A Refresher in Probability Calculus

ulm university universität ulm

VERSION: March 10, 2010

Contents

Chapter 1

Facts form Probability and Measure Theory

We will assume that most readers will be familiar with such things from an elementary course in probability and statistics; for a clear introduction see, e.g. [GW86], or the first few chapters of [GS01]; [Ros97], [Res01], [Dur99], [Ros00] are also useful.

1.1 Measure

The language of modelling financial markets involves that of probability, which in turn involves that of *measure theory*. This originated with Henri Lebesgue (1875-1941), in his thesis, 'Intégrale, longueur, aire' [Leb02]. We begin with defining a measure on $\mathbb{R}$ generalising the intuitive notion of length.
The length $\mu(I)$ of an interval $I = (a,b), [a,b], [a,b)$ or $(a,b]$ should be $b-a$: $\mu(I) = b-a$. The length of the disjoint union $I = \bigcup_{r=1}^{n} I_r$ of intervals I_r should be the sum of their lengths:

$$\mu\left(\bigcup_{r=1}^{n} I_r\right) = \sum_{r=1}^{n} \mu(I_r) \qquad \text{(finite additivity)}.$$

Consider now an infinite sequence $I_1, I_2, \ldots$(*ad infinitum*) of disjoint intervals. Letting n tend to ∞ suggests that length should again be additive over disjoint intervals:

$$\mu\left(\bigcup_{r=1}^{\infty} I_r\right) = \sum_{r=1}^{\infty} \mu(I_r) \qquad \text{(countable additivity)}.$$

For I an interval, A a subset of length $\mu(A)$, the length of the complement $I \setminus A := I \cap A^c$ of A in I should be

$$\mu(I \setminus A) = \mu(I) - \mu(A) \qquad \text{(complementation)}.$$

If $A \subseteq B$ and B has length $\mu(B) = 0$, then A should have length 0 also:

$$A \subseteq B \quad \text{and} \quad \mu(B) = 0 \;\Rightarrow\; \mu(A) = 0 \qquad \text{(completeness)}.$$

The term 'countable' here requires comment. We must distinguish first between finite and infinite sets; then countable sets (like $\mathbb{N} = \{1, 2, 3, \ldots\}$) are the 'smallest', or 'simplest', infinite sets, as distinct from uncountable sets such as $\mathbb{R} = (-\infty, \infty)$.
Let $\mathcal{F}$ be the smallest class of sets $A \subset \mathbb{R}$ containing the intervals, closed under countable disjoint unions and complements, and complete (containing all subsets of sets of length 0 as sets of length 0). The above suggests – what Lebesgue showed – that length can be sensibly defined on the sets $\mathcal{F}$ on the line, but on no others. There are others – but they are hard to construct (in technical language: the axiom of choice, or some variant of it such as Zorn's lemma, is needed to demonstrate the existence of non-measurable sets – but all such proofs are highly non-constructive). So: some but not all subsets of the line have a length. These are called the *Lebesgue-measurable sets*, and form the class $\mathcal{F}$ described above; length, defined on $\mathcal{F}$, is called *Lebesgue measure* μ (on the real line, $\mathbb{R}$). Turning now to the general case, we make the above rigorous. Let Ω be a set.

Definition 1.1.1. *A collection $\mathcal{A}_0$ of subsets of Ω is called an algebra on Ω if:*

(i) $\Omega \in \mathcal{A}_0$,
(ii) $A \in \mathcal{A}_0 \Rightarrow A^c = \Omega \setminus A \in \mathcal{A}_0$,
(iii) $A, B \in \mathcal{A}_0 \Rightarrow A \cup B \in \mathcal{A}_0$.

Using this definition and induction, we can show that an algebra on Ω is a family of subsets of Ω closed under finitely many set operations.

Definition 1.1.2. *An algebra $\mathcal{A}$ of subsets of Ω is called a σ-algebra on Ω if for any sequence $A_n \in \mathcal{A}$, $(n \in \mathbb{N})$, we have*

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$

Such a pair $(\Omega, \mathcal{A})$ is called a measurable space.

Thus a σ-algebra on Ω is a family of subsets of Ω closed under any countable collection of set operations.
The main examples of σ-algebras are σ-algebras generated by a class $\mathcal{C}$ of subsets of Ω, i.e. $\sigma(\mathcal{C})$ is the smallest σ-algebra on Ω containing $\mathcal{C}$.
The Borel σ-algebra $\mathcal{B} = \mathcal{B}(\mathbb{R})$ is the σ-algebra of subsets of $\mathbb{R}$ generated by the open intervals (equivalently, by half-lines such as $(-\infty, x]$ as x varies in $\mathbb{R}$.
As our aim is to define measures on collection of sets we now turn to set functions.

Definition 1.1.3. *Let Ω be a set, $\mathcal{A}_0$ an algebra on Ω and μ_0 a non-negative set function $\mu_0 : \mathcal{A}_0 \to [0, \infty]$ such that $\mu_0(\emptyset) = 0$. μ_0 is called:*

(i) additive, if $A, B \in \mathcal{A}_0, A \cap B = \emptyset \Rightarrow \mu_0(A \cup B) = \mu_0(A) + \mu_0(B)$,
(ii) countably additive, if whenever $(A_n)_{n \in \mathbb{N}}$ is a sequence of disjoint sets in $\mathcal{A}_0$ with $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}_0$ then

$$\mu_0\left(\bigcup_{n=0}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu_0(A_n).$$

Definition 1.1.4. *Let $(\Omega, \mathcal{A})$ be a measurable space. A countably additive map*

$$\mu : \mathcal{A} \to [0, \infty]$$

is called a measure on $(\Omega, \mathcal{A})$. The triple $(\Omega, \mathcal{A}, \mu)$ is called a measure space.

Recall that our motivating example was to define a measure on $\mathbb{R}$ consistent with our geometrical knowledge of length of an interval. That means we have a suitable definition of measure on a family of subsets of $\mathbb{R}$ and want to extend it to the generated σ-algebra. The measure-theoretic tool to do so is the Carathéodory extension theorem, for which the following lemma is an inevitable prerequisite.

Lemma 1.1.1. *Let Ω be a set. Let $\mathcal{I}$ be a π-system on Ω, that is, a family of subsets of Ω closed under finite intersections: $I_1, I_2 \in \mathcal{I} \Rightarrow I_1 \cap I_2 \in \mathcal{I}$. Let $\mathcal{A} = \sigma(\mathcal{I})$ and suppose that μ_1 and μ_2 are finite measures on $(\Omega, \mathcal{A})$ (i.e. $\mu_1(\Omega) = \mu_2(\Omega) < \infty$) and $\mu_1 = \mu_2$ on $\mathcal{I}$. Then*

$$\mu_1 = \mu_2 \quad on \ \ \mathcal{A}.$$

Theorem 1.1.1 (Carathéodory Extension Theorem)**.** *Let Ω be a set, $\mathcal{A}_0$ an algebra on Ω and $\mathcal{A} = \sigma(\mathcal{A}_0)$. If μ_0 is a countably additive set function on $\mathcal{A}_0$, then there exists a measure μ on $(\Omega, \mathcal{A})$ such that*

$$\mu = \mu_0 \quad on \ \ \mathcal{A}_0.$$

If μ_0 is finite, then the extension is unique.

For proofs of the above and further discussion, we refer the reader to Chapter 1 and Appendix 1 of [Wil91] and the appendix in [Dur96].
Returning to the motivating example $\Omega = \mathbb{R}$, we say that $A \subset \mathbb{R}$ belongs to the collection of sets $\mathcal{A}_0$ if A can be written as

$$A = (a_1, b_1] \cup \ldots \cup (a_r, b_r],$$

where $r \in \mathbb{N}$, $-\infty \leq a_1 < b_1 \leq \ldots \leq a_r < b_r \leq \infty$. It can be shown that $\mathcal{A}_0$ is an algebra and $\sigma(\mathcal{A}_0) = \mathcal{B}$. For A as above define

$$\mu_0(A) = \sum_{k=1}^{r} (b_k - a_k).$$

μ_0 is well-defined and countably additive on $\mathcal{A}_0$. As intervals belong to $\mathcal{A}_0$ our geometric intuition of length is preserved. Now by Carathéodory's extension theorem there exists a measure μ on $(\Omega, \mathcal{B})$ extending μ_0 on $(\Omega, \mathcal{A}_0)$. This μ is called Lebesgue measure.
With the same approach we can generalise:

(i) the area of rectangles $R = (a_1, b_1) \times (a_2, b_2)$ – with or without any of its perimeter included – given by $\mu(R) = (b_1 - a_1) \times (b_2 - a_2)$ to Lebesgue measure on Borel sets in $\mathbb{R}^2$;
(ii) the volume of cuboids $C = (a_1, b_1) \times (a_2, b_2) \times (a_3, b_3)$ given by

$$\mu(C) = (b_1 - a_1) \cdot (b_2 - a_2) \cdot (b_3 - a_3)$$

to Lebesgue measure on Borel sets in $\mathbb{R}^3$;
(iii) and similarly in k-dimensional Euclidean space $\mathbb{R}^k$. We start with the formula for a k-dimensional box,

$$\mu\left(\prod_{i=1}^{k}(a_i, b_i)\right) = \prod_{i=1}^{k}(b_i - a_i),$$

and obtain Lebesgue measure μ, defined on $\mathcal{B}$, in $\mathbb{R}^k$.

We are mostly concerned with a special class of measures:

Definition 1.1.5. *A measure $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{A})$ is called a probability measure if*

$$\mathbb{P}(\Omega) = 1.$$

The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is called a probability space.

Observe that the above lemma and Carathéodory's extension theorem guarantee uniqueness if we construct a probability measure using the above procedure. For example the unit cube $[0,1]^k$ in $\mathbb{R}^k$ has (Lebesgue) measure 1. Using $\Omega = [0,1]^k$ as the underlying set in the above construction we find a unique probability (which equals length/area/volume if $k = 1/2/3$).
If a property holds everywhere except on a set of measure zero, we say it holds *almost everywhere* (a.e.). If it holds everywhere except on a set of probability zero, we say it holds *almost surely* (a.s.) (or, with probability one).
Roughly speaking, one uses addition in countable (or finite) situations, integration in uncountable ones. As the key measure-theoretic axiom of countable additivity above concerns addition, countably infinite situations (such as we meet in discrete time) fit well with measure theory. By contrast, uncountable situations (such as we meet in continuous time) do not – or at least, are considerably harder to handle. This is why the discrete-time setting is easier than, and precedes, the continuous-time setting. Our strategy is to do as much as possible to introduce the key ideas – economic, financial and mathematical – in discrete time (which, because we work with a finite time-horizon, the expiry time T, is actually a finite situation), before treating the harder case of continuous time.

1.2 Integral

Let $(\Omega, \mathcal{A})$ be a measurable space. We want to define integration for a suitable class of real-valued functions.

Definition 1.2.1. *Let $f : \Omega \to \mathbb{R}$. For $A \subset \mathbb{R}$ define $f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in A\}$. f is called ($\mathcal{A}$-) measurable if*

$$f^{-1}(B) \in \mathcal{A} \text{ for all } B \in \mathcal{B}.$$

Let μ be a measure on $(\Omega, \mathcal{A})$. Our aim now is to define, for suitable measurable functions, the (Lebesgue) integral with respect to μ. We will denote this integral by

$$\mu(f) = \int_\Omega f d\mu = \int_\Omega f(\omega)\mu(d\omega).$$

We start with the simplest functions. If $A \in \mathcal{A}$ the indicator function $\mathbf{1}_A(\omega)$ is defined by

$$\mathbf{1}_A(\omega) = \begin{array}{ll} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A. \end{array}$$

Then define $\mu(\mathbf{1}_A) = \mu(A)$.
The next step extends the definition to simple functions. A function f is called *simple* if it is a finite linear combination of indicators: $f = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ for constants c_i and indicator functions $\mathbf{1}_{A_i}$ of measurable sets A_i. One then extends the definition of the integral from indicator functions to simple functions by linearity:

$$\mu\left(\sum_{i=1}^{n} c_i \mathbf{1}_{A_i}\right) := \sum_{i=1}^{n} c_i \mu(\mathbf{1}_{A_i}) = \sum_{i=1}^{n} c_i \mu(A_i),$$

for constants c_i and indicators of measurable sets A_i.
If f is a non-negative measurable function, we define

$$\mu(f) := \sup\{\mu(f_0) : f_0 \text{ simple}, f_0 \leq f\}.$$

The key result in integration theory, which we must use here to guarantee that the integral for non-negative measurable functions is well-defined is:

Theorem 1.2.1 (Monotone Convergence Theorem)**.** *If (f_n) is a sequence of non-negative measurable functions such that f_n is strictly monotonic increasing to a function f (which is then also measurable), then $\mu(f_n) \to \mu(f) \leq \infty$.*

We quote that we can construct each non-negative measurable f as the increasing limit of a sequence of simple functions f_n:

$$f_n(\omega) \uparrow f(\omega) \qquad \text{for all } \omega \in \Omega \qquad (n \to \infty), \quad f_n \quad \text{simple}.$$

Using the monotone convergence theorem we can thus obtain the integral of f as

$$\mu(f) := \lim_{n\to\infty} \mu(f_n).$$

Since f_n increases in n, so does $\mu(f_n)$ (the integral is order-preserving), so either $\mu(f_n)$ increases to a finite limit, or diverges to ∞. In the first case, we say f is *(Lebesgue-) integrable* with *(Lebesgue-) integral* $\mu(f) = \lim \mu(f_n)$.
Finally if f is a measurable function that may change sign, we split it into its positive and negative parts, $f_\pm$:

$$\begin{array}{ll} f_+(\omega) := \max(f(\omega), 0), & f_-(\omega) := -\min(f(\omega), 0), \\ f(\omega) = f_+(\omega) - f_-(\omega), & |f(\omega)| = f_+(\omega) + f_-(\omega). \end{array}$$

If both f_+ and f_- are integrable, we say that f is too, and define

$$\mu(f) := \mu(f_+) - \mu(f_-).$$

Thus, in particular, $|f|$ is also integrable, and

$$\mu(|f|) = \mu(f_+) + \mu(f_-).$$

The Lebesgue integral thus defined is, by construction, an absolute integral: f is integrable iff $|f|$ is integrable. Thus, for instance, the well-known formula

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}$$

has no meaning for Lebesgue integrals, since $\int_1^\infty \frac{|\sin x|}{x} dx$ diverges to $+\infty$ like $\int_1^\infty \frac{1}{x} dx$. It has to be replaced by the limit relation

$$\int_0^X \frac{\sin x}{x} dx \to \frac{\pi}{2} \quad (X \to \infty).$$

The class of (Lebesgue-) integrable functions f on Ω is written $\mathcal{L}(\Omega)$ or (for reasons explained below) $\mathcal{L}^1(\Omega)$ – abbreviated to $\mathcal{L}^1$ or $\mathcal{L}$.
For $p \geq 1$, the $\mathcal{L}^p$ space $\mathcal{L}^p(\Omega)$ on Ω is the space of measurable functions f with $\mathcal{L}^p$-norm

$$\|f\|_p := \left(\int_\Omega |f|^p \, d\mu \right)^{\frac{1}{p}} < \infty.$$

The case $p = 2$ gives $\mathcal{L}^2$, which is particular important as it is a *Hilbert space* (Appendix A).
Turning now to the special case $\Omega = \mathbb{R}^k$ we recall the well-known *Riemann integral*. Mathematics undergraduates are taught the Riemann integral (G.B. Riemann (1826–1866)) as their first rigorous treatment of integration theory – essentially this is just a rigorisation of the school integral. It is much easier to set up than the Lebesgue integral, but much harder to manipulate.
For finite intervals $[a, b]$,we quote:

(i) for any function f Riemann-integrable on $[a, b]$, it is Lebesgue-integrable to the same value (but many more functions are Lebesgue integrable);
(ii) f is Riemann-integrable on $[a, b]$ iff it is continuous a.e. on $[a, b]$. Thus the question, 'Which functions are Riemann-integrable?' cannot be answered without the language of measure theory – which gives one the technically superior Lebesgue integral anyway.

Suppose that $F(x)$ is a non-decreasing function on $\mathbb{R}$:

$$F(x) \leq F(y) \qquad \text{if} \quad x \leq y.$$

Such functions can have at most countably many discontinuities, which are at worst jumps. We may without loss redefine F at jumps so as to be right-continuous. We now generalise the starting points above:

- Measure. We take $\mu((a, b]) := F(b) - F(a)$.
- Integral. We have $\mu(\mathbf{1}_{(a,b]}) = \mu((a, b]) = F(b) - F(a)$.

We may now follow through the successive extension procedures used above. We obtain:

- Lebesgue-Stieltjes measure μ_F,

- Lebesgue-Stieltjes integral $\mu_F(f) = \int f d\mu_F$, or even $\int f dF$.

The approach generalises to higher dimensions; we omit further details.
If instead of being monotone non-decreasing, F is the difference of two such functions, $F = F_1 - F_2$, we can define the integrals $\int f dF_1$, $\int f dF_2$ as above, and then define

$$\int f dF = \int f d(F_1 - F_2) := \int f dF_1 - \int f dF_2.$$

If $[a, b]$ is a finite interval and F is defined on $[a, b]$, a finite collection of points, $x_0, x_1, \ldots, x_n$ with $a = x_0 < x_1 < \cdots < x_n = b$, is called a *partition* of $[a, b]$, $\mathcal{P}$ say. The sum $\sum_{i=1}^{n} |F(x_i) - F(x_{i-1})|$ is called the variation of F over the partition. The least upper bound of this over all partitions $\mathcal{P}$ is called the *variation* of F over the interval $[a, b]$, $V_a^b(F)$:

$$V_a^b(F) := \sup_{\mathcal{P}} \sum |F(x_i) - F(x_{i-1})|.$$

This may be $+\infty$; but if $V_a^b(F) < \infty$, F is said to be of *bounded variation* on $[a, b]$, $F \in BV_a^b$. If F is of bounded variation on all finite intervals, F is said to be locally of bounded variation, $F \in BV_{\text{loc}}$; if F is of bounded variation on the real line $\mathbb{R}$, F is of bounded variation, $F \in BV$.
We quote that the following two properties are equivalent:

(i) F is locally of bounded variation,
(ii) F can be written as the difference $F = F_1 - F_2$ of two monotone functions.
So the above procedure defines the integral $\int f dF$ when the integrator F is of bounded variation.

Remark 1.2.1. *(i) When we pass from discrete to continuous time, we will need to handle both 'smooth' paths and paths that vary by jumps – of bounded variation – and 'rough' ones – of unbounded variation but bounded quadratic variation;*
(ii) The Lebesgue-Stieltjes integral $\int g(x)dF(x)$ is needed to express the expectation $\mathbb{E}g(X)$, where X is random variable with distribution function F and g a suitable function.

1.3 Probability

As we remarked in the introduction of this chapter, the mathematical theory of probability can be traced to 1654, to correspondence between Pascal (1623–1662) and Fermat (1601–1665). However, the theory remained both incomplete and non-rigorous until the 20th century. It turns out that the Lebesgue theory of measure and integral sketched above is exactly the machinery needed to construct a rigorous theory of probability adequate for modelling reality (option pricing, etc.) for us. This was realised by Kolmogorov (1903-1987), whose classic book of 1933, *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Foundations of Probability Theory), [Kol33], inaugurated the modern era in probability.
Recall from your first course on probability that, to describe a random experiment mathematically, we begin with the *sample space* Ω, the set of all possible

outcomes. Each point ω of Ω, or *sample point*, represents a possible – random – outcome of performing the random experiment. For a set $A \subseteq \Omega$ of points ω we want to know the probability $I\!P(A)$ (or $\Pr(A), \mathrm{pr}(A)$). We clearly want

(i) $I\!P(\emptyset) = 0, I\!P(\Omega) = 1$,

(ii) $I\!P(A) \geq 0$ for all A,

(iii) If $A_1, A_2, \ldots, A_n$ are disjoint, $I\!P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n I\!P(A_i)$ (finite additivity), which, as above we will strengthen to

(iii)* If $A_1, A_2 \ldots$ (*ad inf.*) are disjoint,

$$I\!P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} I\!P(A_i) \qquad \text{(countable additivity)}.$$

(iv) If $B \subseteq A$ and $I\!P(A) = 0$, then $I\!P(B) = 0$ (completeness).

Then by (i) and (iii) (with $A = A_1,\ \Omega \setminus A = A_2$),

$$I\!P(A^c) = I\!P(\Omega \setminus A) = 1 - I\!P(A).$$

So the class $\mathcal{F}$ of subsets of Ω whose probabilities $I\!P(A)$ are defined (call such A *events*) should be closed under countable, disjoint unions and complements, and contain the empty set $\emptyset$ and the whole space Ω. Therefore $\mathcal{F}$ should be a σ-algebra and $I\!P$ should be defined on $\mathcal{F}$ according to Definition 2.1.5. Repeating this:

Definition 1.3.1. *A probability space, or Kolmogorov triple, is a triple* $(\Omega, \mathcal{F}, I\!P)$ *satisfying Kolmogorov axioms (i),(ii),(iii)*, (iv) above.*

A probability space is a mathematical model of a random experiment.
Often we quantify outcomes ω of random experiments by defining a real-valued function X on Ω, i.e. $X : \Omega \to I\!R$. If such a function is measurable it is called a random variable.

Definition 1.3.2. *Let* $(\Omega, \mathcal{F}, I\!P)$ *be a probability space. A random variable (vector)* X *is a function* $X : \Omega \to I\!R$ $(X : \Omega \to I\!R^k)$ *such that* $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ *for all Borel sets* $B \in \mathcal{B}(I\!R)$ $(B \in \mathcal{B}(I\!R^k))$.

In particular we have for a random variable X that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in I\!R$. Hence we can define the *distribution function* F_X of X by

$$F_X(x) := I\!P\left(\{\omega : X(\omega) \leq x\}\right).$$

The smallest σ-algebra containing all the sets $\{\omega : X(\omega) \leq x\}$ for all real x (equivalently, $\{X < x\}$, $\{X \geq x\}$, $\{X > x\}$) is called the σ-algebra *generated* by X, written $\sigma(X)$. Thus,

$$X \text{ is } \mathcal{F} - \text{measurable (is a random variable) iff } \ \sigma(X) \subseteq \mathcal{F}.$$

The events in the σ-algebra generated by X are the events $\{\omega : X(\omega) \in B\}$, where B runs through the Borel σ-algebra on the line. When the (random) value $X(\omega)$ is *known*, we know *which* of these events have happened.

Interpretation.

Think of $\sigma(X)$ as representing *what we know when we know* X, or in other words *the information contained in* X (or in knowledge of X). This is reflected in the following result, due to J.L. Doob, which we quote:

$$\sigma(X) \subseteq \sigma(Y) \qquad \text{if and only if} \qquad X = g(Y)$$

for some measurable function g. For, knowing Y means we know $X := g(Y)$ – but not vice versa, unless the function g is one-to-one (injective), when the inverse function g^{-1} exists, and we can go back via $Y = g^{-1}(X)$.

Note.

An extended discussion of generated σ-algebras in the finite case is given in Dothan's book [Dot90], Chapter 3. Although technically avoidable, this is useful preparation for the general case, needed for continuous time.
A measure determines an integral. A probability measure $\mathbb{P}$, being a special kind of measure (a measure of total mass one) determines a special kind of integral, called an expectation.

Definition 1.3.3. *The expectation* $\mathbb{E}$ *of a random variable* X *on* $(\Omega, \mathcal{F}, \mathbb{P})$ *is defined by*

$$\mathbb{E}X := \int_\Omega X d\mathbb{P}, \quad \textit{or} \quad \int_\Omega X(\omega) d\mathbb{P}(\omega).$$

The expectation – also called the mean – describes the location of a distribution (and so is called a location parameter). Information about the scale of a distribution (the corresponding scale parameter) is obtained by considering the variance

$$Var(X) := \mathbb{E}\ (X - \mathbb{E}(X))^2\ = \mathbb{E}\ X^2\ - (\mathbb{E}X)^2.$$

If X is real-valued, say, with distribution function F, recall that $\mathbb{E}X$ is defined in your first course on probability by

$$\mathbb{E}X := \int x f(x) dx \quad \text{if } X \text{ has a density } f$$

or if X is discrete, taking values $x_n (n = 1, 2, \ldots)$ with probability function $f(x_n)(\geq 0)$ ($\sum_{x_n} f(x_n) = 1$),

$$\mathbb{E}X := \sum x_n f(x_n).$$

These two formulae are the special cases (for the density and discrete cases) of the general formula

$$\mathbb{E}X := \int_{-\infty}^{\infty} x dF(x)$$

where the integral on the right is a Lebesgue-Stieltjes integral. This in turn agrees with the definition above, since if F is the distribution function of X,

$$\int_\Omega X d\mathbb{P} = \int_{-\infty}^{\infty} x dF(x)$$

follows by the change of variable formula for the measure-theoretic integral, on applying the map $X : \Omega \to \mathbb{R}$ (we quote this: see any book on measure theory, e.g. [Dud89]).
Clearly the expectation operator $\mathbb{E}$ is linear. It even becomes multiplicative if we consider independent random variables.

Definition 1.3.4. *Random variables $X_1, \ldots, X_n$ are independent if whenever $A_i \in \mathcal{B}$ for $i = 1, \ldots n$ we have*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(\{X_i \in A_i\}).$$

Using Lemma 1.1.1 we can give a more tractable condition for independence:

Lemma 1.3.1. *In order for $X_1, \ldots, X_n$ to be independent it is necessary and sufficient that for all $x_1, \ldots x_n \in (-\infty, \infty]$,*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \leq x_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(\{X_i \leq x_i\}).$$

Now using the usual measure-theoretic steps (going from simple to integrable functions) it is easy to show:

Theorem 1.3.1 (Multiplication Theorem)**.** *If $X_1, \ldots, X_n$ are independent and $\mathbb{E}\,|X_i| < \infty$, $i = 1, \ldots, n$, then*

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}(X_i).$$

We now review the distributions we will mainly use in our models of financial markets.

Examples.

(i) *Bernoulli* distribution. Recall our arbitrage-pricing example from §1.4. There we were given a stock with price $S(0)$ at time $t = 0$. We assumed that after a period of time Δt the stock price could have only one of two values, either $S(\Delta t) = e^u S(0)$ with probability p or $S(\Delta t) = e^d S(0)$ with probability $1 - p$ ($u, d \in \mathbb{R}$). Let $R(\Delta t) = r(1)$ be a random variable modelling the logarithm of the stock return over the period $[0, \Delta t]$; then

$$\mathbb{P}(r(1) = u) = p \quad \text{and} \quad \mathbb{P}(r(1) = d) = 1 - p.$$

We say that $r(1)$ is distributed according to a Bernoulli distribution. Clearly $\mathbb{E}(r(1)) = up + d(1-p)$ and $Var(r(1)) = u^2 p + d^2(1-p) - (\mathbb{E}X)^2$.
The standard case of a Bernoulli distribution is given by choosing $u = 1, d = 0$ (which is not a very useful choice in financial modelling).
(ii) *Binomial* distribution. If we consider the logarithm of the stock return over n periods (of equal length), say over $[0, T]$, then subdividing into the periods $1, \ldots, n$ we have

$$\begin{aligned} R(T) &= \log \frac{S(T)}{S(0)} = \log \frac{S(T)}{S(T-\Delta t)} \cdots \frac{S(\Delta t)}{S(0)} \\ &= \log \frac{S(T)}{S(T-\Delta t)} + \ldots + \log \frac{S(\Delta t)}{S(0)} = r(n) + \ldots + r(1). \end{aligned}$$

Assuming that $r(i)$, $i = 1, \ldots, n$ are independent and each $r(i)$ is Bernoulli distributed as above we have that $R(T) = \sum_{i=1}^{n} r(i)$ is binomially distributed. Linearity of the expectation operator and independence yield $\mathbb{E}(R(T)) = \sum_{i=1}^{n} \mathbb{E}(r(i))$ and $Var(R(T)) = \sum_{i=1}^{n} Var(r(i))$.
Again for the standard case one would use $u = 1, d = 0$. The shorthand notation for a binomial random variable X is then $X \sim B(n, p)$ and we can compute

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \quad \mathbb{E}(X) = np, \quad Var(X) = np(1-p).$$

(iii) *Normal* distribution. As we will show in the sequel the limit of a sequence of appropriate normalised binomial distributions is the (standard) normal distribution. We say a random variable X is normally distributed with parameters μ, σ^2, in short $X \sim N(\mu, \sigma^2)$, if X has density function

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

One can show that $\mathbb{E}(X) = \mu$ and $Var(X) = \sigma^2$, and thus a normally distributed random variable is fully described by knowledge of its mean and variance.
Returning to the above example, one of the key results of this text will be that the limiting model of a sequence of financial markets with one-period asset returns modelled by a Bernoulli distribution is a model where the distribution of the logarithms of instantaneous asset returns is normal. That means $S(t + \Delta t)/S(t)$ is lognormally distributed (i.e. $\log(S(t + \Delta t)/S(t))$ is normally distributed). Although rejected by many empirical studies (see [EK95] for a recent overview), such a model seems to be the standard in use among financial practitioners (and we will call it the standard model in the following). The main arguments against using normally distributed random variables for modelling log-returns (i.e. log-normal distributions for returns) are asymmetry and (semi-) heavy tails. We know that distributions of financial asset returns are generally rather close to being symmetric around zero, but there is a definite tendency towards asymmetry. This may be explained by the fact that the markets react differently to positive as opposed to negative information (see [She96] §1.3.4). Since the normal distribution is symmetric it is not possible to incorporate this empirical fact in the standard model. Financial time series also suggest modelling by probability distributions whose densities behave for $x \to \pm\infty$ as

$$|x|^\rho \exp\{-\sigma |x|\}$$

with $\rho \in \mathbb{R}, \sigma > 0$. This means that we should replace the normal distribution with a distribution with heavier tails. Such a model like this would exhibit higher probabilities of extreme events and the passage from ordinary observations (around the mean) to extreme observations would be more sudden. Among suggested (classes of) distributions to be used to address these facts is the class of *hyperbolic* distributions (see [EK95] and §2.12 below), and more general distributions of normal inverse Gaussian type (see [BN98], [Ryd99], [Ryd97]) appear to be very promising.
(iv) *Poisson* distribution. Sometimes we want to incorporate in our model of financial markets the possibility of sudden jumps. Using the standard model

we model the asset price process by a continuous stochastic process, so we need an additional process generating the jumps. To do this we use point processes in general and the *Poisson process* in particular. For a Poisson process the probability of a jump (and no jump respectively) during a small interval Δt are approximately

$$I\!P(\nu(1) = 1) \approx \lambda \Delta t \quad \text{and} \quad I\!P(\nu(1) = 0) \approx 1 - \lambda \Delta t,$$

where λ is a positive constant called the rate or intensity. Modelling small intervals in such a way we get for the number of jumps $N(T) = \nu(1) + \ldots + \nu(n)$ in the interval $[0, T]$ the probability function

$$I\!P(N(T) = k) = \frac{e^{-\lambda T}(\lambda T)^k}{k!}, \quad k = 0, 1, \ldots$$

and we say the process $N(T)$ has a Poisson distribution with parameter λT. We can show $I\!E(N(T)) = \lambda T$ and $Var(N(T)) = \lambda T$.

Glossary.

Table 1.1 summarises the two parallel languages, measure-theoretic and probabilistic, which we have established.

Measure	Probability
Integral	Expectation
Measurable set	Event
Measurable function	Random variable
Almost-everywhere (a.e.)	Almost-surely (a.s.)

Table 1.1: Measure-theoretic and probabilistic languages

1.4 Equivalent Measures and Radon-Nikodým Derivatives

Given two measures $I\!P$ and $I\!Q$ defined on the same σ-algebra $\mathcal{F}$, we say that $I\!P$ is *absolutely continuous* with respect to $I\!Q$, written

$$I\!P << I\!Q$$

if $I\!P(A) = 0$, whenever $I\!Q(A) = 0,\ A \in \mathcal{F}$. We quote from measure theory the vitally important *Radon-Nikodým theorem:*

Theorem 1.4.1 (Radon-Nikodým)**.** *$I\!P << I\!Q$ iff there exists a ($\mathcal{F}$-) measurable function f such that*

$$I\!P(A) = \int_A f dI\!Q \quad \forall A \in \mathcal{F}.$$

(Note that since the integral of anything over a null set is zero, any $\mathbb{P}$ so representable is certainly absolutely continuous with respect to $\mathbb{Q}$ – the point is that the converse holds.)
Since $\mathbb{P}(A) = \int_A d\mathbb{P}$, this says that $\int_A d\mathbb{P} = \int_A f d\mathbb{Q}$ for all $A \in \mathcal{F}$. By analogy with the chain rule of ordinary calculus, we write $d\mathbb{P}/d\mathbb{Q}$ for f; then

$$\int_A d\mathbb{P} = \int_A \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} \quad \forall A \in \mathcal{F}.$$

Symbolically,

$$\text{if} \quad \mathbb{P} << \mathbb{Q}, \quad d\mathbb{P} = \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q}.$$

The measurable function (random variable) $d\mathbb{P}/d\mathbb{Q}$ is called the *Radon-Nikodým derivative* (RN-derivative) of $\mathbb{P}$ with respect to $\mathbb{Q}$.
If $\mathbb{P} << \mathbb{Q}$ and also $\mathbb{Q} << \mathbb{P}$, we call $\mathbb{P}$ and $\mathbb{Q}$ *equivalent* measures, written $\mathbb{P} \sim \mathbb{Q}$. Then $d\mathbb{P}/d\mathbb{Q}$ and $d\mathbb{Q}/d\mathbb{P}$ both exist, and

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = 1/\frac{d\mathbb{Q}}{d\mathbb{P}}.$$

For $\mathbb{P} \sim \mathbb{Q}$, $\mathbb{P}(A) = 0$ iff $\mathbb{Q}(A) = 0$: $\mathbb{P}$ and $\mathbb{Q}$ *have the same null sets.* Taking negations: $\mathbb{P} \sim \mathbb{Q}$ iff $\mathbb{P}, \mathbb{Q}$ have the same sets of positive measure. Taking complements: $\mathbb{P} \sim \mathbb{Q}$ iff $\mathbb{P}, \mathbb{Q}$ have the same sets of probability one (the same a.s. sets). Thus the following are equivalent:

$$\begin{array}{lll} \mathbb{P} \sim \mathbb{Q} & \text{iff} & \mathbb{P}, \mathbb{Q} \text{ have the same null sets,} \\ & \text{iff} & \mathbb{P}, \mathbb{Q} \text{ have the same a.s. sets,} \\ & \text{iff} & \mathbb{P}, \mathbb{Q} \text{ have the same sets of positive measure.} \end{array}$$

Far from being an abstract theoretical result, the Radon-Nikodým theorem is of key practical importance, in two ways:

(a) It is the key to the concept of conditioning, which is of central importance throughout,
(b) The concept of equivalent measures is central to the key idea of mathematical finance, *risk-neutrality*, and hence to its main results, the Black-Scholes formula, fundamental theorem of asset pricing, etc. The key to all this is that prices should be the discounted expected values under an equivalent martingale measure. Thus equivalent measures, and the operation of change of measure, are of central economic and financial importance. We shall return to this later in connection with the main mathematical result on change of measure, Girsanov's theorem.

1.5 Conditional expectation

For basic **events** define

$$\mathbb{P}(A|B) := \mathbb{P}(A \cap B)/\mathbb{P}(B) \quad \text{if} \quad \mathbb{P}(B) > 0. \tag{1.1}$$

From this definition, we get the **multiplication rule**

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B).$$

Using the partition equation $\mathbb{P}(B) = \sum_n \mathbb{P}(B|A_n)\mathbb{P}(A_n)$ with (A_n) a finite or countable partition of Ω, we get the **Bayes rule**

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B|A_j)}.$$

We can always write $\mathbb{P}(A) = \mathbb{E}(\mathbf{1}_A)$ with $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$ and $\mathbf{1}_A(\omega) = 0$ otherwise. Then the above can be written

$$\mathbb{E}(\mathbf{1}_A|B) = \frac{\mathbb{E}(\mathbf{1}_A\mathbf{1}_B)}{\mathbb{P}(B)} \tag{1.2}$$

This suggest defining, for suitable random variables X, the $\mathbb{P}$-average of X over B as

$$\mathbb{E}(X|B) = \frac{\mathbb{E}(X\mathbf{1}_B)}{\mathbb{P}(B)}. \tag{1.3}$$

Consider now **discrete** random variables X and Y. Assume X takes values $x_1, \ldots, x_m$ with probabilities $f_1(x_i) > 0$, Y takes values $y_1, \ldots, y_n$ with probabilities $f_2(y_j) > 0$, while the vector (X, Y) takes values (x_i, y_j) with probabilities $f(x_i, y_j) > 0$. Then the **marginal distributions** are

$$f_1(x_i) = \sum_{j=1} f(x_i, y_j) \quad \text{and} \quad f_2(y_j) = \sum_{i=1} f(x_i, y_j).$$

We can use the standard definition above for the events $\{Y = y_j\}$ and $\{X = x_i\}$ to get

$$\mathbb{P}(Y = y_j|X = x_i) = \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(X = x_i)} = \frac{f(x_i, y_j)}{f_1(x_i)}.$$

Thus conditional on $X = x_i$ (given the information $X = x_i$), Y takes on the values $y_1, \ldots, y_n$ with (conditional) probabilities

$$f_{Y|X}(y_j|x_i) = \frac{f(x_i, y_j)}{f_1(x_i)}.$$

So we can compute its expectation as usual:

$$\mathbb{E}(Y|X = x_i) = \sum_j y_j f_{Y|X}(y_j|x_i) = \frac{\sum_j y_j f(x_i, y_j)}{f_1(x_i)}.$$

Now define the random variable $Z = \mathbb{E}(Y|X)$, the conditional expectation of Y given X, as follows:

$$\text{if } X(\omega) = x_i, \text{ then } Z(\omega) = \mathbb{E}(Y|X = x_i) = z_i \text{ (say)}$$

Observe that in this case Z is given by a 'nice' function of X. However, a more abstract property also holds true. Since Z is constant on the the sets $\{X = x_i\}$ it is $\sigma(X)$-measurable (these sets generate the σ-algebra). Furthermore

$$\begin{aligned} \int_{\{X=x_i\}} Z d\mathbb{P} &= z_i \mathbb{P}(X = x_i) = \sum_j y_j f_{Y|X}(y_j|x_i)\mathbb{P}(X = x_i) \\ &= \sum_j y_j \mathbb{P}(Y = y_j; X = x_i) = \int_{\{X=x_i\}} Y d\mathbb{P}. \end{aligned}$$

Since the $\{X = x_i\}$ generate $\sigma(X)$, this implies

$$\int_G Z d\mathbb{P} = \int_G Y d\mathbb{P} \quad \forall\, G \in \sigma(X).$$

Density case. If the random vector (X, Y) has density $f(x, y)$, then X has (marginal) density $f_1(x) := \int_{-\infty}^{\infty} f(x, y)dy$, Y has (marginal) density $f_2(y) := \int_{-\infty}^{\infty} f(x, y)dx$. The conditional density of Y given $X = x$ is:

$$f_{Y|X}(y|x) := \frac{f(x, y)}{f_1(x)}.$$

Its expectation is

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \frac{\int_{-\infty}^{\infty} y f(x, y)dy}{f_1(x)}.$$

So we define

$$c(x) = \begin{cases} \mathbb{E}(Y|X = x) & \text{if } f_1(x) > 0 \\ 0 & \text{if } f_1(x) = 0, \end{cases}$$

and call $c(X)$ the conditional expectation of Y given X, denoted by $\mathbb{E}(Y|X)$. Observe that on sets with probability zero (i.e $\{\omega : X(\omega) = x; f_1(x) = 0\}$) the choice of $c(x)$ is arbitrary, hence $\mathbb{E}(Y|X)$ is only defined up to a set of probability zero; we speak of different versions in such cases. With this definition we again find

$$\int_G c(X)d\mathbb{P} = \int_G Y d\mathbb{P} \quad \forall\, G \in \sigma(X).$$

Indeed, for sets G with $G = \{\omega : X(\omega) \in B\}$ with B a Borel set, we find by Fubini's theorem

$$\begin{aligned} \int_G c(X)d\mathbb{P} &= \int_{-\infty}^{\infty} \mathbf{1}_B(x)c(x)f_1(x)dx \\ &= \int_{-\infty}^{\infty} \mathbf{1}_B(x)f_1(x) \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_B(x) y f(x, y)dydx = \int_G Y d\mathbb{P}. \end{aligned}$$

Now these sets G generate $\sigma(X)$ and by a standard technique (the π-systems lemma, see [Wil01], §2.3) the claim is true for all $G \in \sigma(X)$.

Example. Bivariate Normal Distribution,

$N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

$$\mathbb{E}(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),$$

the familiar regression line of statistics (linear model).

General case. Here, we follow Kolmogorov's construction using the Radon-Nikodým theorem. Suppose that $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, $\mathcal{G} \subset \mathcal{F}$. If Y is a non-negative random variable with $\mathbb{E}Y < \infty$, then

$$\mathbb{Q}(G) := \int_G Y d\mathbb{P} \qquad (G \in \mathcal{G})$$

is non-negative, σ-additive – because

$$\int_G Y d\mathbb{P} = \sum_n \int_{G_n} Y d\mathbb{P}$$

if $G = \cup_n G_n$, G_n disjoint – and defined on the σ-algebra $\mathcal{G}$, so it is a measure on $\mathcal{G}$.

If $\mathbb{P}(G) = 0$, then $\mathbb{Q}(G) = 0$ also (the integral of anything over a null set is zero), so $\mathbb{Q} \ll \mathbb{P}$.

By the Radon-Nikodým theorem, there exists a Radon-Nikodým derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ on $\mathcal{G}$, which is $\mathcal{G}$-measurable. Following Kolmogorov, we call this Radon-Nikodým derivative the conditional expectation of Y given (or conditional on) $\mathcal{G}$, $\mathbb{E}(Y|\mathcal{G})$, whose existence we now have established. For Y that changes sign, split into $Y = Y^+ - Y^-$, and define $\mathbb{E}(Y|\mathcal{G}) := \mathbb{E}(Y^+|\mathcal{G}) - \mathbb{E}(Y^-|\mathcal{G})$. We summarize:

Definition 1.5.1. *Let Y be a random variable with $\mathbb{E}(|Y|) < \infty$ and $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. We call a random variable Z a version of the conditional expectation $\mathbb{E}(Y|\mathcal{G})$ of Y given $\mathcal{G}$, and write $Z = \mathbb{E}(Y|\mathcal{G})$, a.s., if*

(i) Z is $\mathcal{G}$-measurable;

(ii) $\mathbb{E}(|Z|) < \infty$;

(iii) for every set G in $\mathcal{G}$, we have

$$\int_G Y d\mathbb{P} = \int_G Z d\mathbb{P} \qquad \forall G \in \mathcal{G}. \tag{1.4}$$

Notation. Suppose $\mathcal{G} = \sigma(X_1, \ldots, X_n)$. Then

$$\mathbb{E}(Y|\mathcal{G}) = \mathbb{E}\left(Y|\sigma(X_1, \ldots, X_n)\right) =: \mathbb{E}(Y|X_1, \ldots, X_n),$$

and one can compare the general case with the motivating examples above.

To see the intuition behind conditional expectation, consider the following situation. Assume an experiment has been performed, i.e. $\omega \in \Omega$ has been realized. However, the only information we have is the set of values $X(\omega)$ for every $\mathcal{G}$-measurable random variable X. Then $Z(\omega) = \mathbb{E}(Y|\mathcal{G})(\omega)$ is the expected value of $Y(\omega)$ given this information.

We used the traditional approach to define conditional expectation via the Radon-Nikodým theorem. Alternatively, one can use Hilbert space projection theory ([Nev75] and [JP00] follow this route). Indeed, for $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ one can show that the conditional expectation $Z = \mathbb{E}(Y|\mathcal{G})$ is the least-squares-best $\mathcal{G}$-measurable predictor of Y: amongst all $\mathcal{G}$-measurable random variables it minimises the quadratic distance, i.e.

$$\mathbb{E}[(Y - \mathbb{E}(Y|\mathcal{G}))^2] = \min\{\mathbb{E}[(Y - X)^2] : X \quad \mathcal{G} - \text{measurable}\}.$$

Note.

1. To check that something is a conditional expectation: we have to check that it integrates the right way over the right sets (i.e., as in (1.4)).
2. From (1.4): if two things integrate the same way over all sets $B \in \mathcal{G}$, they have the same conditional expectation given $\mathcal{G}$.
3. For notational convenience, we shall pass between $I\!E(Y|\mathcal{G})$ and $I\!E_{\mathcal{G}}Y$ at will.
4. The conditional expectation thus defined coincides with any we may have already encountered – in regression or multivariate analysis, for example. However, this may not be immediately obvious. The conditional expectation defined above – via σ-algebras and the Radon-Nikodým theorem – is rightly called by Williams ([Wil91], p.84) 'the central definition of modern probability'. It may take a little getting used to. As with all important but non-obvious definitions, it proves its worth in action!

We now discuss the fundamental properties of conditional expectation. From the definition linearity of conditional expectation follows from the linearity of the integral. Further properties are given by

Proposition 1.5.1. *1.* $\mathcal{G} = \{\emptyset, \Omega\}, \quad I\!E(Y|\{\emptyset, \Omega\}) = I\!EY.$

2. If $\mathcal{G} = \mathcal{F}, \quad I\!E(Y|\mathcal{F}) = Y \quad I\!P - a.s..$

3. If Y *is* $\mathcal{G}$*-measurable,* $I\!E(Y|\mathcal{G}) = Y \quad I\!P - a.s..$

4. Positivity. If $X \geq 0$*, then* $I\!E(X|\mathcal{G}) \geq 0 \quad I\!P - a.s..$

5. Taking out what is known. If Y *is* $\mathcal{G}$*-measurable and bounded,* $I\!E(YZ|\mathcal{G}) = Y\,I\!E(Z|\mathcal{G}) \quad I\!P - a.s..$

6. Tower property. If $\mathcal{G}_0 \subset \mathcal{G}$*,* $I\!E[I\!E(Y|\mathcal{G})|\mathcal{G}_0] = I\!E[Y|\mathcal{G}_0] \quad a.s..$

7. Conditional mean formula. $I\!E[I\!E(Y|\mathcal{G})] = I\!EY \quad I\!P - a.s.$

8. Role of independence. If Y *is independent of* $\mathcal{G}$*,* $I\!E(Y|\mathcal{G}) = I\!EY \quad a.s.$

9. Conditional Jensen formula. If $c : I\!R \to I\!R$ *is convex, and* $I\!E|c(X)| < \infty$*, then*

$$I\!E(c(X)|\mathcal{G}) \geq c\,(I\!E(X|\mathcal{G}))\,.$$

Proof. 1. Here $\mathcal{G} = \{\emptyset, \Omega\}$ is the *smallest* possible σ-algebra (*any* σ-algebra of subsets of Ω contains $\emptyset$ and Ω), and represents 'knowing nothing'. We have to check (1.4) for $G = \emptyset$ and $G = \Omega$. For $G = \emptyset$ both sides are zero; for $G = \Omega$ both sides are $I\!EY$.
2. Here $\mathcal{G} = \mathcal{F}$ is the *largest* possible σ-algebra, and represents 'knowing everything'. We have to check (1.4) for *all* sets $G \in \mathcal{F}$. The only integrand that integrates like Y over *all* sets is Y itself, or a function agreeing with Y except on a set of measure zero.
Note. When we condition on $\mathcal{F}$ ('knowing everything'), we *know* Y (because we know everything). There is thus no uncertainty left in Y to average out, so taking the conditional expectation (averaging out remaining randomness) has no effect, and leaves Y unaltered.
3. Recall that Y is *always* $\mathcal{F}$-measurable (this is the definition of Y being a random variable). For $\mathcal{G} \subset \mathcal{F}$, Y may not be $\mathcal{G}$-measurable, but if it is, the

proof above applies with $\mathcal{G}$ in place of $\mathcal{F}$.
Note. To say that Y is $\mathcal{G}$-measurable is to say that Y is known given $\mathcal{G}$ – that is, when we are conditioning on $\mathcal{G}$. Then Y is no longer random (being known when $\mathcal{G}$ is given), and so counts as a constant when the conditioning is performed.
4. Let Z be a version of $I\!E(X|\mathcal{G})$. If $I\!P(Z<0)>0$, then for some n, the set

$$G := \{Z < -n^{-1}\} \in \mathcal{G} \quad \text{and} \quad I\!P(\{Z<-n^{-1}\}) > 0.$$

Thus

$$0 \le I\!E(X\mathbf{1}_G) = I\!E(Z\mathbf{1}_G) < -n^{-1}I\!P(G) < 0,$$

which contradicts the positivity of X.
5. First, consider the case when Y is discrete. Then Y can be written as

$$Y = \sum_{n=1}^{\infty} b_n \mathbf{1}_{B_n},$$

for constants b_n and events $B_n \in \mathcal{G}$. Then for any $B \in \mathcal{G}$, $B \cap B_n \in \mathcal{G}$ also (as $\mathcal{G}$ is a σ-algebra), and using linearity and (1.4):

$$\begin{aligned}
\int_B Y I\!E(Z|\mathcal{G})dI\!P &= \int_B \Big(\sum_{n=1}^{\infty} b_n \mathbf{1}_{B_n}\Big) I\!E(Z|\mathcal{G})dI\!P = \sum_{n=1}^{\infty} b_n \int_{B\cap B_n} I\!E(Z|\mathcal{G})dI\!P \\
&= \sum_{n=1}^{\infty} b_n \int_{B\cap B_n} Z dI\!P = \int_B \sum_{n=1}^{\infty} b_n \mathbf{1}_{B_n} Z dI\!P \\
&= \int_B YZ dI\!P.
\end{aligned}$$

Since this holds for all $B \in \mathcal{G}$, the result holds by (1.4).
For the general case, we approximate to a general random variable Y by a sequence of discrete random variables Y_n, for each of which the result holds as just proved. We omit details of the proof here, which involves the standard approximation steps based on the monotone convergence theorem from measure theory (see e.g. [Wil91], p.90, proof of (j)). We are thus left to show the $I\!E(|ZY|) < \infty$, which follows from the assumption that Y is bounded and $Z \in \mathcal{L}^1$.
6. $I\!E_{\mathcal{G}_0} I\!E_{\mathcal{G}} Y$ is $\mathcal{G}_0$-measurable, and for $C \in \mathcal{G}_0 \subset \mathcal{G}$, using the definition of $I\!E_{\mathcal{G}_0}, I\!E_{\mathcal{G}}$:

$$\int_C I\!E_{\mathcal{G}_0}[I\!E_{\mathcal{G}} Y]dI\!P = \int_C I\!E_{\mathcal{G}} Y dI\!P = \int_C Y dI\!P.$$

So $I\!E_{\mathcal{G}_0}[I\!E_{\mathcal{G}} Y]$ satisfies the defining relation for $I\!E_{\mathcal{G}_0} Y$. Being also $\mathcal{G}_0$-measurable, it *is* $I\!E_{\mathcal{G}_0} Y$ (a.s.).
We also have:
6‘. If $\mathcal{G}_0 \subset \mathcal{G}$, $I\!E[I\!E(Y|\mathcal{G}_0)|\mathcal{G}] = I\!E[Y|\mathcal{G}_0]$ *a.s.*.
Proof. $I\!E[Y|\mathcal{G}_0]$ is $\mathcal{G}_0$-measurable, so $\mathcal{G}$-measurable as $\mathcal{G}_0 \subset \mathcal{G}$, so $I\!E[.|\mathcal{G}]$ has no effect on it, by 3.

Note.

6, 6‘ are the two forms of the *iterated conditional expectations property*. When conditioning on two σ-algebras, one larger (finer), one smaller (coarser), the coarser rubs out the effect of the finer, either way round. This may be thought of as the *coarse-averaging property*: we shall use this term interchangeably with the iterated conditional expectations property ([Wil91] uses the term *tower property*).

7. Take $\mathcal{G}_0 = \{\emptyset, \Omega\}$ in 6. and use 1.
8. If Y is independent of $\mathcal{G}$, Y is independent of $\mathbf{1}_B$ for every $B \in \mathcal{G}$. So by (1.4) and linearity,

$$\begin{aligned}\int_B \mathbb{E}(Y|\mathcal{G})d\mathbb{P} &= \int_B Y d\mathbb{P} = \int_\Omega \mathbf{1}_B Y d\mathbb{P} \\ &= \mathbb{E}(\mathbf{1}_B Y) = \mathbb{E}(\mathbf{1}_B)\mathbb{E}(Y) = \int_B \mathbb{E}Y d\mathbb{P},\end{aligned}$$

using the multiplication theorem for independent random variables. Since this holds for all $B \in \mathcal{G}$, the result follows by (1.4).
9. Recall (see e.g. [Wil91], §6.6a, §9.7h, §9.8h), that for every convex function there exists a countable sequence $((a_n, b_n))$ of points in $\mathbb{R}^2$ such that

$$c(x) = \sup_n (a_n x + b_n), \quad x \in \mathbb{R}.$$

For each fixed n we use 4. to see from $c(X) \geq a_n X + b_n$ that

$$\mathbb{E}[c(X)|\mathcal{G}] \geq a_n \mathbb{E}(X|\mathcal{G}) + b_n.$$

So,

$$\mathbb{E}[c(X)|\mathcal{G}] \geq \sup_n \left(a_n \mathbb{E}(X|\mathcal{G}) + b_n\right) = c\left(\mathbb{E}(X|\mathcal{G})\right).$$

■

Remark 1.5.1. *If in* $6, 6'$ *we take* $\mathcal{G} = \mathcal{G}_0$*, we obtain:*

$$\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{G}] = \mathbb{E}(X|\mathcal{G}).$$

Thus the map $X \to \mathbb{E}(X|\mathcal{G})$ *is idempotent: applying it twice is the same as applying it once. Hence we may identify the conditional expectation operator as a projection.*

1.6 Modes of Convergence

So far, we have dealt with one probability measure – or its expectation operator – at a time. We shall, however, have many occasions to consider a whole sequence of them, converging (in a suitable sense) to some limiting probability measure. Such situations arise, for example, whenever we approximate a financial model in continuous time (such as the continuous-time Black-Scholes model) by a sequence of models in discrete time (such as the discrete-time Black-Scholes model).

In the stochastic-process setting – such as the passage from discrete to continuous Black-Scholes models mentioned above – we need concepts beyond those we have to hand, which we develop later. We confine ourselves here to setting out what we need to discuss convergence of random variables, in the various senses that are useful.
The first idea that occurs to one is to use the ordinary convergence concept in this new setting, of random variables: then if X_n, X are random variables,

$$X_n \to X \quad (n \to \infty)$$

would be taken literally – as if the X_n, X were non-random. For instance, if X_n is the observed frequency of heads in a long series of n independent tosses of a fair coin, $X = 1/2$ the expected frequency, then the above in this case would be the man-in-the-street's idea of the 'law of averages'. It turns out that the above statement is *false* in this case, taken literally: some qualification is needed. However, the qualification needed is absolutely the minimal one imaginable: one merely needs to exclude a set of probability zero – that is, to assert convergence on a set of probability one ('almost surely'), rather than everywhere.

Definition 1.6.1. *If X_n, X are random variables, we say X_n converges to X almost surely –*

$$X_n \to X \quad (n \to \infty) \quad a.s.$$

– if $X_n \to X$ with probability one – that is, if

$$I\!P(\{\omega : X_n(\omega) \to X(\omega) \quad as \quad n \to \infty\}) = 1.$$

The loose idea of the 'law of averages' has as its precise form a statement on convergence almost surely. This is Kolmogorov's *strong law of large numbers* (see e.g. [Wil91], §12.10), which is quite difficult to prove.
Weaker convergence concepts are also useful: they may hold under weaker conditions, or they may be easier to prove.

Definition 1.6.2. *If X_n, X are random variables, we say that X_n converges to X in probability -*

$$X_n \to X \quad (n \to \infty) \quad in\ probability$$

- if, for all $\epsilon > 0$,

$$I\!P\left(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\right) \to 0 \quad (n \to \infty).$$

It turns out that convergence almost surely implies convergence in probability, but not in general conversely. Thus almost-sure convergence is a stronger convergence concept than convergence in probability. This comparison is reflected in the form the 'law of averages' takes for convergence in probability: this is called the *weak law of large numbers*, which as its name implies is a weaker form of the strong law of large numbers. It is correspondingly much easier to prove: indeed, we shall prove it below.
Recall the L^p-spaces of pth-power integrable functions. We similarly define the L^p-spaces of pth-power integrable random variables: if $p \geq 1$ and X is a random variable with

$$\|X\|_p := (I\!E|X|^p)^{1/p} < \infty,$$

we say that $X \in L^p$ (or $L^p(\Omega, \mathcal{F}, \mathbb{P})$ to be precise). For X_n, $X \in L^p$, there is a natural convergence concept: we say that X_n *converges to* X *in* L^p, or *in* pth *mean*,

$$X_n \to X \quad \text{in} \quad L^p,$$

if

$$\|X_n - X\|_p \to 0 \quad (n \to \infty),$$

that is, if

$$\mathbb{E}(|X_n - X|^p) \to 0 \quad (n \to \infty).$$

The cases $p = 1, 2$ are particularly important: if $X_n \to X$ in L^1, we say that $X_n \to X$ *in mean*; if $X_n \to X$ in L^2 we say that $X_n \to X$ *in mean square*. Convergence in pth mean is not directly comparable with convergence almost surely (of course, we have to restrict to random variables in L^p for the comparison even to be meaningful): neither implies the other. Both, however, imply convergence in probability.
All the modes of convergence discussed so far involve the *values* of random variables. Often, however, it is only the *distributions* of random variables that matter. In such cases, the natural mode of convergence is the following:

Definition 1.6.3. *We say that random variables X_n converge to X in distribution if the distribution functions of X_n converge to that of X at all points of continuity of the latter:*

$$X_n \to X \quad \textit{in distribution}$$

if

$$\mathbb{P}(\{X_n \leq x\}) \to \mathbb{P}(\{X \leq x\}) \quad (n \to \infty)$$

for all points x at which the right-hand side is continuous.

The restriction to continuity points x of the limit seems awkward at first, but it is both natural and necessary. It is also quite weak: note that the function $x \mapsto \mathbb{P}(\{X \leq x\})$, being monotone in x, is continuous except for at most countably many jumps. The set of continuity points is thus uncountable: 'most' points are continuity points.
Convergence in distribution is (by far) the weakest of the modes of convergence introduced so far: convergence in probability implies convergence in distribution, but not conversely. There is, however, a partial converse: if the limit X is *constant* (non-random), convergence in probability and in distribution are equivalent.

Weak Convergence.

If $\mathbb{P}_n$, $\mathbb{P}$ are probability measures, we say that

$$\mathbb{P}_n \to \mathbb{P} \quad (n \to \infty) \quad \text{weakly}$$

if

$$\int f d\mathbb{P}_n \to \int f d\mathbb{P} \quad (n \to \infty) \tag{1.5}$$

for all bounded continuous functions f. This definition is given a full-length book treatment in [Bil68], and we refer to this for background and details. For

ordinary (real-valued) random variables, weak convergence of their probability measures is the same as convergence in distribution of their distribution functions. However, the weak-convergence definition above applies equally, not just to this one-dimensional case, or to the finite-dimensional (vector-valued) setting, but also to infinite-dimensional settings such as arise in convergence of stochastic processes. We shall need such a framework in the passage from discrete- to continuous-time Black-Scholes models.

Chapter 2

Basic Probability Background

2.1 Fundamentals

To describe a random experiment we use a *sample space* Ω, the set of all possible outcomes. Each point ω of Ω, or *sample point*, represents a possible random outcome of performing the random experiment.
Examples. Write down Ω for experiments such as flip a coin three times, roll two dice.
For a set $A \subseteq \Omega$ we want to know the probability $I\!P(A)$. The class $\mathcal{F}$ of subsets of Ω whose probabilities $I\!P(A)$ are defined (call such A *events*) should be a σ-algebra , i.e.

(i) $\emptyset, \Omega \in \mathcal{F}$.

(ii) $F \in \mathcal{F}$ implies $F^c \in \mathcal{F}$.

(iii) $F_1, F_2, \ldots \in \mathcal{F}$ then $\bigcup_n F_n \in \mathcal{F}$.

We want a probability measure defined on $\mathcal{F}$

(i) $I\!P(\emptyset) = 0, I\!P(\Omega) = 1$,

(ii) $I\!P(A) \geq 0$ for all A,

(iii) If $A_1, A_2, \ldots,$ are disjoint, $I\!P(\bigcup_i A_i) = \sum_i I\!P(A_i)$ countable additivity.

Definition 2.1.1. *A probability space, or Kolmogorov triple, is a triple* $(\Omega, \mathcal{F}, I\!P)$ *satisfying Kolmogorov axioms (i),(ii) and (iii) above.*

A probability space is a mathematical model of a random experiment.
Examples. Assign probabilities for the above experiments.

Definition 2.1.2. *Let* $(\Omega, \mathcal{F}, I\!P)$ *be a probability space. A random variable (vector)* X *is a function* $X : \Omega \to I\!R(I\!R^k)$ *such that* $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ *for all Borel sets* $B \in \mathcal{B}(\mathcal{B}(I\!R^k))$.

For a random variable X

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. So define the *distribution function* F_X of X by

$$F_X(x) := \mathbb{P}(\{\omega : X(\omega) \leq x\}).$$

Recall: $\sigma(X)$, the σ-algebra *generated* by X.

Some important probability distributions

- Binomial distribution: Number of successes

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

- Geometric distribution: Waiting time

$$\mathbb{P}(N = n) = p(1-p)^{n-1}.$$

- Poisson distribution:

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

- Density of Uniform distribution:

$$f(x) = \frac{1}{b-a} \mathbf{1}_{\{(a,b)\}}.$$

- Density of Exponential distribution:

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{[0,\infty)\}}.$$

- Density of standard Normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Definition 2.1.3. *The expectation $\mathbb{E}$ of a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is defined by*

$$\mathbb{E}X := \int_\Omega X d\mathbb{P}, \quad \text{or} \quad \int_\Omega X(\omega) d\mathbb{P}(\omega).$$

The variance of a random variable is defined as

$$Var(X) := \mathbb{E}\ (X - \mathbb{E}(X))^2\ = \mathbb{E}\ X^2\ - (\mathbb{E}X)^2.$$

If X is real-valued with density f (i.e $f(x) \geq 0 : \int_{\mathbb{R}} f(x)dx = 1$),

$$\mathbb{E}X := \int_{\mathbb{Z}} xf(x)dx$$

or if X is discrete, taking values $x_n (n = 1, 2, \ldots)$ with probability function $f(x_n) (\geq 0)$,

$$\mathbb{E}X := \sum_{\mathbb{X}} x_n f(x_n).$$

Examples. Calculate moments for some of the above distributions.

Definition 2.1.4. *Random variables $X_1, \ldots, X_n$ are independent if whenever $A_i \in \mathcal{B}$ (the Borel σ-algebra) for $i = 1, \ldots n$ we have*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(\{X_i \in A_i\}).$$

Lemma 2.1.1. *In order for $X_1, \ldots, X_n$ to be independent it is necessary and sufficient that for all $x_1, \ldots x_n \in (-\infty, \infty]$,*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \leq x_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(\{X_i \leq x_i\}).$$

Theorem 2.1.1 (Multiplication Theorem)**.** *If $X_1, \ldots, X_n$ are independent and $\mathbb{E}\,|X_i| < \infty$, $i = 1, \ldots, n$, then*

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}(X_i).$$

If X, Y are independent, with distribution functions F, G, define $Z := X + Y$ with distribution function H. We call H the *convolution* of F and G, written $H = F * G$.
Suppose X, Y have densities f, g, then Z has a density h with

$$h(z) = \int_{-\infty}^{\infty} f(z-y)g(y)dy = \int_{-\infty}^{\infty} f(x)g(z-x)dx.$$

Example. Assume $t_1, \ldots, t_n$ are independent random variables that have an exponential distribution with parameter λ. Then $T = t_1 + \ldots + t_n$ has the Gamma(n, λ) density function

$$f(x) = \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x}.$$

Definition 2.1.5. *If X is a random variable with distribution function F, its moment generating function ϕ_X is*

$$\phi(t) := \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} dF(x).$$

The mgf *takes convolution into multiplication*: if X, Y are independent,

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

Observe $\phi^{(k)}(t) = \mathbb{E}(X^k e^{tX})$ and $\phi(0) = \mathbb{E}(X^k)$.
For X on nonnegative integers use the *generating function*

$$\gamma_X(z) = \mathbb{E}(z^X) = \sum_{k=0}^{\infty} z^k \mathbb{P}(Z = k).$$

2.2 Convolution and Characteristic Functions

The most basic operation on numbers is addition; the most basic operation on random variables is addition of independent random variables. If X, Y are independent, with distribution functions F, G, and

$$Z := X + Y,$$

let Z have distribution function H. Then since $X + Y = Y + X$ (addition is commutative), H depends on F and G symmetrically. We call H the *convolution* (German: *Faltung*) of F and G, written

$$H = F * G.$$

Suppose first that X, Y have densities f, g. Then

$$H(z) = I\!P(Z \leq z) = I\!P(X + Y \leq z) = \int\int_{\{(x,y):x+y\leq z\}} f(x)g(y)dxdy,$$

since by independence of X and Y the joint density of X and Y is the product $f(x)g(y)$ of their separate (marginal) densities, and to find probabilities in the density case we integrate the joint density over the relevant region. Thus

$$H(z) = \int_{-\infty}^{\infty} f(x) \left\{ \int_{-\infty}^{z-x} g(y)dy \right\} dx = \int_{-\infty}^{\infty} f(x)G(z-x)dx.$$

If

$$h(z) := \int_{-\infty}^{\infty} f(x)g(z-x)dx,$$

(and of course symmetrically with f and g interchanged), then integrating we recover the equation above (after interchanging the order of integration. This is legitimate, as the integrals are non-negative, by Fubini's theorem, which we quote from measure theory, see e.g. [Wil91], §8.2). This shows that if X, Y are independent with densities f, g, and $Z = X + Y$, then Z has density h, where

$$h(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy.$$

We write

$$h = f * g,$$

and call the density h the *convolution* of the densities f and g.
If X, Y do not have densities, the argument above may still be taken as far as

$$H(z) = I\!P(Z \leq z) = I\!P(X + Y \leq z) = \int_{-\infty}^{\infty} F(x-y)dG(y)$$

(and, again, symmetrically with F and G interchanged), where the integral on the right is the *Lebesgue-Stieltjes integral*. We again write

$$H = F * G,$$

and call the distribution function H the *convolution* of the distribution functions F and G.
In sum: *addition* of *independent* random variables corresponds to *convolution* of distribution functions or densities.
Now we frequently need to add (or average) lots of independent random variables: for example, when forming sample means in statistics – when the bigger the sample size is, the better. But convolution involves integration, so adding n independent random variables involves $n-1$ integrations, and this is awkward to do for large n. One thus seeks a way to transform distributions so as to make the awkward operation of convolution as easy to handle as the operation of addition of independent random variables that gives rise to it.

Definition 2.2.1. *If X is a random variable with distribution function F, its characteristic function ϕ (or ϕ_X if we need to emphasise X) is*

$$\phi(t) := \mathbb{E}(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} dF(x), \quad (t \in \mathbb{R}).$$

Note.

Here $i := \sqrt{-1}$. All other numbers – t, x etc. – are real; all expressions involving i such as e^{itx}, $\phi(t) = \mathbb{E}(e^{itx})$ are complex numbers.
The characteristic function *takes convolution into multiplication*: if X, Y are independent,

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

For, as X, Y are independent, so are e^{itX} and e^{itY} for any t, so by the multiplication theorem (Theorem 1.3.1),

$$\mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX} \cdot e^{itY}) = \mathbb{E}(e^{itX}) \cdot \mathbb{E}(e^{itY}),$$

as required.
We list some properties of characteristic functions that we shall need.

1. $\phi(0) = 1$. For, $\phi(0) = \mathbb{E}(e^{i \cdot 0 \cdot X}) = \mathbb{E}(e^0) = \mathbb{E}(1) = 1$.
2. $|\phi(t)| \leq 1$ for all $t \in \mathbb{R}$.
Proof. $|\phi(t)| = \int_{-\infty}^{\infty} e^{itx} dF(x) \leq \int_{-\infty}^{\infty} |e^{itx}| \, dF(x) = \int_{-\infty}^{\infty} 1 dF(x) = 1$.
Thus in particular the characteristic function *always exists* (the integral defining it is always absolutely convergent). This is a crucial advantage, far outweighing the disadvantage of having to work with complex rather than real numbers (the nuisance value of which is in fact slight).
3. ϕ is continuous (indeed, ϕ is uniformly continuous).
Proof.

$$\begin{aligned} |\phi(t+u) - \phi(t)| &= \int_{-\infty}^{\infty} \{e^{i(t+u)x} - e^{itx}\} dF(x) \\ &= \int_{-\infty}^{\infty} e^{itx}(e^{iux} - 1) dF(x) \leq \int_{-\infty}^{\infty} |e^{iux} - 1| \, dF(x), \end{aligned}$$

for all t. Now as $u \to 0$, $e^{iux} - 1 \to 0$, and $e^{iux} - 1 \leq 2$. The bound on the right tends to zero as $u \to 0$ by Lebesgue's dominated convergence theorem (which we quote from measure theory: see e.g. [Wil91], §5.9), giving continuity; the uniformity follows as the bound holds uniformly in t.
4. (*Uniqueness theorem*): ϕ determines the distribution function F uniquely. Technically, ϕ is the *Fourier-Stieltjes transform* of F, and here we are quoting the uniqueness property of this transform. Were uniqueness not to hold, we would lose information on taking characteristic functions, and so ϕ would not be useful.
5. (*Continuity theorem*): If X_n, X are random variables with distribution functions F_n, F and characteristic functions ϕ_n, ϕ, then convergence of ϕ_n to ϕ,

$$\phi_n(t) \to \phi(t) \quad (n \to \infty) \quad \text{for all} \quad t \in \mathbb{R}$$

is equivalent to convergence in distribution of X_n to X. This result is due to Lévy; see e.g. [Wil91], §18.1.
6. *Moments.* Suppose X has kth moment: $\mathbb{E}|X|^k < \infty$. Take the Taylor (power-series) expansion of e^{itx} as far as the kth power term:

$$e^{itx} = 1 + itx + \cdots + (itx)^k/k! + \boldsymbol{o}\ t^k\ ,$$

where '$\boldsymbol{o}\ t^k$' denotes an error term of smaller order than t^k for small k. Now replace x by X, and take expectations. By linearity, we obtain

$$\phi(t) = \mathbb{E}(e^{itX}) = 1 + it\mathbb{E}X + \cdots + \frac{(it)^k}{k!}\mathbb{E}(X^k) + e(t),$$

where the error term $e(t)$ is the expectation of the error terms (now random, as X is random) obtained above (one for each value $X(\omega)$ of X). It is not obvious, but it is true, that $e(t)$ is still of smaller order than t^k for $t \to 0$:

$$\text{if} \quad \mathbb{E}\ |X|^k\ < \infty, \quad \phi(t) = 1 + it\mathbb{E}(X) + \ldots + \frac{(it)^k}{k!}\mathbb{E}\ X^k\ + \boldsymbol{o}\ t^k \qquad (t \to 0).$$

We shall need the case $k = 2$ in dealing with the central limit theorem below.

Examples

1. Standard Normal Distribution,

$N(0,1)$. For the standard normal density $f(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}$, one has, by the process of 'completing the square' (familiar from when one first learns to solve quadratic equations!),

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{tx} f(x)dx &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\ tx - \frac{1}{2}x^2\ \ dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\ -\frac{1}{2}(x-t)^2 + \frac{1}{2}t^2\ \ dx \\
&= \exp\ \frac{1}{2}t^2\ \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\ -\frac{1}{2}(x-t)^2\ \ dx.
\end{aligned}$$

The second factor on the right is 1 (it has the form of a normal integral). This gives the integral on the left as $\exp\{\frac{1}{2}t^2\}$.
Now replace t by it (legitimate by analytic continuation, which we quote from complex analysis, see e.g. [BB70]). The right becomes $\exp\{-\frac{1}{2}t^2\}$. The integral on the left becomes the characteristic function of the standard normal density – which we have thus now identified (and will need below in §2.8).

2. General Normal Distribution,

$N(\mu, \sigma)$. Consider the transformation $x \mapsto \mu + \sigma x$. Applied to a random variable X, this adds μ to the mean (a change of *location*), and multiplies the variance by σ^2 (a change of *scale*). One can check that if X has the standard normal density above, then $\mu + \sigma X$ has density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \quad -\frac{1}{2}(x-\mu)^2/\sigma^2 \quad ,$$

and characteristic function

$$\begin{aligned} \mathbb{E}e^{it(\mu+\sigma X)} &= \exp\{i\mu t\}\mathbb{E} \quad e^{(i\sigma t)X} \quad = \exp\{i\mu t\} \exp \quad -\frac{1}{2}(\sigma t)^2 \\ &= \exp \quad i\mu t - \frac{1}{2}\sigma^2 t^2 \quad . \end{aligned}$$

Thus the general normal density and its characteristic function are

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \quad -\frac{1}{2}(x-\mu)^2/\sigma^2 \quad , \quad \phi(t) = \exp \quad i\mu t - \frac{1}{2}\sigma^2 t^2 \quad .$$

3. Poisson Distribution,

$P(\lambda)$. Here, the probability mass function is

$$f(k) := \mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!, \quad (k = 0, 1, 2, \ldots).$$

The characteristic function is thus

$$\begin{aligned} \phi(t) &= \mathbb{E} \quad e^{itX} \quad = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \cdot e^{itk} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} (\lambda e^{it})^k/k! = e^{-\lambda} \exp\{\lambda e^{it}\} = \exp\{-\lambda(1 - e^{it})\}. \end{aligned}$$

2.3 The Central Limit Theorem

You will be well aware that

$$1 + \frac{x}{n} \quad ^n \to e^x \quad (n \to \infty) \quad \forall x \in \mathbb{R}.$$

This is the formula governing the passage from discrete to continuous compound interest. Invest one pound (or dollar) for one year at $100x\%$ p.a.; with interest

compounded n times p.a., our capital after one year is $(1+\frac{x}{n})^n$. With continuous compounding, our capital after one year is the exponential e^x: *exponential growth corresponds to continuously compounded interest.*
We need two extensions: the formula still holds with $x \in \mathbb{R}$ replaced by a complex number $z \in \mathbb{C}$:

$$1+\frac{z}{n}\Big)^n \to e^z \quad (n\to\infty) \quad \forall z \in \mathbb{C},$$

and if $z_n \in \mathbb{C}$, $z_n \to z$,

$$1+\frac{z_n}{n}\Big)^n \to e^z \quad (n\to\infty) \quad (z_n \to z \in \mathbb{C}).$$

As a first illustration of the power of transform methods, we prove the weak law of large numbers:

Theorem 2.3.1 (Weak Law of Large Numbers)**.** *If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad (n\to\infty) \quad \textit{in probability.}$$

Proof. If the X_i have characteristic function ϕ, then by the moment property of §2.8 with $k=1$,

$$\phi(t) = 1 + i\mu t + o(t) \quad (t\to 0).$$

Now using the i.i.d. assumption, $\frac{1}{n}\sum_{i=1}^{n} X_i$ has characteristic function

$$\begin{aligned}
\mathbb{E}\left(\exp\left(it\cdot\frac{1}{n}\sum_{1}^{n} X_j\right)\right) &= \mathbb{E}\left(\prod_{i=1}^{n} \exp\left(it\cdot\frac{1}{n}X_j\right)\right) \\
&= \prod_{i=1}^{n} \mathbb{E}\left(\exp\left(\frac{it}{n}X_j\right)\right) = (\phi(t/n))^n \\
&= \left(1+\frac{i\mu t}{n}+o(1/n)\right)^n \to e^{i\mu t} \quad (n\to\infty),
\end{aligned}$$

and $e^{i\mu t}$ is the characteristic function of the constant μ (for fixed t, $o(1/n)$ is an error term of smaller order than $1/n$ as $n\to\infty$). By the continuity theorem,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{in distribution,}$$

and as μ is constant, this says (see §2.6) that

$$\frac{1}{n}\sum_{1}^{n} X_i \to \mu \quad \text{in probability.}$$

■

The main result of this section is the same argument carried one stage further.

Theorem 2.3.2 (Central Limit Theorem)**.** *If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and variance σ^2, then with $N(0,1)$ the standard normal distribution,*

$$\frac{\sqrt{n}}{\sigma}\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)/\sigma \to N(0,1) \quad (n\to\infty) \quad \textit{in distribution.}$$

That is, for all $x \in \mathbb{R}$,

$$\mathbb{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)/\sigma \le x\right\} \to \Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}y^2}dy \quad (n\to\infty).$$

Proof. We first centre at the mean. If X_i has characteristic function ϕ, let $X_i - \mu$ have characteristic function ϕ_0. Since $X_i - \mu$ has mean 0 and second moment $\sigma^2 = Var(X_i) = \mathbb{E}[(X_i - \mathbb{E}X_i)^2] = \mathbb{E}[(X_i-\mu)^2]$, the case $k = 2$ of the moment property of §2.7 gives

$$\phi_0(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2) \qquad (t \to 0).$$

Now $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu)/\sigma$ has characteristic function

$$\begin{aligned}
& \mathbb{E}\left\{\exp\left\{it\cdot\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{j=1}^{n}X_j-\mu\right)\right\}\right\} \\
= \; & \mathbb{E}\left\{\prod_{j=1}^{n}\exp\left\{\frac{it(X_j-\mu)}{\sigma\sqrt{n}}\right\}\right\} = \prod_{j=1}^{n}\mathbb{E}\left\{\exp\left\{\frac{it}{\sigma\sqrt{n}}(X_j-\mu)\right\}\right\} \\
= \; & \phi_0\left(\frac{t}{\sigma\sqrt{n}}\right)^n = \left(1-\frac{\frac{1}{2}\sigma^2t^2}{\sigma^2 n}+o\left(\frac{1}{n}\right)\right)^n \to e^{-\frac{1}{2}t^2} \quad (n\to\infty),
\end{aligned}$$

and $e^{-\frac{1}{2}t^2}$ is the characteristic function of the standard normal distribution $N(0,1)$. The result follows by the continuity theorem. ∎

Note.

In Theorem 2.3.2, we:

(i) *centre* the X_i by subtracting the mean (to get mean 0);
(ii) *scale* the resulting $X_i - \mu$ by dividing by the standard deviation σ (to get variance 1). Then if $Y_i := (X_i - \mu)/\sigma$ are the resulting *standardised* variables, $\frac{1}{\sqrt{n}}\sum_1^n Y_i$ converges in distribution to standard normal.

Example: the Binomial Case.

If each X_i is Bernoulli distributed with parameter $p \in (0,1)$,

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = 0) = q := 1 - p$$

– so X_i has mean p and variance pq - $S_n := \sum_{i=1}^{n} X_i$ is binomially distributed with parameters n and p:

$$\mathbb{P}\left\{\sum_{i=1}^{n}X_i = k\right\} = \binom{n}{k}p^k q^{n-k} = \frac{n!}{(n-k)!k!}p^k q^{n-k}.$$

A direct attack on the distribution of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - p)/\sqrt{pq}$ can be made via

$$\mathbb{P}\left(a \le \sum_{i=1}^{} X_i \le b \right) = \sum_{k: np + a\sqrt{npq} \le k \le np + b\sqrt{npq}} \frac{n!}{(n-k)!k!} p^k q^{n-k}.$$

Since n, k and $n-k$ will all be large here, one needs an approximation to the factorials. The required result is *Stirling's formula* of 1730:

$$n! \sim \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} \quad (n \to \infty)$$

(the symbol $\sim$ indicates that the ratio of the two sides tends to 1). The argument can be carried through to obtain the sum on the right as a Riemann sum (in the sense of the Riemann integral: §2.2) for $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$, whence the result. This, the earliest form of the central limit theorem, is the de Moivre-Laplace limit theorem (Abraham de Moivre, 1667–1754; P.S. de Laplace, 1749–1827). The proof of the de-Moivre-Laplace limit theorem sketched above is closely analogous to the passage from the discrete to the continuous Black-Scholes formula.

Local Limit Theorems.

The central limit theorem as proved above is a *global limit theorem*: it relates to *distributions* and convergence thereof. The de Moivre-Laplace limit theorem above, however, deals directly with individual probabilities in the discrete case (the sum of a large number of which is shown to approximate an integral). A limit theorem dealing with *densities* and convergence thereof in the density case, or with the discrete analogues of densities – such as the individual probabilities $\mathbb{P}(S_n = k)$ in the binomial case above – is called a *local limit theorem.*

Poisson Limit Theorem.

The de Moivre-Laplace limit theorem – convergence of binomial to normal – is only one possible limiting regime for binomial models. The next most important one has a *Poisson* limit in place of a normal one.
Suppose we have a sequence of binomial models $B(n,p)$, where the success probability $p = p_n$ varies with n, in such a way that

$$np_n \to \lambda \quad > 0, \quad (n \to \infty). \tag{2.1}$$

Thus $p_n \to 0$ – indeed, $p_n \sim \lambda/n$. This models a situation where we have a large number n of Bernoulli trials, each with small probability p_n of success, but such that np_n, the expected total number of successes, is 'neither large nor small, but intermediate'. Binomial models satisfying condition (2.1) converge to the Poisson model $P(\lambda)$ with parameter $\lambda > 0$.
This result is sometimes called the law of small numbers. The Poisson distribution is widely used to model statistics of accidents, insurance claims and the like, where one has a large number n of individuals at risk, each with a small probability p_n of generating an accident, insurance claim etc. ('success probability' seems a strange usage here!).

Chapter 3

Statistics Background

3.1 Simple Random Sampling

Each particular sample of size n has the same probability of occurrence (and each member of the population appears at most once).

Definition 3.1.1. *The random variables $X_1, \ldots, X_n$ are called a random sample of size n from the population $f(x)$ if $X_1, \ldots X_n$ are mutually independent rvs and the probability density (mass) function is the same function $f(x)$.*

We can compute the joint pdf of $X_1, \ldots X_n$.

$$f(x_1, \ldots, X_n) = f(x_1) \cdot \ldots \cdot f(x_n) = \prod_{i=1} f(x_i).$$

We are interested in summaries of the values of $X_1 = x_1, \ldots X_n = x_n$. Any such summary may be expressed as a suitable function $T(x_1, \ldots x_n)$.

Definition 3.1.2. *Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued (or vector-valued) function whose domain includes the sample space of $(X_1, \ldots X_n)$. Then the random variable $Y = T(X_1, \ldots X_n)$ is called a statistic. The probability distribution of a statistic Y is called the sampling distribution of Y.*

Definition 3.1.3. *The sample mean is the arithmetic average of the values in a random sample. It is denoted by*

$$\bar{X} = \frac{1}{n} \sum_{i=1} X_i.$$

Definition 3.1.4. *The sample variance is the statistic defined by*

$$S^2 = \frac{1}{n-1} \sum_{i=1} (X_i - \bar{X})^2.$$

The sample deviation is the statistic defined by

$$S = \sqrt{S^2}.$$

Theorem 3.1.1. *Let $X_1, \dots X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

(i) $I\!E(\bar{X}) = \mu$,

(ii) $Var(\bar{X}) = \frac{\sigma^2}{n}$,

(iii) $I\!E(S^2) = \sigma^2$.

The relationships (i) and (iii) mean that $\bar{X}$ is an *unbiased* estimator of μ and S^2 is an *unbiased* estimator of σ^2.

3.2 The sampling distribution of $\bar{X}$

3.2.1 Characteristic Functions

First note the following useful relationship for characteristic functions:

$$\phi_{\bar{X}}(t) = (\phi_{X_1}(t/n))^n.$$

This relation can be used to derive the sampling distribution of $\bar{X}$ from the uniqueness theorem

Example. Let $X_1, \dots X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, then $\bar{X}$ has a $N(\mu, \sigma^2/n)$ distribution.

3.2.2 Normal Approximation

In the general case, i.e. distribution of X_i unknown, we can use the Central Limit Theorem to obtain an approximation of the distribution of $\bar{X}$. By the CLT we have for a fixed number z

$$I\!P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z) \quad (n \to \infty).$$

We use the CLT to approximate the probability that the error made by estimating μ by $\bar{X}$ is less than some constant δ:

$$\begin{aligned} I\!P(\,|\bar{X} - \mu| \leq \delta) &= I\!P(-\delta \leq \bar{X} - \mu \leq \delta) \\ &= I\!P\left(-\frac{\delta}{\sigma_{\bar{X}}} \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq \frac{\delta}{\sigma_{\bar{X}}}\right) \\ &\approx \Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right) - \Phi\left(-\frac{\delta}{\sigma_{\bar{X}}}\right) \\ &= 2\Phi\left(\frac{\delta}{\sigma_{\bar{X}}}\right) - 1. \end{aligned}$$

We can now derive a *confidence interval* for the population mean μ. A confidence interval for a population parameter, θ, is a random interval, calculated from the sample, that contains θ with a specified probability. For example, a 95 % confidence interval for μ is a random interval that contains μ with probability 0.95.

For $0 \leq \alpha \leq 1$, let $z(\alpha)$ be that number such that the area under the standard normal density function to the right of $z(\alpha)$ is α. By symmetry $z(1-\alpha) = z(\alpha)$. So

$$I\!P(-z(\alpha/2) \leq Z \leq z(\alpha/2)) = 1 - \alpha.$$

Using the CLT we get

$$I\!P(-z(\alpha/2) \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq z(\alpha/2)) \approx 1 - \alpha$$

or

$$I\!P(\bar{X} - z(\alpha/2)\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z(\alpha/2)\sigma_{\bar{X}}) \approx 1 - \alpha$$

3.3 Estimation of Parameters

Many families of probability laws depend only on a small number of parameters (Poisson, Normal, Gamma, ..). These parameters must be estimated from data in order to fit the probability law. After parameters have chosen the model should be compared to the actual data to see if the fit is reasonable. We present two methods of finding such parameters and evaluate their properties.
The model setting is as follows: Observed data will be regarded as realisations of random variables $X_1, X_2, \ldots, X_n$, whose joint distribution depends on an unknown parameter θ. The X_i will be modeled as independent random variables all having the same distribution $f(x|\theta)$. An estimate of θ will be a statistic and as such be a random variable with a probability distribution called its **sampling distribution**. We will use the standard deviation (called **standard error**) of the sampling distribution to assess the variability of our estimate.

3.3.1 The Method of Moments

Let $X \sim f$, the kth moment of the probability law f is defined as

$$\mu_k = I\!E(X^k).$$

Then the kth **sample moment** is defined as

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1} X_i^k.$$

We view $\hat{\mu}_k$ as an estimate of μ_k. The method of moments estimates parameters by finding expressions for them in terms of the lowest possible order moments and then substituting sample moments into the expressions.
For example: $\theta = (\theta_1, \theta_2)$ and

$$\theta_1 = h_1(\mu_1, \mu_2) \quad \theta_2 = h_2(\mu_1, \mu_2)$$

then the method of moments estimates are

$$\hat{\theta}_1 = h_1(\hat{\mu}_1, \hat{\mu}_2) \quad \hat{\theta}_2 = h_2(\hat{\mu}_1, \hat{\mu}_2).$$

Poisson Distribution. $X \sim Po(\lambda)$. Now $I\!E(X) = \lambda$, so

$$\hat{\lambda} = \hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1} X_i.$$

What is the sampling distribution? From $X_i \sim Po(\lambda)$ we know $\sum X_i \sim Po(n\lambda)$. So

$$I\!E(\hat{\lambda}) = \frac{1}{n}n\lambda = \lambda; \quad Var(\hat{\lambda}) = \frac{1}{n^2}\, Var(\sum X_i) = \frac{1}{n}\lambda.$$

The standard error of $\hat{\lambda}$ is $\sigma_{\hat{\lambda}} = \sqrt{\lambda/n}$. Since we do not know λ we can only calculate an estimated standard error as

$$s_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}}.$$

If n is large the CLT yields $\hat{\lambda} \sim N(\lambda, \lambda/n)$.

Normal Distribution. The first and second moments are

$$\begin{aligned} \mu_1 &= I\!E(X) = \mu \\ \mu_2 &= I\!E(X^2) = \mu^2 + \sigma^2. \end{aligned}$$

Therefore

$$\begin{aligned} \mu &= \mu_1 \\ \sigma^2 &= \mu_2 - \mu_1^2. \end{aligned}$$

The estimates from the sample moments are

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1} (X_i - \bar{X})^2. \end{aligned}$$

We know $\bar{X} \sim N(\mu, \sigma^2/n), n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$. (See §2.5.2.)
Method of moment estimates can be proved to be consistent if the functions relating the estimates to the sample moments are continuous.

Definition 3.3.1. *Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be **consistent** in probability if for any $\varepsilon > 0$,*

$$I\!P(|\hat{\theta} - \theta| > \varepsilon) \to 0, \quad (n \to \infty).$$

Consistency of $\hat{\theta}_n$ justifies the approximation of the standard error with $s_{\hat{\theta}} = \sigma(\hat{\theta})/\sqrt{n}$.

3.3.2 Method of Maximum Likelihood

Construction of Maximum Likelihood Estimators

Suppose that random variables $X_1, \ldots, X_n$ have a joint density or frequency function $f(x_1, \ldots, x_n|\theta)$. Given observed values $X_1 = x_1, \ldots, X_n = x_n$ the likelihood of θ as a function of $x_1, \ldots, x_n$ is defined as

$$L(\theta) = f(x_1, \ldots, x_n|\theta).$$

So L is a function of θ. If the distribution is discrete the likelihood function gives the probability of observing the given data as a function of θ. The **maximum likelihood estimate (MLE)** of θ is that value of θ that maximises the likelihood – that is, makes the observed data *most probable* or *most likely.*

The statistics is under i.i.d. assumptions

$$L(\theta) = \prod_{i=1} f(X_i|\theta)$$

and the **log likelihood** is

$$l(\theta) = \log(L(\theta)) = \sum_{i=1} \log(f(X_i|\theta)).$$

Poisson Distribution. $X \sim Po(\lambda)$ so

$$I\!P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

and (under i.i.d.)

$$\begin{aligned} l(\theta) &= \sum_{i=1} (X_i \log \lambda - \lambda - \log(X_i!)) \\ &= \log \lambda \sum_{i=1} X_i - n\lambda - \sum_{i=1} \log(X_i!). \end{aligned}$$

To find the MLE we have to solve

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1} X_i - n = 0$$

so again

$$\hat{\lambda} = \bar{X}.$$

Normal Distribution. Here

$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_{i=1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right).$$

So the likelihood function is

$$l(\mu, \sigma) = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1} (X_i - \mu)^2.$$

We have to solve

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1} (X_i - \mu) = 0 \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1} (X_i - \mu)^2 = 0. \end{aligned}$$

We obtain for the MLE

$$\hat{\mu} = \bar{X} \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1} (X_i - \bar{X})^2}.$$

Large Sample Theory for Maximum Likelihood Estimates

We consider i.i.d. samples, then the log likelihood is

$$l(\theta) = \sum_{i=1} f(X_i|\theta).$$

The true value of θ is θ_0. We assume throughout this section that the function f satisfies *technical smoothness conditions.*
We now develop approximations to the sampling distribution of MLEs by using limiting arguments as the sample size increases.

Theorem 3.3.1. *The MLE from an i.i.d. sample is consistent.*

Lemma 3.3.1. *Define $I(\theta)$ by*

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2.$$

$I(\theta)$ may also be expressed as

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right].$$

The large sample distribution of a maximum likelihood estimate is approximately normal with mean θ_0 and variance $1/nI(\theta_0)$. We say that the MLE is **asymptotically unbiased** and refer to the variance of the limiting normal distribution as the **asymptotic variance of the MLE**. Formally

Theorem 3.3.2. *The probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to the normal distribution.*

3.4 Construction of Maximum Likelihood Estimators

Suppose that random variables $X_1, \ldots, X_n$ have a joint density or frequency function $f(x_1, \ldots, x_n|\theta)$. Given observed values $X_1 = x_1, \ldots, X_n = x_n$ the likelihood of θ as a function of $x_1, \ldots, x_n$ is defined as

$$L(\theta) = f(x_1, \ldots, x_n|\theta).$$

So L is a function of θ. If the distribution is discrete the likelihood function gives the probability of observing the given data as a function of θ. The **maximum likelihood estimate (MLE)** of θ is that value of θ that maximises the likelihood – that is, makes the observed data *most probable* or *most likely.*
The statistics is under i.i.d. assumptions

$$L(\theta) = \prod_{i=1} f(X_i|\theta)$$

and the **log likelihood** is

$$l(\theta) = \log(L(\theta)) = \sum_{i=1} \log(f(X_i|\theta)).$$

Poisson Distribution. $X \sim Po(\lambda)$ so

$$I\!P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

and (under i.i.d.)

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} (X_i \log \lambda - \lambda - \log(X_i!)) \\ &= \log \lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!). \end{aligned}$$

To find the MLE we have to solve

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n = 0$$

so again

$$\hat{\lambda} = \bar{X}.$$

Normal Distribution. Here

$$f(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right).$$

So the likelihood function is

$$l(\mu, \sigma) = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

We have to solve

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0 \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (X_i - \mu)^2 = 0. \end{aligned}$$

We obtain for the MLE

$$\hat{\mu} = \bar{X} \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

3.5 Large Sample Theory for Maximum Likelihood Estimates

We consider i.i.d. samples, then the log likelihood is

$$l(\theta) = \sum_{i=1}^{n} f(X_i | \theta).$$

The true value of θ is θ_0. We assume throughout this section that the function f satisfies *technical smoothness conditions.*
We now develop approximations to the sampling distribution of MLEs by using limiting arguments as the sample size increases.

Theorem 3.5.1. *The MLE from an i.i.d. sample is consistent.*

Lemma 3.5.1. *Define $I(\theta)$ by*

$$I(\theta) = I\!E \left[\frac{\partial}{\partial\theta} \log f(X|\theta) \right]^2.$$

$I(\theta)$ may also be expressed as

$$I(\theta) = -I\!E \left[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta) \right].$$

The large sample distribution of a maximum likelihood estimate is approximately normal with mean θ_0 and variance $1/nI(\theta_0)$. We say that the MLE is **asymptotically unbiased** and refer to the variance of the limiting normal distribution as the **asymptotic variance of the MLE**. Formally

Theorem 3.5.2. *The probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to the normal distribution.*

3.6 Confidence Intervals for Maximum Likelihood Estimates

Normal Sample. Based on §2.5.2 we have

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}.$$

Let $t_{n-1}(\alpha/2)$ denote the point beyond which the t distribution with $n-1$ degrees of freedom has probability $\alpha/2$. Since the t distribution is symmetric about 0, the probability to the left of $-t_{n-1}(\alpha/2)$ is also $\alpha/2$. So

$$I\!P\left[-t_{n-1}(\alpha/2) \le \frac{\sqrt{n}(\bar{X} - \mu)}{S} \le t_{n-1}(\alpha/2) \right] = 1 - \alpha.$$

From this we find

$$I\!P\left[\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}} \right] = 1 - \alpha.$$

A confidence interval for σ^2 can be constructed using (§2.5.2)

$$\frac{n\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-1}.$$

So

$$I\!P\left[\chi^2_{n-1}(1-\alpha/2) \le \frac{n\hat{\sigma^2}}{\sigma^2} \le \chi^2_{n-1}(\alpha/2) \right] = 1 - \alpha.$$

Manipulation of the inequalities yield

$$\mathbb{P}\left[\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1-\alpha/2)}\right] = 1-\alpha.$$

General i.i.d. Case. We use the large sample theory outlines above: $\sqrt{nI(\hat{\theta})}(\hat{\theta}-\theta)$ is approximately normal distributed. So

$$\mathbb{P}\left[-z(\alpha/2) \leq \sqrt{nI(\hat{\theta})}(\hat{\theta}-\theta) \leq z(\alpha/2)\right] = 1-\alpha.$$

So

$$\hat{\theta} \pm z(\alpha/2)\sqrt{\frac{1}{nI(\hat{\theta})}}$$

is an approximate $100(1-\alpha)\%$ confidence interval.

3.7 Efficiency and the Cramer-Rao Lower Bound

In general the above methods do not lead to the same estimator, so the question arises how to evaluate estimators. Qualitatively, it would be sensible to choose that estimate whose sampling distribution was most highly concentrated about the true parameter value. Quantitatively we use the **Mean Squared Error** as a measure of concentration

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta}-\theta_0)^2 = Var(\hat{\theta}) - (\mathbb{E}(\hat{\theta})-\theta_0)^2.$$

If $\hat{\theta}$ is unbiased, them $MSE(\hat{\theta}) = Var(\hat{\theta})$.
Given two estimates, $\hat{\theta}$ and $\tilde{\theta}$, the **efficiency** of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be

$$\text{eff}(\hat{\theta},\tilde{\theta}) = \frac{Var(\tilde{\theta})}{Var(\hat{\theta})}.$$

In searching for an optimal estimate, we might ask whether there is a lower bound for the MSE of *any* estimate. An estimate achieving such a lower bound could not be improved on. The Cramer-Rao inequality provides such a lower bound for *unbiased* estimators.

Theorem 3.7.1 (Cramer-Rao Inequality)**.** *Let $X_1,\ldots,X_n$ be i.i.d. with density $f(x|\theta)$ and T an unbiased estimate of θ. Then*

$$Var(T) \geq \frac{1}{nI(\theta)}.$$

An unbiased estimate whose variance ac hieves the lower bound is said to be **efficient**. Since the asymptotic variance of a maximum likelihood estimate is equal to the lower bound, maximum likelihood estimates are said to be **asymptotically efficient**.

3.8 Sufficiency

The concept of sufficiency arises as an attempt to answer the following question: Is there a statistic $T(X_1, \ldots X_n)$, which contains all the information in the sample about θ? If so a reduction of the original data to this statistic without loss of information is possible.

Definition 3.8.1. *A statistic $T(X_1, \ldots, X_n)$ is said to be **sufficient** for θ if the conditional distribution of $X_1, \ldots, X_n$, given $T = t$, does not depend on θ for any t.*

So, given the value of a **sufficient statistic** T, we can gain no more knowledge about θ from knowing more about the probability distribution of $X_1, \ldots, X_n$. Sufficient statistics can be identified more easily by

Theorem 3.8.1 (Factorisation Theorem)**.** *A necessary and sufficient condition for $T(X_1, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint probability density (mass) function factors in the form*

$$f(x_1, \ldots, x_n|\theta) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n).$$

Recall that one-parameter exponential families have density functions of the form

$$f(x|\theta) = h(x)c(\theta) \exp\left(\omega(\theta)T(x)\right) \mathbf{1}_A(x).$$

So an i.i.d. sample from such a family has joint distribution function

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= \prod_{i=1} h(x_i)c(\theta) \exp\left(\omega(\theta)T(x_i)\right) \mathbf{1}_A(x_i) \\ &= \exp\left(\omega(\theta) \sum_{i=1} T(x_i) + n \log c(\theta)\right) \exp\left(\sum_{i=1} \log h(x_i)\right) \prod_{i=1} \mathbf{1}_A(x_i). \end{aligned}$$

So by the factorisation theorem $\sum_{i=1}^{n} T(x_i)$ is a sufficient statistic for θ.

Corollary 3.8.1. *If T is sufficient for θ, the maximum likelihood estimate is a function of T.*

If an estimator is not a function of a sufficient statistic it can be improved!

Theorem 3.8.2 (Rao-Blackwell Theorem)**.** *Let $\hat{\theta}$ be an estimator of θ with $\mathbb{E}(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and let $\tilde{\theta} = \mathbb{E}(\hat{\theta}|T)$. Then, for all θ,*

$$\mathbb{E}(\tilde{\theta} - \theta)^2 \leq \mathbb{E}(\hat{\theta} - \theta)^2.$$

The inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

3.9 Distributions Derived from the Normal Distribution

3.9.1 The χ^2, F, t Distributions

Definition 3.9.1. *Given $X_1, \ldots, X_n$ independent $N(0, 1)$ distributed random variables. The distribution of the sum*

$$Y = X_1^2 + \ldots + X_n^2$$

is called χ^2-distribution with n degrees of freedom, χ^2_n.

Theorem 3.9.1. *The χ^2_n distribution has the density*

$$g_n(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} \exp\{-y/2\}, \quad y > 0$$

and $g_n(y) = 0$ elsewhere.

- If $X \sim \chi^2_n$ then $\mathbb{E}[X] = n$, $Var[X] = 2n$.
- If $X \sim \chi^2_n$ and $Y \sim \chi^2_m$, then $X + Y \sim \chi^2_{n+m}$ (convolution property).

Definition 3.9.2. *Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$ be independent. Then the distribution of the quotient*

$$\frac{Y/m}{X/n}$$

is called a F-distribution with (m, n) degrees of freedom, $F_{m,n}$.

Theorem 3.9.2. *The density of the F-distribution is given by*

$$f_{m,n}(y) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} m^{m/2} n^{n/2} \frac{y^{m/2-1}}{(n+my)^{(m+n)/2}}, \quad y > 0$$

and $f_{m,n} = 0$ elsewhere.

Since $X \sim F_{m,n}$ implies $1/X \sim F_{n,m}$ we have for the quantile-function $F^{-1}_{n,m}(q) = 1/F^{-1}_{m,n}(1-q)$.

Definition 3.9.3. *Let X and Y be independent random variables with $X \sim N(0,1)$ and $Y \sim \chi^2_n$ distributed. The distribution of*

$$\frac{X}{\sqrt{Y/n}}$$

is called t distribution with n degrees of freedom, t_n.

Theorem 3.9.3. *The t_n distribution has density*

$$h_n(y) = \frac{\Gamma((m+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2}, \quad y > 0$$

and $h_n = 0$ elsewhere

We also need the non-central versions of the above distributions

Definition 3.9.4. *(i) Given $X_1, \ldots, X_n$ independent $N(\mu_i, 1), i = 1, \ldots n$ distributed random variables. The distribution of the sum*

$$Y = X_1^2 + \ldots + X_n^2$$

is called non-central χ^2-distribution with n degrees of freedom and non-centrality parameter $\lambda = \sum_{i=1}^n \mu_i^2$, $\chi^2_{n,\lambda}$.

(ii) Let $X \sim \chi^2_n$ and $Y \sim \chi^2_{m,\lambda}$ be independent. Then the distribution of the quotient

$$\frac{Y/m}{X/n}$$

is called a non-central F-distribution with (m,n) degrees of freedom and the non-centrality parameter λ, $F_{m,n,\lambda}$.

(iii) Let X and Y be independent random variables with $X \sim N(\lambda,1)$ and $Y \sim \chi^2_n$ distributed. The distribution of

$$\frac{X}{\sqrt{Y/n}}$$

is called non-central t distribution with n degrees of freedom and non-centrality parameter λ, $t_{n,\lambda}$.

Proposition 3.9.1. *(i) If $X \sim \chi^2_{n,\lambda}$ then $\mathbb{E}[X] = n+\lambda$ and $Var[X] = 2n+4\lambda$.*

(ii) If $T \sim t_{n,\lambda}$, then $\mathbb{E}[T] = \lambda\sqrt{n/2}\Gamma((n-1)/2)/\Gamma(n/2), n>1$ and $Var(T) = n(1+\lambda^2)/(n-2) - (\mathbb{E}[T])^2, n>2$.

(iii) If $F \sim F_{m,n,\lambda}$, then $\mathbb{E}[F] = \frac{n(m+\lambda)}{m(n-2)}, n>2$.

3.9.2 Sample Mean and Sample Variance

Given a set of i.i.d. $N(\mu,\sigma^2)$ distributed random variables $X_1,\dots X_n$ we compute the distribution of the sample mean

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

and of the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

and of $(\bar{X}_n, S_n^2)$.

Theorem 3.9.4. *(i) $\bar{X}_n$ and S_n^2 are independent;*

(ii) $(n-1)S_n^2/\sigma^2$ is χ^2_{n-1} distributed;

(iii) $\bar{X}_n$ is $N(\mu,\sigma^2/n)$ distributed;

(iv) $\sqrt{n}(\bar{X}_n - \mu)/S_n$ is t_{n-1} distributed.

Bibliography

[BB70] J.C. Burkill and H. Burkill. *A second course in mathematical analysis*. Cambridge University Press, Cambridge, 1970.

[Bil68] P. Billingsley. *Convergence of probability measures*. Wiley, New York, 1968.

[BN98] O.E. Barndorff-Nielsen. Processes of normal inverse Gaussian type. *Finance and Stochastics*, 2(1):41–68, 1998.

[Dot90] M. U. Dothan. *Prices in financial markets*. Oxford University Press, Oxford, 1990.

[Dud89] R.M. Dudley. *Real analysis and probability*. Wadsworth, Pacific Grove, 1989.

[Dur96] R. Durrett. *Probability: Theory and examples*. Duxbury Press at Wadsworth Publishing Company, 2nd edition, 1996.

[Dur99] R. Durrett. *Essentials of stochastic processes*. Springer Texts in Statistics. Springer, 1999.

[EK95] E. Eberlein and U. Keller. Hyperbolic distributions in finance. *Bernoulli*, 1:281–299, 1995.

[GS01] G. R. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, Oxford, 3rd edition, 2001. 1st ed. 1982, 2nd ed. 1992.

[GW86] G.R. Grimmett and D.J.A. Welsh. *Probability: An introduction*. Oxford University Press, Oxford, 1986.

[JP00] J. Jacod and P. Protter. *Probability essentials*. Springer, 2000.

[Kol33] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933. English translation: *Foundations of probability theory*, Chelsea, New York, (1965).

[Leb02] H. Lebesgue. Intégrale, longueur, aire. *Annali di Mat.*, 7(3):231–259, 1902.

[Nev75] J. Neveu. *Discrete-parameter martingales*. North-Holland, Amsterdam, 1975.

[Res01] S. Resnick. *A probability path*. Birkhäuser, 2001. 2nd printing.

[Ros97] S.M. Ross. *Probability models.* Academic Press, 6th edition, 1997.

[Ros00] J.S. Rosenthal. *A first look at rigorous probability theory.* World Scientific, Singapore, 2000.

[Ryd97] T.H. Rydberg. The normal inverse Gaussian lévy process: Simulation and approximation. Research Report, Department of Theoretical Statistics, Institute of Mathematics, University of Århus University, 1997.

[Ryd99] T.H. Rydberg. Generalized hyperbolic diffusions with applications towards finance. *Mathematical Finance*, 9:183–201, 1999.

[She96] N. Shephard. Statistical aspects of ARCH and stochastic volatility. In D.R. Cox, D.V. Hinkley, and O.E. Barndorff-Nielsen, editors, *Time Series Models - in econometrics, finance and other fields*, pages 1–67. Chapman & Hall, London, 1996.

[Wil91] D. Williams. *Probability with martingales.* Cambridge University Press, Cambridge, 1991.

[Wil01] D. Williams. *Weighting the odds.* Cambridge University Press, Cambridge, 2001.