Chapter I. PROBABILITY BACKGROUND.

§1. MEASURE

Measure-theoretic Probability Theory is based on Measure Theory. This originated with Henri LEBESGUE (1875-1941), in his 1902 thesis, ‘Intégrale, longueur, aire’. We begin with the simplest case.

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:

$$\mu \left( \bigcup_{r=1}^{n} I_r \right) = \sum_{r=1}^{n} \mu(I_r)$$

(finite additivity).

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

$$\mu \left( \bigcup_{r=1}^{\infty} I_r \right) = \sum_{r=1}^{\infty} \mu(I_r)$$

(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

$$\mu(I \setminus A) = \mu(I) - \mu(A)$$

(complementation).

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

$$A \subseteq B \& \mu(B) = 0 \Rightarrow \mu(A) = 0$$

(completeness).

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* $\mu$ (on the real line, $\mathbb{R}$).

**Area.**

The area of a rectangle $R = (a_1, b_1) \times (a_2, b_2)$ – with or without any of its perimeter included – should be $\mu(R) = (b_1 - a_1) \times (b_2 - a_2)$. The area of a finite or countably infinite union of disjoint rectangles should be the sum of their areas:

$$
\mu \left( \bigcup_{n=1}^{\infty} R_n \right) = \sum_{n=1}^{\infty} \mu(R_n) \quad \text{(countable additivity)}.
$$

For $R$ a rectangle and $A \subseteq R$ with area $\mu(A)$, its complement should have area

$$
\mu(R \setminus A) = \mu(R) - \mu(A) \quad \text{(complementation)}.
$$

If $B \subseteq A$ and $A$ has area 0, $B$ should have area 0:

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

Let $\mathcal{F}$ be the smallest class of sets, containing the rectangles, closed under finite or countably infinite unions, closed under complements, and complete (containing all subsets of sets of area 0 as sets of area 0). Lebesgue showed that area can be sensibly defined on the sets in $\mathcal{F}$ and no others. The sets $A \in \mathcal{F}$ are called the *Lebesgue-measurable sets* in the plane $\mathbb{R}^2$; area, defined on $\mathcal{F}$, is called *Lebesgue measure* in the plane. So: some but not all sets in the plane have an area.

**Volume.**

Similarly in 3-dimensional space $\mathbb{R}^3$, starting with the volume of a cuboid $C = (a_1, b_1) \times (a_2, b_2) \times (a_3, b_3)$ as

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

**Euclidean Space.**

Similarly in $k$-dimensional Euclidean space $\mathbb{R}^k$. We start with

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

and obtain the class $\mathcal{F}$ of *Lebesgue-measurable sets* in $\mathbb{R}^k$, and *Lebesgue measure* $\mu$ in $\mathbb{R}^k$. 2
Probability.

The unit cube \([0, 1]^k\) in \(\mathbb{R}^k\) has Lebesgue measure 1. It can be used to model the uniform distribution (density \(f(x) = 1\) if \(x \in [0, 1]^k\), 0 otherwise), with probability = length/area/volume if \(k = 1/2/3\).

Note. If a property holds everywhere except on a set of measure zero, we say it holds almost everywhere (a.e.) [French: presque partout, p.p.; German: fast überall, f.u.]. If it holds everywhere except on a set of probability zero, we say it holds almost surely (a.s.) [or, with probability one].

§2 INTEGRAL.

1. Indicators.

We start in dimension \(k = 1\) for simplicity, and consider the simplest calculus formula \(\int_a^b 1 \, dx = b - a\). We rewrite this as

\[ I(f) := \int_{-\infty}^{\infty} f(x) \, dx = b - a \text{ if } f(x) = I_{[a,b]}(x), \]

the indicator function of \([a, b]\) (1 in \([a, b]\), 0 outside it), and similarly for the other three choices about end-points.

2. Simple functions.

A function \(f\) is called simple if it is a finite linear combination of indicators: \(f = \sum_{i=1}^n c_i f_i\) for constants \(c_i\) and indicator functions \(f_i\) of intervals \(I_i\). One then extends the definition of the integral from indicator functions to simple functions by linearity:

\[ I \left( \sum_{i=1}^n c_i f_i \right) := \sum_{i=1}^n c_i I(f_i) \]

for constants \(c_i\) and indicators \(f_i\) of intervals \(I_i\).

3. Non-negative measurable functions.

Call \(f\) a (Lebesgue-) measurable function if, for all \(c\), the sets \(\{x : f(x) \leq c\}\) is a Lebesgue-measurable set (§1). If \(f\) is a non-negative measurable function, we quote that it is possible to construct \(f\) as the increasing limit of a sequence of simple functions \(f_n\):

\[ f_n(x) \uparrow f(x) \text{ for all } x \in \mathbb{R} \quad (n \to \infty), \quad f_n \text{ simple}. \]

We then define the integral of \(f\) as

\[ I(f) := \lim_{n \to \infty} I(f_n) \leq \infty \]

3
(we quote that this does indeed define $I(f)$: the value does not depend on which approximating sequence $(f_n)$ we use). Since $f_n$ increases in $n$, so does $I(f_n)$ (the integral is order-preserving), so either $I(f_n)$ increases to a finite limit, or diverges to $\infty$. In the first case, we say $f$ is \textit{(Lebesgue-) integrable} with \textit{(Lebesgue-) integral} $I(f) = \lim I(f_n)$, or $\int f(x) \, dx = \lim \int f_n(x) \, dx$, or simply $\int f = \lim \int f_n$.

\textbf{4. Measurable functions.}

If $f$ is a measurable function that may change sign, we split it into its positive and negative parts, $f_\pm$:

\[ f_+(x) := \max(f(x), 0), \quad f_-(x) := -\min(f(x), 0), \]

\[ f(x) = f_+(x) - f_-(x), \quad |f(x)| = f_+(x) + f_-(x) \]

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

\[ \int f := \int f_+ - \int f_- \]

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

\[ \int |f| = \int f_+ + \int f_- \]

\textit{Note.} The Lebesgue integral thus defined is, by construction, an \textit{absolute integral}: $f$ is integrable iff $|f|$ is integrable. Thus, for instance,

\[ \int_0^\infty \frac{\sin x}{x} \, dx = \frac{\pi}{2} \]

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

\[ \int_0^X \frac{\sin x}{x} \, dx \to \frac{\pi}{2} \quad \text{as} \quad X \to \infty. \]

The class of (Lebesgue-) integrable functions $f$ on $\mathbb{R}$ is written $L(\mathbb{R})$ or (for reasons explained below) $L_1(\mathbb{R})$ – abbreviated to $L_1$ or $L$.

\textbf{Higher dimensions.}
In $\mathbb{R}^k$, we start instead from the $k$-dimensional boxes in place of intervals. If $f$ is the indicator of a box $B = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_k, b_k],$
\[
\int f := \prod_{i=1}^{k} (b_i - a_i).
\]
We then extend to simple functions (linear combinations of indicators of boxes) by linearity, to non-negative measurable functions by taking increasing limits, and to measurable functions by taking positive and negative parts.

$L_p$ spaces.

For $p \geq 1$, the $L_p$ spaces $L_p(\mathbb{R}^k)$ on $\mathbb{R}^k$ are the spaces of measurable functions $f$ with $L_p$-norm
\[
\|f\|_p := \left( \int |f|^p \right)^{\frac{1}{p}} < \infty.
\]

Riemann integrals.

We first meet the ‘Sixth-Form integral’, taught non-rigorously at school. Mathematics undergraduates are taught a rigorous integral (in their 1st or 2nd years), the Riemann integral [G.B. RIEMANN (1826-1866)] – essentially just a rigourization of the school integral. It is much easier to set up than the Lebesgue integral, but much harder to handle.

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 then gives one the technically superior Lebesgue integral anyway.

Note. Integration is like summation (which is why Leibniz gave us the integral sign $\int$, as an elongated S). Lebesgue was a very practical man – his father was a tradesman – and used to think about integration in the following way. Think of a shopkeeper totalling up his day’s takings. The Riemann integral is like adding up the takings – notes and coins – in the order in which they arrived. By contrast, the Lebesgue integral is like totalling up the takings in order of size – from the smallest coins up to the largest notes. This is obviously better! In mathematical effect, it exchanges ‘integrating by $x$-values’ (abscissae) with ‘integrating by $y$-values (ordinates).
Lebesgue-Stieltjes integral.

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

\[
F(x) \leq F(y) \quad \text{if} \quad x \leq y
\]

(prime example: \( F \) a probability distribution function). Such functions can have at most countably many discontinuities, which are at worst jumps. We may without loss re-define \( F \) at jumps so as to be right-continuous.

We now generalize the starting points above:

(i) Measure. We take \( \mu((a,b]) := F(b) - F(a) \).

(ii) Integral. We take \( \int_a^b 1 := F(b) - F(a) \).

Now follow through the successive extension procedures above. We obtain:

(i) Lebesgue-Stieltjes measure \( \mu \), or \( \mu_F \).

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

Similarly in higher dimensions; we omit further details.

Finite variation (FV).

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 \( [a,b] \), \( V^b_a(F) \):

\[
V^b_a(F) := \sup_{\mathcal{P}} \sum_{i} |F(x_i) - F(x_{i-1})|.
\]

This may be +\( \infty \); but if \( V^b_a(F) < \infty \), \( F \) is said to be of finite variation on \( [a,b] \), \( F \in FV^b_a \). If \( F \) is of bounded variation on all finite intervals, \( F \) is said to be locally of finite variation, \( F \in FV^\text{loc} \); if \( F \) is of finite variation on the real line \( \mathbb{R} \), \( F \) is of finite variation, \( F \in FV \).

We quote that the following two properties are equivalent:

(i) \( F \) is locally of finite variation,

(ii) \( F \) is 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 finite variation.

§3 PROBABILITY. Probability spaces.

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 till 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 the great Russian mathematician and probabilist A.N.KOLMOGOROV (1903-1987), whose classic book of 1933, Grundbegriffe der Wahrscheinlichkeitsrechnung (Foundations of probability theory) [Kol] 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 ⊆ Ω of points ω we want to know the probability P(A) (or pr(A)). We need

1. P(∅) = 0, P(Ω) = 1,
2. P(A) ≥ 0 for all A,
3. If A_1, A_2, …, A_n are disjoint, P(∪_{i=1}^{n} A_i) = ∑_{i=1}^{n} P(A_i) (finite additivity), which, as above we will strengthen to 3*. If A_1, A_2 … (ad inf.) are disjoint,

\[ P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \]  (countable additivity).

4. If B ⊆ A and P(A) = 0, then P(B) = 0 (completeness).

Then by 1 and 3 (with A = A_1, Ω \ A = A_2),

\[ P(A^c) = P(\Omega \setminus A) = 1 - P(A). \]

So the class \( \mathcal{F} \) of subsets of Ω whose probabilities \( P(A) \) are defined should be closed under countable, disjoint unions and complements, and contain the empty set ∅ and the whole space Ω. Such a class is called a \( \sigma \)-field of subsets of Ω [or sometimes a \( \sigma \)-algebra, which one would write \( \mathcal{A} \)]. For each \( A \in \mathcal{F} \), \( P(A) \) should be defined (and satisfy 1, 2, 3*, 4 above). So, \( P : \mathcal{F} \to [0, 1] \) is a set-function,

\[ P : A \mapsto P(A) \in [0, 1] \quad (A \in \mathcal{F}). \]
The sets \( A \in \mathcal{F} \) are called events. Finally, 4 says that all subsets of null-sets (events) with probability zero (we will call the empty set \( \emptyset \) empty, not null) should be null-sets (completeness). A probability space, or Kolmogorov triple, is a triple \((\Omega, \mathcal{F}, P)\) satisfying these Kolmogorov axioms 1,2,3*,4 above. A probability space is a mathematical model of a random experiment.

Random variables.

Next, recall random variables \( X \) from your first probability course. Given a random outcome \( \omega \), you can calculate the value \( X(\omega) \) of \( X \) (a scalar – a real number, say; similarly for vector-valued random variables, or random vectors). So, \( X \) is a function from \( \Omega \) to \( \mathbb{R} \), \( X: \Omega \rightarrow \mathbb{R} \), \( X: \omega \rightarrow X(\omega) \) (\( \omega \in \Omega \)).

Recall also that the distribution function of \( X \) is defined by

\[ F(x), \quad \text{or} \quad F_X(x) := P\left(\{\omega : X(\omega) \leq x\}\right), \quad \text{or} \quad P(X \leq x), \quad (x \in \mathbb{R}). \]

We can only deal with functions \( X \) for which all these probabilities are defined. So, for each \( x \), we need \( \{\omega : X(\omega) \leq x\} \in \mathcal{F} \). We summarize this by saying that \( X \) is measurable with respect to the \( \sigma \)-field \( \mathcal{F} \) (of events), briefly, \( X \) is \( \mathcal{F} \)-measurable. Then, \( X \) is called a random variable [non-\( \mathcal{F} \)-measurable \( X \) cannot be handled, and so are left out]. So,

(i) a random variable \( X \) is an \( \mathcal{F} \)-measurable function on \( \Omega \),

(ii) a function on \( \Omega \) is a random variable (is measurable) iff its distribution function is defined.

Generated \( \sigma \)-fields.

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

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

When the (random) value \( X(\omega) \) is known, we know which of the events in the \( \sigma \)-field generated by \( X \) have happened: these are the events \( \{\omega : X(\omega) \in B\} \), where \( B \) runs through the Borel \( \sigma \)-field [generated by the intervals] on \( \mathbb{R} \).

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 (1910-2004), which we quote:

\[ \sigma(X) \subseteq \sigma(Y) \quad \text{iff} \quad X = g(Y) \]

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 \( Y = g^{-1}(X) \).

**Expectation.**

A measure (§1) determines an integral (§2). A probability measure \( P \), being a special kind of measure [a measure of total mass one] determines a special kind of integral, called an expectation.

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

\[
EX := \int_{\Omega} X \, dP, \quad \text{or} \quad \int_{\Omega} X(\omega) \, dP(\omega).
\]

If \( X \) is real-valued, say, with distribution function \( F \), recall that \( EX \) is defined in your first course on probability by

\[
EX := \int xf(x) \, dx \quad \text{if} \quad X \quad \text{has a density} \quad f
\]

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

\[
EX := \sum x_n f(x_n).
\]

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

\[
EX := \int_{-\infty}^{\infty} x \, dF(x)
\]

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 \( X \),

\[
\int_{\Omega} X \, dP = \int_{-\infty}^{\infty} x \, dF(x)
\]

follows by the change of variable formula for the measure-theoretic integral, on applying the map \( X : \Omega \to \mathbb{R} \) (we quote this).

**Glossary.**

We now have two parallel languages, measure-theoretic and probabilistic:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral</td>
<td>Expectation</td>
</tr>
<tr>
<td>Measurable set</td>
<td>Event</td>
</tr>
<tr>
<td>Measurable function</td>
<td>Random variable</td>
</tr>
<tr>
<td>almost-everywhere (a.e.)</td>
<td>almost-surely (a.s.)</td>
</tr>
</tbody>
</table>
§4. Equivalent measures and Radon-Nikodym derivatives.

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

$$P \ll Q,$$

if $P(A) = 0$ whenever $Q(A) = 0$, $A \in \mathcal{F}$. We quote from measure theory the vitally important Radon-Nikodym theorem: $P \ll Q$ iff there exists a ($\mathcal{F}$-)measurable function $f$ such that

$$P(A) = \int_A f \, dQ \quad \forall A \in \mathcal{F}$$

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

$$\int_A dP = \int_A \frac{dP}{dQ} \, dQ \quad \forall A \in \mathcal{F} : \quad \text{if } P \ll Q, \quad dP = \frac{dP}{dQ} \, dQ.$$

The measurable function [random variable] $dP/dQ$ is called the Radon-Nikodym derivative [RN-derivative] of $P$ with respect to $Q$.

If $P \ll Q$ and also $Q \ll P$, we call $P$ and $Q$ equivalent measures, written $P \sim Q$. Then $dP/dQ$ and $dQ/dP$ both exist, and $dP/dQ = 1/(dQ/dP)$.

For $P \sim Q$, $P(A) = 0$ iff $Q(A) = 0$: $P$ and $Q$ have the same null sets.

Taking negations: $P \sim Q$ iff $P, Q$ have the same sets of positive measure.

Taking complements: $P \sim Q$ iff $P, Q$ have the same sets of probability one [the same a.s. sets]. Thus the following are equivalent: $P \sim Q$ iff $P, Q$ have the same null sets/the same a.s. sets/the same sets of positive measure.

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

(a) It is the key to conditioning, which is of central importance.

(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, the Fundamental Theorem of Asset Pricing, etc. The key to all this is that prices should be the discounted expected values under the equivalent martingale measure. Thus equivalent measures, and the operation of change of measure, are of central economic and financial importance. Mathematically, change of measure is effected by Girsanov’s theorem.