{N}$. In fact $E_1, E_5, E_9, \ldots, E_{4n+1}, \ldots$ are independent and $\sum_{k=0}^{\infty} P(E_{4k+1}) = \infty$. So by the Borel-Cantelli lemma, $P\left(\limsup_{n\to\infty} E_{4n+1}\right) = 1$. So $1 = P\left(\limsup_{n\to\infty} E_{4n+1}\right) \leq P\left(\limsup_{n\to\infty} E_n\right) \leq 1$ and so $P\left(\limsup_{n\to\infty} E_n\right) = 1$ as is required.

Let (A_n) be a sequence of events in $\mathcal{F}$. The *tail σ-algebra* associated to (A_n) is

$$\tau = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \ldots,).$$

Clearly $\liminf_{n\to\infty} A_n \in \tau$ and $\limsup_{n\to\infty} A_n \in \tau$ (Why?) The next result may appear quite surprising. It is called the *Kolmogorov $0-1$ law* in honour of A.N.Kolmogorov.

Theorem 4.3.3 *[Kolmogorov's $0-1$ law.] Let (A_n) be a sequence of independent events in $\mathcal{F}$ and τ be the tail σ-algebra that they generate. If $A \in \tau$ then either $P(A) = 0$ or $P(A) = 1$.*

Proof. If $A \in \tau$ then $A \in \sigma(A_n, A_{n+1}, \ldots,)$ for all $n \in \mathbb{N}$. Then by Proposition 4.2.1 A is independent of $A_1, A_2, \ldots, A_{n-1}$ for all $n = 2, 3, 4, \ldots$. Since independence is only defined in terms of finite subcollections of sets, it follows that A is independent of $\{A_1, A_2, \ldots\}$. But $A \in \tau \subseteq \sigma(A_1, A_2, \ldots,)$. Hence A is independent of itself. So $P(A) = P(A \cap A) = P(A)^2$ and hence $P(A) = 0$ or 1. $\square$

In the light of the Kolmogorov $0-1$ law, at least the second part of the Borel-Cantelli lemma should no longer seem so surprising.

4.4 Convergence of Random Variables

Let (X_n) be a sequence of random variables, all of which are defined on the same probability space $(\Omega, \mathcal{F}, P)$. There are various different ways in which we can examine the convergence of this sequence to a random variable X (which is also defined on $(\Omega, \mathcal{F}, P)$.

We say that (X_n) converges to X

- *in probability* if given any $a > 0, \lim_{n\to\infty} P(|X_n - X| > a) = 0$,
- *in mean square* if $\lim_{n\to\infty} \mathbb{E}(|X_n - X|^2) = 0$,
- *almost surely* if there exists $\Omega' \in \mathcal{F}$ with $P(\Omega') = 1$ so that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega'$.

When (X_n) converges to X almost surely we sometimes write $X_n \to X$ (a.s.) as $n \to \infty$.

There are some relationships between these different modes of convergence.

Theorem 4.4.1 *1. Convergence in mean square implies convergence in probability.*

2. Convergence almost surely implies convergence in probability.

Proof.

1. This follows from Chebychev's inequality (Problems 25 and 26)[1] since for any $a > 0$,

$$P(|X_n - X| > a) \leq \frac{\mathbb{E}(|X_n - X|^2)}{a^2} \to 0 \text{ as } n \to \infty.$$

2. Let $\epsilon > 0$ be arbitrary and let $A_n = \{\omega \in \Omega, \text{there exists } m > n \text{ for which } |X_m(\omega) - X(\omega)| > \epsilon\}$. As $X_n \to X$ (a.s) as $n \to \infty$, (A_n) is a decreasing sequence of events. Let $A = \bigcap_{n=1}^{\infty} A_n$. If $\omega \in A$ then $X_n(\omega)$ cannot converge to $X(\omega)$ as $n \to \infty$ and so $P(A) \leq P(\Omega - \Omega') = 0$. By continuity of probability (Theorem 4.2.1 (b)), $\lim_{n\to\infty} P(A_n) = 0$. But for all $m > n, P(|X_m - X| > \epsilon) \leq P(A_n)$ and the result follows. □.

[1]Strictly speaking, we are using the inequality $P(|Y| > a) \leq \frac{\mathbb{E}(|Y|^2)}{a^2}$, which is proved in the same way as Chebychev's.

We have a partial converse to Theorem 4.4.1 (2)

Theorem 4.4.2 *If $X_n \to X$ in probability as $n \to \infty$ then there is a subsequence of (X_n) that converges to X almost surely.*

Proof. If (X_n) converges in probability to X, for all $c > 0$, given any $\epsilon > 0$, there exists $N(c) \in \mathbb{N}$ so that for all $n > N(c)$,

$$P(|X_n - X| > c) < \epsilon.$$

In order to find our subsequence, first choose, $c = 1$ and $\epsilon = 1$, then for $n > N(1)$,

$$P(|X_n - X| > 1) < 1.$$

Next choose $c = 1/2$ and $\epsilon = 1/4$, then for $n > N(2)$,

$$P(|X_n - X| > 1/2) < 1/4,$$

and for $r \geq 3$, $c = 1/r$ and $\epsilon = 1/2^r$, then for $n > N(r)$,

$$P(|X_n - X| > 1/r) < 1/2^r,$$

Now choose the numbers $k_r = N(r) + 1$, for $r \in \mathbb{N}$ to obtain a subsequence (X_{k_r}) so that for all $r \in \mathbb{N}$,

$$P(|X_{k_r} - X| > 1/r) < 1/2^r.$$

Since $\sum_{r=1}^{\infty} \frac{1}{2^r} < \infty$, by the first Borel-Cantelli lemma (Theorem 4.3.2 (i)) we have

$$P(\limsup_{r\to\infty} |X_{k_r} - X| > 1/r) = 0,$$

and so

$$P(\liminf_{r\to\infty} |X_{k_r} - X| \leq 1/r) = 1.$$

Then for any $\omega \in \liminf_{r\to\infty}(|X_{k_r} - X| \leq 1/r)$, given $\epsilon > 0$, we can find $r(\omega)$ such that for $m > r(\omega)$, $|X_{k_m}(\omega) - X(\omega)| \leq \frac{1}{m} < \epsilon$, and the result follows. □

Note that there is no simple relationship between a.s. convergence and convergence in mean square.

4.5 Laws of Large Numbers

Let (X_n) be a sequence of random variables all defined on the same probability space, that have the following properties,

- they are independent (see section 4.2)
- they are identically distributed , i.e. $p_{X_n} = p_{X_m}$ for all $n \neq m$. In other words, for all $A \in \mathcal{B}(\mathbb{R})$,

$$P(X_1 \in A) = P(X_2 \in A) = \cdots = P(X_n \in A) = \cdots$$

Such a sequence is said to be *i.i.d.*. I.i.d. sequences are very important in probability modelling (consider the steps of a random walk) and also statistics (consider a sequence of idealised experiments carried out under identical conditions.) We can form a new sequence of random variables $(\overline{X_n})$ where $\overline{X_n}$ is the *empirical mean*

$$\overline{X_n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

If X_n is integrable for some (and hence all) $n \in \mathbb{N}$ then $\mathbb{E}(X_n) = \mu$ is finite. It also follows that $\overline{X_n}$ is integrable, and by linearity $\mathbb{E}(\overline{X_n}) = \mu$. If $\mathbb{E}(X_n^2) < \infty$ then (see Problem 53) $\mathrm{Var}(X_n) = \sigma^2 < \infty$ for some (and hence all) $n \in \mathbb{N}$, and it follows by elementary properties of the variance that $\mathrm{Var}(\overline{X_n}) = \frac{\sigma^2}{n}$. It is extremely important to learn about the asymptotic behaviour of $\overline{X_n}$ as $n \to \infty$. Two key results are the *weak law of large numbers* or WLLN and the *strong law of large numbers* or SLLN. In fact the second of these implies the first but its much harder to prove. Later in this chapter we will study the *central limit theorem* or CLT.

Theorem 4.5.1 *[WLLN] Let (X_n) be a sequence of integrable i.i.d. random variables with $\mathbb{E}(X_n) = \mu$ for all $n \in \mathbb{N}$. Suppose also that $\mathbb{E}(X_n^2) < \infty$ for all $n \in \mathbb{N}$. Then $\overline{X_n} \to \mu$ in probability as $n \to \infty$.*

Proof. Let $\sigma^2 = \mathrm{Var}(X_n)$ for all $n \in \mathbb{N}$. Then by Chebychev's inequality, for all $a > 0$,

$$\begin{aligned} P(|\overline{X_n} - \mu| > a) &\leq \frac{\mathrm{Var}(\overline{X_n})}{a^2} \\ &= \frac{\sigma^2}{na^2} \to 0 \text{ as } n \to \infty. \quad \square \end{aligned}$$

Theorem 4.5.2 *[SLLN] Let (X_n) be a sequence of integrable i.i.d. random variables with $\mathbb{E}(X_n) = \mu$ for all $n \in \mathbb{N}$. Suppose also that $\mathbb{E}(X_n^2) < \infty$ for all $n \in \mathbb{N}$. Then $\overline{X_n} \to \mu$ almost surely as $n \to \infty$.*

Before we discuss the proof we make an observation. SLLN $\Rightarrow$ WLLN by Theorem 4.4.1 (2). The full proof of the SLLN is a little difficult for this course (see e.g. Rosenthal pp.47-9). We'll give a manageable proof by making an assumption on the fourth moments of the sequence (X_n).

Assumption 4.1 $\mathbb{E}((X_n - \mu)^4) = b < \infty$ for all $n \in \mathbb{N}$.

Proof of SLLN under Assumption 4.1. Assume that $\mu = 0$. If not we can just replace X_n throughout the proof with $Y_n = X_n - \mu$. Let $S_n = X_1 + X_2 + \cdots + X_n$ so that $S_n = n\overline{X_n}$ for all $n \in \mathbb{N}$. Consider $\mathbb{E}(S_n^4)$. It contains many terms of the form $\mathbb{E}(X_j X_k X_l X_m)$ (with distinct indices) and these all vanish by independence. A similar argument disposes of terms of the form $\mathbb{E}(X_j X_k^3)$ and $\mathbb{E}(X_j X_k X_l^2)$. The only terms with non-vanishing expectation are n terms of the form X_i^4 and $\binom{n}{2}.\binom{4}{2} = 3n(n-1)$ terms of the form $X_i^2 X_j^2$ with $i \neq j$. Now by Problem 44, X_i^2 and X_j^2 are independent for $i \neq j$ and so

$$\mathbb{E}(X_i^2 X_j^2) = \mathbb{E}(X_i^2)\mathbb{E}(X_j^2) = \mathrm{Var}(X_i^2)\mathrm{Var}(X_j^2) = \sigma^4.$$

We then have

$$\begin{aligned} \mathbb{E}(S_n^4) &= \sum_{i=1}^{n} \mathbb{E}(X_i^4) + \sum_{i \neq j} \mathbb{E}(X_i^2 X_j^2) \\ &= nb + 3n(n-1)\sigma^4 \leq Kn^2, \end{aligned}$$

where $K = b + 3\sigma^4$. Then for all $a > 0$, by Markov's inequality (Lemma 3.3.1)

$$\begin{aligned} P(|\overline{X_n}| > a) &= P(S_n^4 > a^4 n^4) \\ &\leq \frac{\mathbb{E}(S_n^4)}{a^4 n^4} \\ &\leq \frac{Kn^2}{a^4 n^4} = \frac{K}{a^4 n^2}. \end{aligned}$$

But $\sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$ and so by the first Borel-Cantelli lemma, $P(\limsup_{n\to\infty} |\overline{X_n}| > a) = 0$ and so $P(\liminf_{n\to\infty} |\overline{X_n}| \leq a) = 1$. By a similar argument to the last part of Theorem 4.4.2 we deduce that $\overline{X_n} \to 0$ a.s. as required. $\square$

Notes

1. The last part of the proof skated over some details. In fact you can show that for any sequence (Y_n) of random variables $P(\limsup_{n\to\infty} |Y_n - Y| \geq a) = 0$ for all $a > 0$ implies that $Y_n \to Y$ (a.s.) as $n \to \infty$. See Lemma 5.2.2 in Rosenthal p.45.

2. The proof in the general case without Assumption 4.1 uses a *truncation argument* and defines $Y_n = X_n \mathbf{1}_{\{X_n \leq n\}}$. Then $Y_n \leq n$ for all n and so $\mathbb{E}(Y_n^k) \leq n^k$ for all k. If $X_n \geq 0$ for all n, $\mathbb{E}(Y_n) \to \mu$ by monotone convergence. Roughly speaking we can prove a SLLN for the $\overline{Y_n}$s. We then need a clever probabilistic argument to transfer this to the $\overline{X_n}$s. The assumption in Theorem 4.5.2 that all the random variables have a finite second moment may also be dropped.

4.6 Characteristic Functions and Weak Convergence

In this section, we introduce two tools that we will need to prove the central limit theorem.

4.6.1 Characteristic Functions

Let (S, Σ, m) be a measure space and $f : S \to \mathbb{C}$ be a complex-valued function. Then we can write $f = f_1 + if_2$ where f_1 and f_2 are real-valued functions. We say that f is measurable/integrable if both f_1 and f_2 are. Define $|f|(x) = |f(x)| = \sqrt{f_1(x)^2 + f_2(x)^2}$ for each $x \in S$. It is not difficult to see that $|f|$ is measurable, using e.g. Problem 13. In Problem 56, you can prove that f is integrable if and only if $|f|$ is integrable. The Lebesgue dominated convergence theorem continues to hold for sequences of measurable functions from S to $\mathbb{C}$.

Now let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$. Its *characteristic function* $\phi_X : \mathbb{R} \to \mathbb{C}$ and is defined, for each $u \in \mathbb{R}$, by

$$\phi_X(u) = \mathbb{E}(e^{iuX}) = \int_{\mathbb{R}} e^{iuy} p_X(dy).$$

Note that $y \to e^{iuy}$ is measurable since $e^{iuy} = \cos(uy) + i\sin(uy)$ and integrability holds since $|e^{iuy}| \leq 1$ for all $y \in \mathbb{R}$ and in fact we have $|\phi_X(u)| \leq 1$ for all $u \in \mathbb{R}$.

Example $X \sim N(\mu, \sigma^2)$ means that X has a normal or Gaussian distribution with mean μ and variance σ^2 so that for all $x \in \mathbb{R}$,

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} dy.$$

In Problem 56, you can show for yourself that in this case, for all $u \in \mathbb{R}$

$$\phi_X(u) = \exp\left\{i\mu u - \frac{1}{2}\sigma^2 u^2\right\}.$$

Characteristic functions have many interesting properties. Here is one of the most useful. It is another instance of the "independence means multiply" philosophy.

Theorem 4.6.1 *If X and Y are independent random variables then for all $u \in \mathbb{R}$,*

$$\phi_{X+Y}(u) = \phi_X(u)\phi_Y(u).$$

Proof.

$$\phi_{X+Y}(u) = \mathbb{E}(e^{iu(X+Y)}) = \mathbb{E}(e^{iuX}e^{iuY}) = \mathbb{E}(e^{iuX})\mathbb{E}(e^{iuY}) = \phi_X(u)\phi_Y(u),$$

by Problem 44. □

The following result is also important but we omit the proof. It tells us that the probability law of a random variable is uniquely determined by its characteristic function.

Theorem 4.6.2 *If X and Y are two random variables for which $\phi_X(u) = \phi_Y(u)$ for all $u \in \mathbb{R}$ then $p_X = p_Y$.*

The characteristic function is the Fourier transform of the law p_X of the random variable X and we have seen that it always exists. In elementary probability theory courses we often meet the Laplace transform $\mathbb{E}(e^{uX})$ of X which is called the *moment generating function*. This exists in some nice cases (e.g. when X is Gaussian), but will not do so in general as $y \to e^{uy}$ may not be integrable since it becomes unbounded as $y \to \infty$ (when $u > 0$) and as $y \to -\infty$ (when $u < 0$.)

We will now develop an important inequality that we will need to prove the central limit theorem. Let $x \in \mathbb{R}$ and let $R_n(x)$ be the remainder term of the series expansion at $n \in \mathbb{N} \cup \{0\}$ in e^{ix}, i.e.

$$R_n(x) = e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}.$$

Note that $R_0(x) = e^{ix} - 1 = \begin{cases} \int_0^x ie^{iy}dy & \text{if } x > 0 \\ -\int_x^0 ie^{iy}dy & \text{if } x < 0 \end{cases}$. From the last two identities, we have $|R_0(x)| \leq \min\{|x|, 2\}$. Then you should check that $R_n(x) = \begin{cases} \int_0^x iR_{n-1}(y)dy & \text{if } x > 0 \\ -\int_x^0 iR_{n-1}(y)dy & \text{if } x < 0 \end{cases}$. Finally using induction, we can deduce the useful inequality:

$$|R_n(x)| \leq \min\left\{\frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!}\right\}.$$

Now let X be a random variable with characteristic function ϕ_X for which $E(|X|^n) < \infty$ for some given $n \in \mathbb{N}$. Then integrating the last inequality yields for all $y \in \mathbb{R}$

$$\left|\phi_X(y) - \sum_{k=0}^{n} \frac{(iy)^k \mathbb{E}(X^k)}{k!}\right| \leq \mathbb{E}\left[\min\left\{\frac{2|yX|^n}{n!}, \frac{|yX|^{n+1}}{(n+1)!}\right\}\right]. \tag{4.6.2}$$

When we prove the CLT we will want to apply this in the case $n = 2$ to a random variable that has $\mathbb{E}(X) = 0$. Then writing $\mathbb{E}(X^2) = \sigma^2$ we deduce that for all $u \in \mathbb{R}$.

$$\left|\phi_X(y) - 1 + \frac{1}{2}\sigma^2 y^2\right| \leq \theta(y), \tag{4.6.3}$$

where $\theta(y) = y^2\mathbb{E}\left[\min\left\{|X|^2, |y|\frac{|X|^3}{6}\right\}\right]$. Note that

$$\min\left\{|X|^2, |y|\frac{|X|^3}{6}\right\} \leq |X|^2$$

which is integrable by assumption. Also we have

$$\lim_{y\to 0} \min\left\{|X|^2, |y|\frac{|X|^3}{6}\right\} = 0,$$

and so by the dominated convergence theorem we can deduce the following important property of θ which is that

$$\lim_{y\to 0} \frac{\theta(y)}{y^2} = 0. \tag{4.6.4}$$

4.6.2 Weak Convergence

A sequence (μ_n) of probability measures on $\mathbb{R}$ is said to converge *weakly* to a probability measure μ if

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} f(x)\mu(dx)$$

for all bounded continuous functions f defined on $\mathbb{R}$. In the case where there is a sequence of random variables (X_n) and instead of μ_n we have p_{X_n} and also μ is the law p_X of a random variable X we say that (X_n) *converges in distribution* to X; so that convergence in distribution means the same thing as weak convergence of the sequence of laws. It can be shown that (X_n) converges to X in distribution if and only if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at every continuity point x of the c.d.f. F.

It can be shown that convergence in probability implies convergence in distribution (and so, by Theorem 4.4.1 (2), almost sure convergence also implies convergence in distribution.) For a proof, see Proposition 10.0.3 on p.98 in Rosenthal.

There is an important link between the concepts of weak convergence and characteristic functions which we present next.

Theorem 4.6.3 *[The Continuity Theorem] Let (X_n) be a sequence of random variables, where each X_n has characteristic function ϕ_n and let X be a random variable having characteristic function ϕ. Then (X_n) converges to X in distribution if and only if $\lim_{n\to\infty} \phi_n(u) = \phi(u)$ for all $u \in \mathbb{R}$.*

Proof. We'll only do the easy part here. Suppose (X_n) converges to X in distribution. Then

$$\begin{aligned} \phi_n(u) &= \int_{\mathbb{R}} \cos(uy)p_{X_n}(dy) + i\int_{\mathbb{R}} \sin(uy)p_{X_n}(dy) \\ &\to \int_{\mathbb{R}} \cos(uy)p_X(dy) + i\int_{\mathbb{R}} \sin(uy)p_X(dy) = \phi(u), \end{aligned}$$

as $n \to \infty$, since both $y \to \cos(uy)$ and $y \to \sin(uy)$ are bounded continuous functions. See e.g. Rosenthal pp.108-9 for the converse [2]. □

4.7 The Central Limit Theorem

Let (X_n) be a sequence of i.i.d. random variables having finite mean μ and finite variance σ^2. We have already met the SLLN which tells us that $\overline{X_n}$

[2] This reference is to the first edition. You'll find it on pp. 132-3 in the second edition

converges to μ a.s. as $n \to \infty$. Note that the standard deviation (i.e. the square root of the variance) of $\overline{X_n}$ is $\sigma/\sqrt{n}$ which also converges to zero as $n \to \infty$. Now consider the sequence (Y_n) of standardised random variables defined by

$$Y_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \tag{4.7.5}$$

Then $\mathbb{E}(Y_n) = 0$ and $\mathrm{Var}(Y_n) = 1$ for all $n \in \mathbb{N}$.

Its difficult to underestimate the importance of the next result. It shows that the normal distribution has a universal character as the attractor of the sequence (Y_n). From a modelling point of view, it tells us that as you combine together many i.i.d. different observations then they aggregate to give a normal distribution. This is of vital importance in applied probability and statistics. Note however that if we drop our standing assumption that all the X_n's have a finite variance, then this would no longer be true.

Theorem 4.7.1 *[Central Limit Theorem] Let (X_n) be a sequence of i.i.d. random variables each having finite mean μ and finite variance σ^2. Then the corresponding sequence (Y_n) of standardised random variables converges in distribution to the standard normal $Z \sim N(0,1)$, i.e. for all $a \in \mathbb{R}$*

$$\lim_{n\to\infty} P(Y_n \leq a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}y^2} dy.$$

Before we give the proof, we state a known fact from elementary analysis. We know that for all $y \in \mathbb{R}$,

$$\lim_{n\to\infty} \left(1 + \frac{y}{n}\right)^n = e^y.$$

Now for all $y \in \mathbb{R}$, let $(\alpha_n(y))$ be a sequence of real (or complex) numbers for which $\lim_{n\to\infty} \alpha_n(y) = 0$. Then we also have that for all $y \in \mathbb{R}$

$$\lim_{n\to\infty} \left(1 + \frac{y + \alpha_n(y)}{n}\right)^n = e^y \tag{4.7.6}$$

You may want to try your hand at proving this rigorously.

Proof. For convenience we assume that $\mu = 0$ and $\sigma = 1$. Indeed if it isn't we can just replace X_n everywhere by $(X_n - \mu)/\sigma$. Let ψ be the common characteristic function of the X_ns so that in particular $\psi(u) = \mathbb{E}(e^{iuX_1})$ for

all $u \in \mathbb{R}$. Let ϕ_n be the characteristic function of Y_n for each $n \in \mathbb{N}$. Then for all $u \in \mathbb{R}$, using Theorem 4.6.1 we find that

$$\begin{aligned}
\phi_n(u) &= \mathbb{E}(e^{iuS_n/\sqrt{n}}) \\
&= \mathbb{E}\left(e^{iu(\frac{1}{\sqrt{n}}(X_1+X_2+\cdots+X_n)}\right) \\
&= \psi(u/\sqrt{n})^n \\
&= \mathbb{E}\left(e^{i\frac{u}{\sqrt{n}}X_1}\right)^n \\
&= \left(1+\frac{iu}{\sqrt{n}}\mathbb{E}(X_1) - \frac{u^2}{2n}\mathbb{E}(X_1^2) + \frac{\theta_n(u)}{n}\right)^n,
\end{aligned}$$

where by (4.6.3) and the same argument we used to derive (4.6.4),

$$|\theta_n(u)| \leq u^2\mathbb{E}\left[\min\left\{|X_1|^2, \frac{|u|.|X_1|^3}{6\sqrt{n}}\right\}\right] \to 0$$

as $n \to \infty$, for all $u \in \mathbb{R}$.

Now we use (4.7.6) to find that

$$\begin{aligned}
\phi_n(u) &= \left(1 - \frac{\frac{u^2}{2} - \theta_n(u)}{n}\right)^n \\
&\to e^{-\frac{1}{2}u^2} \text{ as } n \to \infty.
\end{aligned}$$

The result then follows by the Lévy continuity theorem (Theorem 4.6.3). □

The CLT may be extensively generalised. We mention just two results here. If the i.i.d. sequence (X_n) is such that $\mu = 0$ and $\mathbb{E}(|X_n|^3) = \rho^3 < \infty$, the *Berry-Esseen theorem* gives a useful bound for the difference between the cdf of the normalised sum and the cdf Φ of the standard normal. To be precise we have that for all $x \in \mathbb{R}, n \in \mathbb{N}$:

$$\left|P\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) - \Phi(x)\right| \leq C\frac{\rho}{\sqrt{n}\sigma^3},$$

where $C > 0$.

We can also relax the requirement that the sequence (X_n) be i.i.d.. Consider the *triangular array* $(X_{nk}, k = 1, \ldots, n, n \in \mathbb{N})$ of random variables which we may list as follows:

$$\begin{array}{ccccccc}
X_{11} & & & & & & \\
X_{21} & X_{22} & & & & & \\
X_{31} & X_{32} & X_{33} & & & & \\
\vdots & \vdots & \vdots & & & & \\
X_{n1} & X_{n2} & X_{n3} & \ldots & X_{nn} & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & &
\end{array}$$

We assume that each row comprises independent random variables. Assume further that $\mathbb{E}(X_{nk}) = 0$ and $\sigma_{nk}^2 = \mathbb{E}(X_{nk}^2) < \infty$ for all k, n. Define the row sums $S_n = X_{n1} + X_{n2} + \cdots + X_{nn}$ for all $n \in \mathbb{N}$ and define $\tau_n = \mathrm{Var}(S_n) = \sum_{k=1}^n \sigma_{nk}^2$. *Lindeburgh's central limit theorem* states that if we have the asymptotic tail condition

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{\tau_n^2} \int_{|X_{nk}|\geq \epsilon\tau_n} X_{nk}^2(\omega) dP(\omega) = 0,$$

for all $\epsilon > 0$ then $\frac{S_n}{\tau_n}$ converges in distribution to a standard normal as $n \to \infty$.

The highlights of this last chapter have been the proofs of the law of large numbers and central limit theorem. There is a third result that is often grouped together with the other two as one of the key results about sums of i.i.d. random variables. It is called the *law of the iterated logarithm* and it gives bounds on the fluctuations of S_n for an i.i.d sequence with $\mu = 0$ and $\sigma = 1$. The result is quite remarkable. It states that

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log(n)}} = 1 \text{ a.s.} \tag{4.7.7}$$

This means that (with probability one) if $c > 1$ then only finitely many of the events $S_n > c\sqrt{2n\log\log(n)}$ occur but if $c < 1$ then infinitely many of such events occur. You should be able to deduce from (4.7.7) that

$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log(n)}} = -1 \text{ a.s.}$$