Conceptual foundations of probability theory

Danny Nygård Hansen

3rd April 2023

1 Introduction

Theorem 1.1: The strong law of large numbers

Let $(\mathsf{X}_n)_{n\in\mathbb{N}}$ *be a sequence of i.i.d. integrable random variables on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathsf{X}_i(\omega) = \mathbb{E}[\mathsf{X}_1]$$

for $\mathbb{P}$*-almost all* $\omega \in \Omega$*.*

Proof. Bauer (1995, Theorem 12.1) or Billingsley (1995, Theorem 22.1). □

In the usual measure-theoretical formulation of probability theory, the following result is a corollary of the Law of Large Numbers:

Theorem 1.2: The frequency interpretation of probability

Let $\mathsf{X}, \mathsf{X}_1, \mathsf{X}_2, \ldots$ *be i.i.d. real-valued random variables on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. For every* $B \in \mathcal{B}(\mathbb{R})$ *we have*

$$\mathbb{P}(\mathsf{X} \in B) = \lim_{n\to\infty} \frac{|\{j \in \{1,\ldots,n\} \mid \mathsf{X}_j(\omega) \in B\}|}{n}$$

for $\mathbb{P}$*-almost all* $\omega \in \Omega$*.*

Proof. This follows directly by applying Theorem 1.1 to the sequence $(\mathbf{1}_B(\mathsf{X}_n))_{n\in\mathbb{N}}$. □

That is, given a sequence $(\mathsf{X}_n)_{n\in\mathbb{N}}$, and one extra X, of i.i.d. random variables, the probability that X lies in some Borel set B can be thought of as the proportion of the X_n that lie in B, as n tends to infinity. In other words, probability is a measure of the *frequency* with which an outcome of a random experiment obtains, if we repeat the experiment many times.

Whether or not this is the correct interpretation of probability as it occurs in the natural world we will not discuss here. Nonetheless the above result is an uncontroversial consequence of the theory, and it certainly aligns with our intuitive understanding of probability.

In this note we turn this result on its head and attempt to use it to motivate the formalisation of probability theory in terms of measure spaces. As we shall see, this is not entirely successful and will require some leaps that are not entirely justified by our conceptual grasp of probability.

2 Preliminaries

2.1 • Boolean algebras

We begin by reviewing some of the purely algebraic properties of Boolean algebras.

Definition 2.1: Boolean algebras

A ***Boolean algebra*** is a structure $\langle B; \vee, \wedge, ', 0, 1\rangle$ such that

(i) $\langle B; \vee, \wedge\rangle$ is a distributive lattice,

(ii) 0 and 1 are elements of B such that $x \vee 0 = x$ and $x \wedge 1 = x$ for all $x \in B$, and

(iii) $'$ is a unary operation such that $x \vee x' = 1$ and $x \wedge x' = 0$ for all $x \in B$.

The binary operations $\vee$ and $\wedge$ are called ***join*** and ***meet***, respectively. For $x \in B$ the element x' is called the ***complement*** of x. In a general bounded lattice L, an element $y \in L$ such that $x \vee y = 1$ and $x \wedge y = 0$ is called a complement of $x \in L$. If L is distributive, complements are unique. Recall also that the lattice structure on B induces a partial order $\leq$ such that $x \leq y$ if and only if $x \vee y = y$ for $x, y \in B$.

If $x \wedge y = 0$, then x and y are said to be ***disjoint***. A collection $\{x_i\}_{i\in I}$ of elements in B is called ***pairwise disjoint*** if x_i and x_j are disjoint for any choice of indices $i \neq j$.

Let B be a Boolean algebra. For $x, y \in B$ we define the ***symmetric difference*** between x and y by

$$x \triangle y = (x \wedge y') \vee (y \wedge x').$$

If $x \triangle y = 0$, then it is easy to show that $x = y$.

Before proceeding we note the following technical result that we shall need later:

Lemma 2.2

Let $\langle B; \vee, \wedge, ', 0, 1\rangle$ be a Boolean algebra. Let $\{x_i\}_{i\in I}$ be a collection of pairwise disjoint elements in B. If $\bigvee_{i\in I} x_i \in B$, then $\bigvee_{i\in J} x_i \in B$ for any cofinite[1] *$J \subseteq I$.*

[1] Recall that a subset J of a set I is called *cofinite* if the complement $I \setminus J$ is finite.

Proof. Let $(x_i)_{i\in I}$ be such a collection of elements, and let $J \subseteq I$ be cofinite. It suffices to prove the lemma in the case $I \setminus J = \{i_0\}$, since the general case then follows by induction. We claim that

$$\bigvee_{i\in J} x_i = x'_{i_0} \wedge \bigvee_{i\in I} x_i, \tag{2.1}$$

in which case the claim would follow. Let $j \in J$ and notice that, since $x_{i_0} \wedge x_j = 0$,

$$x'_{i_0} \wedge x_j = (x'_{i_0} \wedge x_j) \vee (x_{i_0} \wedge x_j) = (x'_{i_0} \vee x_{i_0}) \wedge x_j = 1 \wedge x_j = x_j,$$

i.e. $x_j \leq x'_{i_0}$. Now because also $x_j \leq \bigvee_{i\in I} x_i$ we get

$$x_j \leq x'_{i_0} \wedge \bigvee_{i\in I} x_i.$$

Since $j \in J$ was arbitrary, the inequality $\leq$ in (2.1) follows. Conversely, suppose that $s \in B$ is such that $x_j \leq s$ for all $j \in J$. Then $x_i \leq x_{i_0} \vee s$ for all $i \in I$, so

$$\bigvee_{i\in I} x_i \leq x_{i_0} \vee s.$$

It follows that

$$x'_{i_0} \wedge \bigvee_{i\in I} x_i \leq x'_{i_0} \wedge (x_{i_0} \vee s) = 0 \vee (x_{i_0} \wedge s) \leq s,$$

which implies the inequality $\geq$ in (2.1). □

2.2 • Abstract measure spaces

Definition 2.3: Generalised abstract measure spaces

A ***measure*** on a Boolean algebra B is a map $\mu\colon B \to [0,\infty)$ such that

$$\mu(x \vee y) = \mu(x) + \mu(y) \tag{2.2}$$

for all disjoint $x, y \in B$. If μ is a measure on a Boolean algebra B, then we call the pair (B,μ) a ***generalised abstract measure space***. If $x \neq 0$ implies that $\mu(x) > 0$, then μ is called ***positive definite***. If $\mu(1) = 1$, then we call μ a ***probability measure***.

It is clear that $\mu(\emptyset) = 0$ and that μ is increasing. It follows that $\mu(x) \leq \mu(1)$ for all $x \in B$. Notice that we require that μ is finite, but this is no restriction since we are ultimately interested in the case where μ is a probability measure.

The property (2.2) is called ***(finite) additivity*** of μ, since an easy induction argument extends it to all finite joins. We will later define more restrictive structures and measures upon them, hence the adjective 'generalised'.

Proposition 2.4: Boole's inequality

Let (B, μ) be a generalised abstract measure space. Then for any $x, y \in B$ we have

$$\mu(x \vee y) \leq \mu(x) + \mu(y).$$

Similar to (2.2), Boole's inequality may be extended to all finite joins by induction.

Proof. Notice that

$$(x \wedge y') \wedge y = x \wedge (y' \wedge y) = 0,$$

so $x \wedge y'$and y are disjoint, and that

$$(x \wedge y') \vee y = (x \vee y) \wedge (y' \vee y) = x \vee y.$$

It follows by additivity of μ that

$$\mu(x \vee y) = \mu\Big((x \wedge y') \vee y\Big) = \mu(x \wedge y') + \mu(y) \leq \mu(x) + \mu(y),$$

as desired. □

Definition 2.5: Metric Boolean algebras

A ***pseudometric Boolean algebra*** is a tuple (B, ρ), where B is a Boolean algebra and ρ is a pseudometric on B such that the maps $x \mapsto x'$, $(x, y) \mapsto x \vee y$, and $(x, y) \mapsto x \wedge y$ are continuous.

If ρ is a metric, then (B, ρ) is called a ***metric Boolean algebra***.

Next we equip generalised abstract measure spaces with a canonical pseudometric. If (B, μ) is a generalised abstract measure space, define a map $\rho_\mu \colon B \times B \to [0, \infty)$ by

$$\rho_\mu(x, y) = \mu(x \triangle y) \tag{2.3}$$

for $x, y \in B$. The next proposition shows that ρ_μ is in fact a pseudometric. We will always equip a generalised abstract measure space with this pseudometric.

Proposition 2.6

Given a generalised abstract measure space (B, μ), the map ρ_μ defined in (2.3) makes (B, ρ_μ) into a pseudometric Boolean algebra. Furthermore, ρ_μ is a metric if and only if μ is positive definite.

Proof. *ρ_μ is a pseudometric:* We only need to prove the triangle inequality. To this end, let $x, y, z \in B$ and notice that

$$\begin{aligned} x \wedge z' &= (x \wedge z) \wedge (y' \vee y) \\ &= (x \wedge z' \wedge y') \vee (x \wedge z' \wedge y) \\ &\leq (x \wedge y') \vee (y \wedge z'). \end{aligned}$$

Similarly we have $z \wedge x' \leq (z \wedge y') \vee (y \wedge x')$. It follows that

$$\begin{aligned} x \triangle z &= (x \wedge z') \vee (z \wedge x') \\ &\leq (x \wedge y') \vee (y \wedge x') \vee (y \wedge z') \vee (z \wedge y') \\ &= (x \triangle y) \vee (y \triangle z). \end{aligned}$$

Now Boole's inequality implies that

$$\rho_\mu(x, z) = \mu(x \triangle y) \leq \mu(x \triangle y) + \mu(y \triangle z) = \rho_\mu(x, y) + \rho_\mu(y, z),$$

as desired.

Continuity of lattice operations: Let $x \in B$, and let $(x_n)_{n \in \mathbb{N}}$ be a sequence in B that converges to x. Notice that $x_n' \triangle x' = x_n \triangle x$, so $\rho_\mu(x_n', x') = \rho_\mu(x_n, x)$. Hence the complementation map $x \mapsto x'$ is continuous.

Let further $(y_n)_{n \in \mathbb{N}}$ be a sequence converging to a point $y \in B$. A short calculation shows that

$$\begin{aligned} (x_n \vee y_n) \triangle (x \vee y) &= (x_n \wedge x' \wedge y') \vee (y_n \wedge x' \wedge y') \vee (x \wedge x_n' \wedge y_n') \vee (y \wedge x_n' \wedge y_n') \\ &\leq (x_n \wedge x') \vee (x \wedge x_n') \vee (y_n \wedge y') \vee (y \wedge y_n') \\ &= (x_n \triangle x) \vee (y_n \triangle y). \end{aligned}$$

Thus Boole's inequality shows that

$$\rho_\mu(x_n \vee y_n, x \vee y) \leq \rho_\mu(x_n, x) + \rho_\mu(y_n, y), \tag{2.4}$$

which implies continuity of the join map $(x, y) \mapsto x \vee y$.

Finally, continuity of the meet map $(x, y) \mapsto x \wedge y$ follows since

$$x \wedge y = \left(x' \vee y'\right)',$$

so it is a composition of continuous functions.

Positive definiteness: The last claim follows directly from the fact that $x \triangle y = 0$ if and only if $x = y$, for all $x, y \in B$. □

Remark 2.7. Notice that the measure μ can be written in terms of ρ_μ, since $\mu(x) = \rho_\mu(x, 0)$. Furthermore, since (pseudo)metrics are continuous, it follows that $\mu(x_n) \to \mu(x)$ whenever $x_n \to x$ in B. ⌟

It is well-known that any (pseudo)metric space has a completion, i.e. can be isometrically embedded as a dense subset of a complete (pseudo)metric space. See for instance Corollary 24.5 in Willard (1970). A natural question is then: If (B, ρ) is a (pseudo)metric Boolean algebra with metric completion $(\overline{B}, \overline{\rho})$, does $\overline{B}$ also carry the structure of a Boolean algebra?

This is indeed the case, and we sketch the construction: Let $x, y \in \overline{B}$, and let (x_n) and (y_n) be sequences in B that converge to x and y, respectively. Then these are Cauchy sequences in B, and the calculation leading to (2.4) shows that $(x_n \vee y_n)$ is also a Cauchy sequence. Thus it converges to some element of $\overline{B}$. Denote it $x \vee y$. We define x' and $x \wedge y$ similarly. It is easy to check that these operations satisfy the conditions in Definition 2.1. Furthermore, the completion $\overline{\rho}$ of the pseudometric ρ makes $(\overline{B}, \overline{\rho})$ into a pseudometric Boolean algebra.

This takes care of the metric structure. The next proposition shows that we can also extend the measure on a generalised abstract measure space to its completion.

Proposition 2.8

Let (B, μ) be a generalised abstract measure space, and let $(\overline{B}, \overline{\rho}_\mu)$ be the completion of (B, ρ_μ). Define a map $\overline{\mu}\colon \overline{B} \to [0, \infty)$ by

$$\overline{\mu}(x) = \lim_{n \to \infty} \mu(x_n), \tag{2.5}$$

*where $(x_n)_{n \in \mathbb{N}}$ is any sequence in B that converges to x. Then $\overline{\mu}$ is a well-defined measure on $\overline{B}$. The generalised abstract measure space $(\overline{B}, \overline{\mu})$ is called the **completion** of (B, μ).*

Proof. First notice that for any $x \in \overline{B}$ there does in fact exist a sequence in B converging to x. If $(x_n)_{n \in \mathbb{N}}$ is such a sequence, it is a Cauchy sequence in B, and the reverse triangle inequality shows that $(\mu(x_n))$ is a Cauchy sequence in $\mathbb{R}$, hence convergent. Thus the limit on the right-hand side of (2.5) exists.

Now let (y_n) be another sequence in B that approximates x. Another application of the reverse triangle inequality then shows that

$$|\mu(x_n) - \mu(y_n)| \leq \rho_\mu(x_n, y_n) \leq \rho_\mu(x_n, x) + \rho_\mu(y_n, x) \to 0.$$

Hence $\mu(x_n)$ and $\mu(y_n)$ converge to the same value, and thus $\overline{\mu}$ is well-defined.

Next we show that $\overline{\mu}$ is finitely additive. Let $x, y \in \overline{B}$ with $x \wedge y = 0$ and choose approximating sequences (x_n) and (y_n) in B. Then

$$(x_n \vee y_n) \wedge (x_n \wedge y_n)' = \Big(x_n \wedge (x_n \wedge y_n)'\Big) \vee \Big(y_n \wedge (x_n \wedge y_n)'\Big)$$

is the join of disjoint elements of B, so

$$\mu\Big((x_n \vee y_n) \wedge (x_n \wedge y_n)'\Big) = \mu\Big(x_n \wedge (x_n \wedge y_n)'\Big) + \mu\Big(y_n \wedge (x_n \wedge y_n)'\Big).$$

By continuity of the lattice operations we have $(x_n \wedge y_n)' \to 1$, so the three elements given as arguments to μ above are elements in approximating sequences for $x \vee y$, x and y respectively. By definition of $\overline{\mu}$ it follows that

$$\overline{\mu}(x \vee y) = \overline{\mu}(x) + \overline{\mu}(y)$$

as desired. □

Lemma 2.9

Let (B, μ) be a generalised abstract measure space. Every monotonic sequence in B is a Cauchy sequence.

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be an increasing sequence in B. Then since $x_n \leq 1$ we also have $\mu(x_n) \leq \mu(1) < \infty$ for all $n \in \mathbb{N}$. Thus the sequence $(\mu(x_n))$ is a bounded increasing sequence in $\mathbb{R}$, hence it converges to some $\alpha \geq 0$. For $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that $\mu(x_n) \in (\alpha - \varepsilon, \alpha]$ for all $n \geq N$.

If then $m, n \geq N$ with $m \leq n$, then $x_m \leq x_n$ and so

$$x_m \wedge x_n' \leq x_n \wedge x_n' = 0.$$

Hence $x_m \triangle x_n = x_n \wedge x_m'$, so it follows that

$$\rho_\mu(x_m, x_n) = \mu(x_n \wedge x_m') = \mu(x_n) - \mu(x_m) < \varepsilon.$$

Thus (x_n) is indeed a Cauchy sequence. The case where (x_n) is decreasing is similar. □

Proposition 2.10

Let (B, μ) be a generalised abstract measure space with completion $(\overline{B}, \overline{\mu})$. Then any sequence $(x_n)_{n\in\mathbb{N}}$ in B has a join s in $\overline{B}$, and

$$\bigvee_{i\leq n} x_i \to s$$

as $n \to \infty$. Similarly for meets.

Proof. The sequence $(\bigvee_{i\leq n} x_i)_{n\in\mathbb{N}}$ is increasing, so by Lemma 2.9 it has a limit $s \in \overline{B}$. We show that s is the join of (x_n). For $k \in \mathbb{N}$ and $n \geq k$ we have $x_k \leq \bigvee_{i\leq n} x_i$, i.e.

$$x_k \vee \bigvee_{i\leq n} x_i = \bigvee_{i\leq n} x_i.$$

Taking the limit as $n \to \infty$, continuity of (binary) joins implies that $x_k \vee s = s$ in $\overline{B}$, or $x_k \leq s$. Thus s is an upper bound of the sequence (x_n).

On the other hand, if $t \in \overline{B}$ is an upper bound of (x_n), then $x_k \leq t$. We have just seen that taking limits preserves inequalities, so this implies that $s \leq t$, and hence s is the least upper bound.

The corresponding result for meets follows similarly, or from the fact that complementation is continuous. □

2.3 • Abstract σ-algebras and continuity

Definition 2.11: Axiom of continuity

A generalised abstract measure space (B, μ) is said to satisfy the ***axiom of continuity*** if it has the following property: If $(x_n)_{n\in\mathbb{N}}$ is a decreasing sequence of elements in B such that $\bigwedge_{n\in\mathbb{N}} x_n$ exists and equals 0, then $\lim_{n\to\infty} \mu(x_n) = 0$.

This is an analogue of the continuity (from above) of ordinary countably additive measures on concrete σ-algebras.

Lemma 2.12

Let (B, μ) be a generalised abstract measure space with μ positive definite. Then (B, μ) satisfies the axiom of continuity.

As far as I know, the assumption that μ be positive definite is necessary for this to hold in general.

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a decreasing sequence in B such that $\bigwedge_{n\in\mathbb{N}} x_n = 0$. A similar argument to the one in the proof of Proposition 2.10 shows that x_n converges to 0 as $n \to \infty$ in B, hence also in the completion $\overline{B}$. Furthermore, since the sequence is decreasing we have $x_n = \bigwedge_{i\leq n} x_i$, so Proposition 2.10 also implies that the sequence converges to its meet s in $\overline{B}$. But since μ is positive definite, limits in $\overline{B}$ are unique, so $s = 0$. By Remark 2.7, $\mu(x_n)$ converges to $\mu(0) = 0$, proving the claim. □

Proposition 2.13: The generalised addition theorem

Let (B, μ) be a generalised abstract measure space satisfying the axiom of continuity. If $(x_n)_{n\in\mathbb{N}}$ is a sequence of pairwise disjoint elements in B such that $x = \bigvee_{n\in\mathbb{N}} x_n \in B$, then

$$\mu(x) = \sum_{n=1}^{\infty} \mu(x_n).$$

Proof. By Lemma 2.2, $r_n = \bigvee_{i>n} x_i$ exists in B, and we claim that $\bigwedge_{n\in\mathbb{N}} r_n = 0$. Let $t \in B$ be a lower bound of r_n for $n \in \mathbb{N}$. Then for $n \in \mathbb{N}$ we have

$$t \vee \bigvee_{i>n} x_i = \bigvee_{i>n} x_i,$$

and taking the meet of each side with x_n' yields $t \wedge x_n' = 0$. Hence $\bigvee_{n\in\mathbb{N}} t \wedge x_n' = 0$. Now notice that $\bigwedge_{n\in\mathbb{N}} x_n = 0$, since $x_n \wedge x_m = 0$ when $n \neq m$. It follows by taking complements that $\bigvee_{n\in\mathbb{N}} x_n' = 1$, and so

$$t = t \wedge \bigvee_{n\in\mathbb{N}} x_n' = \bigvee_{n\in\mathbb{N}} t \wedge x_n' = 0.$$

Thus $\bigwedge_{n\in\mathbb{N}} r_n = 0$ as claimed. It now follows from finite additivity of μ and the axiom of continuity that

$$\mu(x) = \sum_{i=1}^{n} \mu(x_i) + \mu(r_n) \to \sum_{i=1}^{\infty} \mu(x_i)$$

as $n \to \infty$ as desired. □

Proposition 2.14: Boole's inequality

Let (B, μ) be a generalised abstract measure space, and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in B. If $x = \bigvee_{n\in\mathbb{N}} x_n \in B$, then

$$\mu(x) \leq \sum_{n=1}^{\infty} \mu(x_n).$$

Proof. First notice that, by the finite Boole equality,

$$\mu\Big(\bigvee_{i\leq n} x_i\Big) \leq \sum_{i=1}^{n} \mu(x_i) \leq \sum_{i=1}^{\infty} \mu(x_i)$$

for all $n \in \mathbb{N}$. By Proposition 2.10, $\bigvee_{i\leq n} x_i \to x$ as $n \to \infty$, so $\mu(\bigvee_{i\leq n} x_i) \to \mu(x)$. The claim follows □

Of course, the join of a sequence of elements may not exist. In the case where we are ensured the existence of countable joins we use the following terminology:

Definition 2.15: Abstract σ-algebra

An ***abstract σ-algebra*** is a Boolean algebra B with countable joins. That is, if $(x_n)_{n\in\mathbb{N}}$ is a sequence of elements in B, then their join $\bigvee_{n\in\mathbb{N}} x_n$ exists.

If the join $\bigvee_{n\in\mathbb{N}} x_n$ exists, then it follows by taking complements that the meet $\bigwedge_{n\in\mathbb{N}} x_n'$ also exists. Hence an abstract σ-algebra also has countable meets. In the context of abstract measure spaces we obtain the following:

Definition 2.16: Abstract measure spaces

An ***abstract measure space*** is a generalised abstract measure space (B, μ) that satisfies the axiom of continuity, and where B is an abstract σ-algebra.

Lemma 2.17

If (B, μ) is a generalised abstract measure space with μ positive definite, then the completion $(\overline{B}, \overline{\mu})$ is an abstract measure space.

Proof. Since the completion of a metric space is a metric space, $\overline{\mu}$ is positive definite, so Lemma 2.12 implies that $(\overline{B}, \overline{\mu})$ satisfies the axiom of continuity.

On the other hand, Proposition 2.10 implies that every sequence in $\overline{B}$ has a join, so $\overline{B}$ is an abstract σ-algebra. □

3 The algebra of probability spaces

3.1 • Motivation

If the probability of an event is supposed to be a measure of how often this event occurs on repetitions of the random experiment in question, then it seems reasonable to assume that we are, at least in principle, able to distinguish when the event does and does not obtain. For example, after rolling a six-sided die the state of affairs 'the result of the die roll is three' is an event, since we can determine the outcome of the roll just by looking at the die. To take another example, after throwing a ball the state of affairs 'the ball was thrown further than 50 metres' is also an event: That is, we can determine whether or not the length of the throw was strictly greater than 50 metres.

One might take a different view: Say that one grants that it is possible to *affirm* that the length of the throw, measured in metres, lies in the interval $(50, \infty)$. If the length L in metres does in fact lie in the above interval, we can simply use a ruler whose subdivisions are smaller than $L - 50$ in metres. Still one might disagree that it is possible to *refute* that $L \in (50, \infty)$. For if L is exactly 50 metres, then since any measurement of L carries some error, it is in practice impossible to determine whether L is 50 (or slightly smaller), or whether it is slightly larger than 50. We will not pursue this line further but refer the reader to Vickers (1989) for more on this ***logic of affirmative assertions***.

To be precise, after performing the relevant random experiment, we will assume that we are always able to decide whether or not the event has occurred or not. In particular, if E is an event, then the state of affairs 'E does not obtain' is also an event, denoted $\overline{E}$: If E obtains, then $\overline{E}$ does not. And conversely, if E does not obtain, then $\overline{E}$ does obtain. We call $\overline{E}$ the ***complement of*** or the ***complementary event to E***.[1] Evidently, the complement of a complement is just the event we started with.

Next consider two events E_1 and E_2. Since we are able to decide whether each of them have obtained, the same is true for the event 'both E_1 and E_2 have obtained' and the event 'at least one of E_1 and E_2 has obtained'. The first is called the ***conjunction*** of E_1 and E_2 and is denoted $E_1 \wedge E_2$, and the second is the ***disjunction*** $E_1 \vee E_2$ of E_1 and E_2.

[1] In contrast, in the logic of affirmative assertions we do not allow complementation (i.e. negation). Hence it may not be surprising that this logic ends up being closely tied to topology, the relevant analogy being that the complement of an open set need not be open.

Finally it seems natural to allow an 'impossible event' 0 which never occurs, as well as a 'sure event' 1 that always occurs. Clearly 0 and 1 are each other's complements. If E_1 and E_2 are events with $E_1 \wedge E_2 = 0$, then this manifestly means that E_1 and E_2 cannot obtain simultaneously: The two events are ***incompatible***.

We collect all the relevant events in a set $\mathcal{F}$ and postulate that the structure $\langle \mathcal{F}; \vee, \wedge, \bar{\cdot}, 0, 1 \rangle$ is a Boolean algebra. We leave it to the reader to reflect on the reasonability of this assumption. This leads naturally to the following definition:

Definition 3.1: Generalised abstract probability spaces

A ***generalised abstract probability space*** is a generalised abstract measure space $(\mathcal{F}, P)$, where P is a positive definite probability measure.

The partial order $\leq$ induced by the lattice structure then has the interpretation that if $E \leq F$, then E implies F. For recall that this means that $E \vee F = F$, i.e. if F has already occurred then no information is gained by observing that E has also occurred. Conversely, if F has *not* occurred, then E is impossible.

We have justified every part of this definition except for the positive definiteness of P. In the usual measure theoretical formulation of probability theory, there may (and often do) exist events that are not empty but still have probability zero. Kolmogorov had the following to say in critique of this approach:

> [W]e are forced to give up the principle, formulated in numerous classical works in probability theory, according to which an event of probability zero is absolutely impossible. More precisely, one must allow that an event of positive probability can be decomposed into a (possible continuous) infinity of variants of which each has probability zero. (Kolmogorov and Jeffrey 1995)

Furthermore, this also has the technical benefit that the pseudometric ρ_P induced by P is in fact a metric. We will return to the consequences of this choice in the next section.

3.2 • Continuity of probability measures

Of course, the map P above is supposed to be analogous to a probability measure on a (concrete) σ-algebra. But ordinary measures are *countably* additive, not just finitely so. It is however difficult to justify extending the finite additivity to sequences of disjoint events purely on conceptual or operational groups. In fact, according to Kolmogorov himself:

> Since the new axiom [the axiom of continuity] is essential for infinite fields of probability only, it is almost impossible to elucidate its empirical meaning. (...) For, in describing any observable random process we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes. *We limit ourselves, arbitrarily, to only those models which satisfy Axiom VI* [the axiom of continuity]. This limitation has been found expedient in researches of the most diverse sort. (Kolmogorov 1956)

And furthermore:

> [S]omewhat more complicated problems require, if the theory is to be simple and tractable, that probability be subject to the *axiom of denumerable additivity*. However, the justification of that axiom remains purely empirical, in that we have not yet encountered any interesting problem for which we have not been able to construct a probability field conforming to the axiom in question. (Kolmogorov and Jeffrey 1995)

The axiom of continuity is, as we saw in Proposition 2.13, equivalent to countable additivity in our formalism. While countable additivity is usually preferred in the definition of (probability) measures today, Kolmogorov (1956) instead assumed the axiom of continuity and then proved the generalised addition theorem, as we have done above.

Kolmogorov could apparently only find justification for assuming the axiom in the success of the theory, and not in its conceptual underpinnings. Luckily, in the present context we can avoid taking a stance: We have already heard Kolmogorov argue that the absense of non-empty events of probability zero is conceptually problematic; hence we have disallowed them by assuming our probability measures to be positive definite. And according to Lemma 2.12, in this setting we get the axiom of continuity for free.

We are still not assured that our probability spaces are closed under countable joins. However, Lemma 2.17 tells us that taking the completion of a generalised abstract probability space solves the problem. And according to Kolmogorov, this can be done without issue:

> The use of denumerable additivity, with its great concomitant freedom, is not generally possible except in *complete* metric Boolean algebras. It is therefore natural, in probability theory, always to assume that the algebra of events is *complete* by adjunction of ideal elements. As has been pointed out, this situation is always realizable.

> So, the use of denumerably additive probabilities is seen to be legitimate and to impose no additional restrictions on the nature of the problems which fall under the scope of the general theory. (Kolmogorov and Jeffrey 1995)

Notice that he calls the adjoined elements 'ideal'. Thus he seems to think of these new events as not (necessarily) corresponding to real, observable events as described above, or at least not events that we can gain any information about. Instead, he seems only to be concerned with how their adjunction can improve the flexibility of the theory without restricting which classes of problems the theory can address. Indeed, he had the following to say:

> Even if the sets (events) A of $\mathfrak{F}$ can be interpreted as actual and (perhaps only approximately) observable events, it does not, of course, follow from this that the sets of the extended field $B\mathfrak{F}$ [the σ-algebra generated by $\mathfrak{F}$] reasonably admit of such an interpretation.
>
> Thus there is the possibility that while a field of probability $(\mathfrak{F}, \mathsf{P})$ may be regarded as the image (idealized, however) of actual random events, the extended field of probability $(B\mathfrak{F}, \mathsf{P})$ will still remain merely a mathematical structure.
>
> Thus sets of $B\mathfrak{F}$ are generally merely ideal events to which nothing corresponds in the outside world. However, if reasoning which utilizes the probabilities of such ideal events leads us to a determination of the probability of an actual event of $\mathfrak{F}$, then, from an empirical point of view also, this determination will automatically fail to be contradictory. (Kolmogorov 1956)

And indeed, upon taking the completion of a probability space we always have an isometric copy of the original space inside the new one.

We summarise the result of the above discussion in the following definition, which is the one we will be concerned with going forward:

Definition 3.2: Abstract probability spaces

An ***abstract probability space*** is a generalised abstract probability space $(\mathcal{F}, P)$, where $\mathcal{F}$ is an abstract σ-algebra.

4 Set-theoretical probability theory

4.1 • Basic definitions and properties

We begin by recalling the standard definitions in order to compare them to the abstract versions we have considered above:

Definition 4.1: Measurable spaces

A ***(concrete) σ-algebra*** in a set Ω is a collection $\mathcal{F}$ of subsets of Ω such that

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

(ii) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, and

(iii) if $(A_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{F}$, then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

The sets in $\mathcal{F}$ are called ***$\mathcal{F}$-measurable***, and the pair $(\Omega, \mathcal{F})$ is called a ***measurable space***.

Note that a σ-algebra by this definition is indeed an *abstract* σ-algebra in the sense of Definition 2.15. Also contrast the definition of a σ-algebra with that of a *set algebra*: in the latter case we also require it to contain Ω and be closed under complementation, but we only require it to be closed under *finite* unions. In particular a set algebra is a Boolean algebra.

Definition 4.2: Measure spaces and probability spaces

Let $(\Omega, \mathcal{F})$ be a measurable space. A ***measure*** on $(\Omega, \mathcal{F})$ is a map $\mu\colon \mathcal{F} \to [0, \infty]$ such that

(i) $\mu(\emptyset) = 0$, and

(ii) μ is countably additive, i.e. for every sequence $(A_n)_{n\in\mathbb{N}}$ of pairwise disjoint sets in $\mathcal{F}$ we have

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n).$$

The triple $(\Omega, \mathcal{F}, \mu)$ is called a ***(concrete) measure space***. If $\mu(\Omega) = 1$ then μ is called a ***probability measure*** and the triple $(\Omega, \mathcal{F}, \mu)$ a ***(concrete) probability space***.

Again, if $(\Omega, \mathcal{F}, \mu)$ is a measure space thus defined, then $(\mathcal{F}, \mu)$ is also an *abstract* measure space of the type considered in Definition 2.16. However, even if μ is a probability measure, $(\mathcal{F}, \mu)$ is not necessarily an *abstract* probability space as defined in Definition 3.2.

The trouble is that $\mu(A) = 0$ does not necessarily imply that $A = \emptyset$, i.e. μ need not be positive definite. In an abstract measure space (perhaps of the generalised kind) $(\mathcal{F}, \mu)$, the only information we have about the relationship between two elements $x, y \in \mathcal{F}$ is whether they are comparable and, if so, which one is greater.[1] Or in other terms for events E and F, whether one of them implies the other or not.

By contrast, elements of a concrete σ-algebra are subsets of an underlying set. This has several important implications: First of all, this vastly increases the number of objects we have to contend with, namely *every* subset of Ω, not just those lying in $\mathcal{F}$. This may have benefits of a technical nature; certainly it would have simplified many of the arguments in the previous section on Boolean algebras.

Secondly, it forces us to take the very elements of the *sample space* Ω seriously. Usually these are referred to as *outcomes* or, e.g. by Kolmogorov, *elementary events*. But this may be conceptually problematic:

> [T]he notion of an elementary event is an artificial superstructure imposed on the concrete notion of an event. In reality, events are not composed of elementary events, but elementary events originate in the dismemberment of composite events. (Kolmogorov and Jeffrey 1995)

In other words, arriving at the idea of an elementary event requires *analysis* of the collection of events that is given in some random experiment.

A simple example may help elucidate this point: Consider the random experiment consisting of rolling a six-sided die. Say this die is loaded in such a way that it is impossible to roll a 6, and that it is very likely to roll a 1 or a 2. We construct an abstract probability space $(\mathcal{F}, P)$ to model the outcome of the die roll. The event space might be

$$\mathcal{F} = \{0, A_1, \ldots, A_5, E, F, 1\},$$

where we interpret the event A_i as 'the result was i' for $i = 1, \ldots, 5$, E as 'the result was even' and F as 'the result was odd'. There is no event A_6 since it is impossible to roll a 6. Certainly we would require that

(1) $A_1, \ldots, A_5$ be pairwise disjoint,

[1] Recall that the ordering on a lattice completely determines the lattice structure.

(2) $E = A_2 \vee A_4$ and $F = A_1 \vee A_3 \vee A_5$ (from which it follows that E and F are disjoint), and

(3) $E \vee F = 1$.

Assigning probabilities to each event we might find that $P(A_1) = P(A_2) = 0.35$, and that $P(A_3) = P(A_4) = P(A_5) = 0.1$. Additivity of P would then imply that $P(E) = 0.45$ and $P(F) = 0.55$.

It is tempting to *identify* each of E and F with the (possible) events that imply them, i.e. identify E with the set $\{A_2, A_4\}$ and F with the set $\{A_1, A_3, A_4\}$ of events. Maybe we are even tempted to include a hypothetical event A_6 in the former set. But notice that this is an *analysis* of the events E and F. It would be quite possible to grasp the meaning of the event E without immediately enumerating each of the 'elementary' events it consists of. Indeed events in even set-theoretic probability theory (and objects in mathematics as a whole, for that matter) are often defined in terms of their properties, not of their constituents. Or to put it in other terms: by their *intension* and not their *extension*.

On the other hand, the set-theoretic approach and the possibility of non-empty null sets offer certain conceptual advantages as well. Perhaps we would like there to be an event A_6 even if it is impossible. Certainly our model of the die seems incomplete without it; the die is, after all, six-sided! Admittedly this picture seems rather artificial since the probability of rolling a 6 using a physical die is surely positive, however small it happens to be. So imagine that this die appears in a video game or in a fantasy novel where this might be a more easily digestible proposition.

Furthermore, if we actually do want to use our theory of probability to *model* physical phenomena, then it might be perfectly reasonable to include some event in the model on conceptual grounds, even if this event happens to be assigned probability zero in the model. Perhaps this fact is due to numerical approximation, missing information or a state of affairs that comes about after each event under consideration has been identified. Indeed, we may consider a (concrete or abstract) σ-algebra and interpret its elements as events, even if there is no particular probability measure defined on it. Thus the system of events seems in some sense prior to the assignment of probabilities to the events.

4.2 • Extending premeasures to σ-algebras

Recall that we in Lemma 2.17 proved that the completion of a generalised abstract measure space (B, μ) is an abstract measure space under the assumption that μ is positive definite. Since this is no longer the case we need another

way of extending a finitely additive probability measure on a set algebra to a σ-algebra. Furthermore, in Lemma 2.12 we showed that (B, μ) automatically satisfies the axiom of continuity if μ is positive definite. We will thus also have to overcome this obstacle.

We begin with the latter: Let $\mathcal{A}$ be an algebra on a set Ω, and let $\mu_0 \colon \mathcal{A} \to [0, \infty]$ be a ***premeasure*** on $\mathcal{A}$. Recall that this means that $\mu_0(\emptyset) = 0$, and that

$$\mu_0\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n=1}^{\infty} \mu_0(A_n)$$

for any sequence $(A_n)_{n\in\mathbb{N}}$ of pairwise disjoint sets in $\mathcal{A}$ such that $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$. Notice that if μ_0 is finite, this says exactly that $(\mathcal{A}, \mu_0)$ is a generalised abstract measure space. But if μ_0 is finite, then countable additivity of the above sort follows from Proposition 2.13 if only $(\mathcal{A}, \mu_0)$ satisfies the axiom of continuity.

We will attempt to motivate the axiom of continuity for a generalised abstract probability space $(\mathcal{F}, P)$. In this setting the axiom says that, given a decreasing sequence of events $(E_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$ such that $\bigwedge_{n\in\mathbb{N}} E_n = 0$, we have $\lim_{n\to\infty} P(E_n) = 0$.

If (E_n) is eventually constant, E_n must equal 0 for large enough n, in which case the axiom is obvious. If instead all E_n are possible, the assumption that $\lim_{n\to\infty} P(E_n) = 0$ says that it is still impossible for all E_n to obtain. Since the sequence is decreasing, this must mean that the E_n become increasingly less likely to occur, and the probability that E_n occurs must get vanishingly small as n tends to infinity. In other words, $P(E_n)$ must approach zero in the limit $n \to 0$.

Of course this is not a proof of the axiom of continuity, but it illustrates that it a quite natural assumption. Thus we will henceforth assume that μ_0 is indeed a premeasure on a set algebra $\mathcal{A}$.

Next, recall that if $\mathcal{J}$ is a collection of subsets of Ω containing $\emptyset$ and $\mu \colon \mathcal{J} \to [0, \infty]$ satisfies $\mu(\emptyset) = 0$, then

$$\mu^*(A) = \inf\Big\{\sum_{n=1}^{\infty} \mu(B_n) \;\Big|\; (B_n)_{n\in\mathbb{N}} \subseteq \mathcal{J} \text{ and } A \subseteq \bigcup_{n\in\mathbb{N}} B_n\Big\} \tag{4.1}$$

defines an outer measure on Ω. We denote the σ-algebra of μ^*-measurable sets by $\mathcal{M}(\mu^*)$. In the case $\mathcal{J} = \mathcal{A}$ and $\mu = \mu_0$, the following theorem guarantees that we may extend μ_0 to an almost unique countably additive measure:

Theorem 4.3

Let μ_0 be a premeasure on an algebra $\mathcal{A}$ in a set Ω, and let μ^* be given by (4.1) with $\mu = \mu_0$ and $\mathcal{J} = \mathcal{A}$. Then $\mathcal{A} \subseteq \mathcal{M}(\mu^*)$, and $\mu^*|_{\mathcal{A}} = \mu_0$.

In particular, μ^* restricts to a measure μ on $\mathcal{F} = \sigma(\mathcal{A})$ whose restriction to $\mathcal{A}$ is μ_0. If ν is another measure on $\mathcal{F}$ that extends μ_0, then $\nu(A) \leq \mu(A)$ for all $A \in \mathcal{F}$,

with equality when $\mu(A) < \infty$. If μ_0 is σ-finite, then $\mu = \nu$.

Proof. Folland (2007, Theorem 1.14). □

Thus in the case where μ_0 is a probability premeasure, there exists a *unique* extension to a probability measure P on the σ-algebra $\mathcal{F}$ generated by $\mathcal{A}$.

To sum up: Given a finitely additive probability measure μ_0 on a set algebra $\mathcal{A}$ in a set Ω, the pair $(\mathcal{A}, \mu_0)$ is a generalised abstract measure space. However, since μ_0 is not necessarily positive definite, it does not immediately extend to a measure on an abstract σ-algebra, the issue being that $(\mathcal{A}, \mu_0)$ does not satisfy the axiom of continuity. Thus we must impose this axiom, and we have argued that this is reasonable. Under this assumption μ_0 becomes a premeasure and thus extends uniquely to a measure P on $\mathcal{F} = \sigma(\mathcal{A})$.

Thus we arrive at the usual conception of a probability space as a measure space $(\Omega, \mathcal{F}, P)$ with P a countably additive probability measure. However, we still don't really understand how to think about σ-algebras. Can we interpret σ-algebras in some way, or are they merely a convenience that doesn't mean anything in practice?

4.3 • Conditional probability and independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Given events $A, B \in \mathcal{F}$ with $P(B) > 0$, the ***conditional probability of A given B*** is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Notice that the map $A \mapsto P(A \mid B)$ is a probability measure on B. We interpret it as the probability that A obtains given the knowledge that B has already occurred. In the frequentist interpretation of probability, if the appropriate random experiment is performed N times, and B occurs $n(B)$ times and $A \cap B$ $n(A \cap B)$ times, then we would expect that the probability that A occurs *given* B to be approximated by the proportion of 'successful' outcomes:

$$P(A \mid B) \approx \frac{n(A \cap B)}{n(A)} = \frac{n(A \cap B)/N}{n(A)/N} \approx \frac{P(A \cap B)}{P(B)}.$$

Another way to 'derive' the formula for $P(A \mid B)$ is to consider the measure P_B on the measurable space $(B, \mathcal{F}_B)$, where $\mathcal{F}_B = \{A \cap B \mid A \in \mathcal{F}\}$, given by $P_B(A) = kP(A \cap B)$, where $k > 0$ is a constant to be determined. Interpreting $P_B(A)$ to be the conditional probability of A given B, we would like P_B to be a probability measure. This requires that

$$1 = P_B(B) = kP(B \cap B) = kP(B),$$

or equivalently that $k = 1/P(B)$, thus recovering the formula above.

In the case that knowing that B has occurred yields no knowledge about the likelihood of A, we would expect that $P(A \mid B) = P(A)$, or equivalently

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

That is, A and B are *independent*:

Definition 4.4: Independence I

Let I be a non-empty index set. A family $(A_i)_{i \in I}$ of events from $\mathcal{F}$ is called ***independent*** relative to P if

$$P\Big(\bigcap_{\nu=1}^{n} A_{i_\nu}\Big) = \prod_{\nu=1}^{n} P(A_{i_\nu}),$$

for any finite subset $\{i_1, \ldots, i_n\}$ of distinct elements of I.

The intuition being that the knowledge that some A_i has obtained does not affect our knowledge of the other A_i. This is clear in the case when I contains two elements. To understand the extension to arbitrary collections of events, let us first look at the family $\{A, B, C\}$ of events from $\mathcal{F}$.

If the events A and B are independent, then we interpret this as meaning that knowing that A has occurred gives us no information about whether B has occurred or not. That is, it does not allow us to approximate $P(B)$ any better. Furthermore, if $\{A, B, C\}$ is independent, then we would similarly interpret this to mean that knowing that A has occurred does not give us information about whether B or C have occurred. But certainly, if the collection $\{A, B, C\}$ is independent, then e.g. the subcollection $\{A, B\}$ should also be independent. So if $\{A, B, C\}$ is independent, then we would in particular have $P(A \cap B) = P(A)P(B)$.

Next we attempt to interpret, or derive, the identity $P(A \cap B \cap C) = P(A)P(B)P(C)$. We imagine doing an experiment to determine whether A has occurred, and then doing another experiment to determine whether B has occurred. Then we would know whether $A \cap B$ has occurred, so if neither A nor B can provide information about C, then it seems reasonable to think that $A \cap B$ cannot either.

Hence, if $\{A, B, C\}$ is to be independent, then $\{A \cap B, C\}$ is also to be independent. Furthermore, we already argued that $\{A, B\}$ should be independent, so we find that

$$P(A \cap B \cap C) = P\Big((A \cap B) \cap C\Big) = P(A \cap B)P(C) = P(A)P(B)P(C).$$

Hence this identity is a natural consequence of our interpretation of events and of independence.

This obviously extends to any finite collection of events. But what does it mean for an infinite collection of events to be independent? Above we justified requiring the identity $P(A \cap B \cap C) = P(A)P(B)P(C)$ in the definition of independence by appealing to an ability to perform experiments to judge whether the events A, B, C had occurred. While it is clearly possible, at least in principle, to perform any finite number of experiments (as we argued in our discussion of the generalised abstract probability space of events), it is not so clear that we should be able to perform *infinitely* many such experiments. Hence for an infinite collection $\mathcal{A}$ of events to be independent, it seems to be sufficient that any finite subcollection of $\mathcal{A}$ be independent, which is equivalent to Definition 4.4.

We may generalise this definition of independence in the following way:

Definition 4.5: Independence II

Let $(\mathcal{C}_i)_{i\in I}$ be a family of sets $\mathcal{C}_i \subseteq \mathcal{F}$ of events. This family is called ***independent*** relative to P if

$$P\Big(\bigcap_{\nu=1}^{n} A_{i_\nu}\Big) = \prod_{\nu=1}^{n} P(A_{i_\nu}),$$

for any choice of events $A_{i_\nu} \in \mathcal{C}_{i_\nu}$ and any finite subset $\{i_1, \dots, i_n\}$ of distinct elements of I.

Remark 4.6. Notice that, since the intersections and products above are finite, $(\mathcal{C}_i)_{i\in I}$ is independent if and only if $(\mathcal{C}_i)_{i\in J}$ is independent for every finite subset $J \subseteq I$. ⌟

This definition reduces to the first one if all $\mathcal{C}_i$ are singletons. But notice what this definition says: Choosing a single event A_i from each $\mathcal{C}_i$, these A_i are independent. The definition does not imply any relationship between the events in each $\mathcal{C}_i$. An example, adapted from Example 6.1 in Bauer (1995), will illustrate this point:

Example 4.7. Consider two throws of a fair die. The sample space is then $\Omega = \{1, \dots, 6\}^2$, and we consider the probability space $(\Omega, 2^\Omega, P)$, where P assigns equal probability $1/36$ to each elementary event $\{(i, j)\}$. Let A_1 and A_2 be the events that on the first throw an even or odd number showed up, respectively, and let A_3 be the event that the sum of the two throws is odd. Now collect the three events in $\mathcal{C}_1 = \{A_1, A_2\}$ and $\mathcal{C}_2 = \{A_3\}$. Then $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent even through A_1 and A_2 are not. ⌟

The question then becomes, why are we interested in this generalisation of independence? Independence of two collections $\mathcal{C}_1$ and $\mathcal{C}_2$ is supposed to mean that the information contained in each is in some way independent of

the information contained in the other. By 'information' we mean something like: the ability to better predict the outcome of a random experiment, or in other words improved knowledge of the probabilities of events.

Given that $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent, is there any more we can say? Is it the case, for instance, that each event $\mathcal{C}_1$ is independent of the (finite or countable) intersection of events in $\mathcal{C}_2$? What about unions or complements? In the case of intersections the answer is negative:

Example 4.8. Let $(\Omega, 2^\Omega, P)$ be the probability space defined in Example 4.7, and let A_1 and A_3 be the same events as before. But now let A_2' be the event that the *second* throw yielded an odd number, and put $\mathcal{C}_1' = \{A_1, A_2'\}$. Again $\mathcal{C}_1'$ and $\mathcal{C}_2$ are easily seen to be independent, and indeed so are A_1 and A_2'. However, $A_1 \cap A_2'$ and A_3 are manifestly not independent: For while $P(A_1 \cap A_2') = 1/4$ by independence, we have

$$P\Big((A_1 \cap A_2') \cap A_3\Big) = 0,$$

since if each throw turns up odd, the sum must be even.

Incidentally, this also shows that the requirement in Definition 4.4 that P be multiplicative for *every* finite subset of I is necessary; it is not enough that the events be *pairwise* independent. ⌟

As for unions and complementation, we have slightly more success. To do this discussion justice we introduce another piece of terminology:

Definition 4.9: δ-systems

A collection $\mathcal{D}$ of subsets of a set X is called a ***δ-system*** in X if

(i) $X \in \mathcal{D}$,

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

(iii) $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{D}$ for every increasing sequence $(A_n)_{n\in\mathbb{N}}$ of sets in $\mathcal{D}$.

A δ-system is also variously called a Dynkin class, Dynkin system, d-system, or λ-system. Clearly every σ-algebra is a δ-system. If $\mathcal{S}$ is a collection of subsets of X, then there is a smallest δ-system in X that contains $\mathcal{S}$, namely the intersection of all such δ-systems. We denote this by $\delta(\mathcal{S})$ and say that it is ***generated*** by $\mathcal{S}$.

The motivation for considering δ-systems is twofold. First of all they are significantly simpler than σ-algebras, and working with δ-systems instead of σ-algebras can often we done with no loss of generality, as the following fundamental result shows:

Theorem 4.10: Dynkin's Lemma

Let $\mathcal{S}$ be a collection of subsets of a set X that is closed under finite intersections. Then

$$\delta(\mathcal{S}) = \sigma(\mathcal{S}).$$

Also known as ***Dynkin's π-λ theorem*** since a non-empty collection of sets that is closed under finite intersections is also called a π-system.

Proof. Cohn (2001, Theorem 1.6.2). See also Bauer (2001, Theorem 2.3), though note that Bauer uses a slightly different definition of δ-systems. □

Another source of motivation comes from the following result about finite measures which is important when proving uniqueness of properties of finite or σ-finite measures, but will also be of interest to us below:

Lemma 4.11

Let μ and ν be finite measures on a measurable space $(X, \mathcal{E})$ such that $\mu(X) = \nu(X)$. The family $\mathcal{D} \subseteq \mathcal{E}$ of sets on which μ and ν agree is a δ-system.

Proof. By assumption $X \in \mathcal{D}$. Let $A_1, A_2 \in \mathcal{D}$ with $A_1 \subseteq A_2$. Then

$$\mu(A_2 \setminus A_1) = \mu(A_2) - \mu(A_1) = \nu(A_2) - \nu(A_1) = \nu(A_2 \setminus A_1),$$

since μ and ν are finite. Finally assume that $(A_n)_{n\in\mathbb{N}}$ is an increasing sequence of elements in $\mathcal{D}$. Then by continuity we have

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \nu(A_n) = \nu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big).$$

Thus $\mathcal{D}$ is a δ-system as claimed. □

Our motivation for considering δ-systems, however, is the following result:

Proposition 4.12

Let $(\mathcal{C}_i)_{i\in I}$ be an independent family of sets of events from $\mathcal{F}$. Then the family $(\delta(\mathcal{C}_i))_{i\in I}$ is also independent. In particular, if the $\mathcal{C}_i$ are closed under intersection, the family $(\sigma(\mathcal{C}_i))_{i\in I}$ is independent.

Proof. Fix an index $i_0 \in I$, and choose sets $A_{i_\nu} \in \mathcal{C}_{i_\nu}$ for distinct indices $i_1, \dots, i_n \in I \setminus \{i_0\}$. Then define measures P_1 and P_2 on $\mathcal{F}$ by

$$P_1(A) = P\Big(A \cap \bigcap_{\nu=1}^{n} A_{i_\nu}\Big) \quad \text{and} \quad P_2(A) = P(A)\prod_{\nu=1}^{n} P(A_{i_\nu}),$$

for $A \in \mathcal{F}$. Notice that P_1 and P_2 agree on $\mathcal{C}_{i_0}$ by independence, so since $P_1(\Omega) = P_2(\Omega)$, Lemma 4.11 implies that they also agree on $\delta(\mathcal{C}_{i_0})$. It follows that

$$P\Big(A \cap \bigcap_{\nu=1}^{n} A_{i_\nu}\Big) = P(A)\prod_{\nu=1}^{n} P(A_{i_\nu})$$

for all choices of sets $A_{i_\nu} \in \mathcal{C}_{i_\nu}$. But this precisely expresses the independence of the family $(\mathcal{C}_i)_{i\in I}$ with $\mathcal{C}_{i_0}$ replaced by $\delta(\mathcal{C}_{i_0})$.

By Remark 4.6 we may assume that I is finite. Thus performing a finite number of such replacements, once for each index in I, proves the first claim. The second claim follows by Dynkin's lemma. □

Proposition 4.13: Combining σ-algebras

Let $(\mathcal{C}_i)_{i\in I}$ be an independent family of $\cap$-stable sets $\mathcal{C}_i \subseteq \mathcal{F}$, and let $(I_j)_{j\in J}$ be a partition of I. If $\mathcal{G}_j = \sigma(\bigcup_{i\in I_j} \mathcal{C}_i)$, then $(\mathcal{G}_j)_{j\in J}$ is independent.

Proof. Let $\tilde{\mathcal{C}}_j$ denote the collection of sets

$$A_{i_1} \cap \cdots \cap A_{i_n},$$

where $i_1, \ldots, i_n$ are distinct elements from I_j, and $A_{i_\nu} \in \mathcal{C}_{i_\nu}$. Then $\tilde{\mathcal{C}}_j$ is $\cap$-stable, and the family $(\tilde{\mathcal{C}}_j)_{j\in I_j}$ is independent. We clearly have $\mathcal{G}_j = \sigma(\tilde{\mathcal{C}}_j)$, so Proposition 4.12 implies that $(\mathcal{G}_j)_{j\in J}$ is independent. □

4.4 • σ-algebras and information

We interpret Proposition 4.12 as follows: Say that we are interested in an event $A \in \mathcal{C}_1$. The independence of $\mathcal{C}_1$ and $\mathcal{C}_2$ tells us that no single event $B \in \mathcal{C}_2$ can provide us with information about A. The result above then implies that neither can any event in $\delta(\mathcal{C}_2)$. In particular, taking complements and increasing countable unions cannot give us information that was available in a single event in $\mathcal{C}_2$ to begin with.

But if we are thinking of $\mathcal{C}_2$ as information, then surely it makes sense to consider two different events $B_1, B_2 \in \mathcal{C}_2$ simultaneously. After all, knowing whether B_1 and B_2 each have occurred we can conclude whether $B_1 \cup B_2$ and $B_1 \cap B_2$ have occurred as well. In other words, $\mathcal{C}_2$ must be a set algebra if it is to model information. But then Proposition 4.12 implies that $\mathcal{C}_1$ and $\sigma(\mathcal{C}_2)$ are independent: no information in $\sigma(\mathcal{C}_2)$ can improve our knowledge of A if no single event in the algebra $\mathcal{C}_2$ can.

In some sense then, if $\mathcal{A}$ is a set algebra (so in particular $\mathcal{A}$ is closed under intersection) in Ω, the generated σ-algebra $\sigma(\mathcal{A})$ carries no more information than $\mathcal{A}$. This picture is not quite complete, of course: If $\mathcal{A}$ is unable to give

us *any information whatsoever* about some event E, then $\sigma(\mathcal{A})$ cannot either. But why should that mean that there is no increase in information at all when passing from $\mathcal{A}$ to $\sigma(\mathcal{A})$? Might there not be some other way of cashing out this idea of 'information' than the ability to predict the probability of events?

To further probe the interpretation of σ-algebras as information, we introduce the following terminology:

Definition 4.14

Let $\mathcal{C}$ be a collection of subsets of a set Ω. We say that points $\omega, \omega' \in \Omega$ are ***$\mathcal{C}$-equivalent*** if, for every $A \in \mathcal{C}$, they both lie in A or A^c, i.e. if $\mathbf{1}_A(\omega) = \mathbf{1}_A(\omega')$. In this case we write $\omega \sim_{\mathcal{C}} \omega'$.

The relation $\sim_{\mathcal{C}}$ induces a partition of Ω called the ***$\mathcal{C}$-partition***.

Another way to put this is that ω and ω' lie in all the same sets in $\mathcal{C}$. Compare this with the *topological indistinguishability* relation from point-set topology: Two points x, y in a topological space $(X, \mathcal{T})$ are called topologically indistinguishable if they have all the same (open) neighbourhoods, i.e. if every open set either contains both x and y or contains neither. This is then the same as x and y being $\mathcal{T}$-equivalent, or in the notation above, $x \sim_{\mathcal{T}} y$.

Lemma 4.15

Given points $\omega, \omega' \in \Omega$ we have $\omega \sim_{\mathcal{C}} \omega'$ if and only if $\omega \sim_{\sigma(\mathcal{C})} \omega'$. In particular, the $\mathcal{C}$- and $\sigma(\mathcal{C})$-partitions of Ω coincide.

Proof. The latter clearly implies the former, so assume that $\omega \sim_{\mathcal{C}} \omega'$. The collection of subsets $A \subseteq \Omega$ such that $\mathbf{1}_A(\omega) = \mathbf{1}_A(\omega')$ is clearly a σ-algebra, and it contains $\mathcal{C}$ by assumption. But then it must also contain $\sigma(\mathcal{C})$, so $\omega \sim_{\sigma(\mathcal{C})} \omega'$.□

Now consider drawing a random element ω from Ω. If an observer has the information $\mathcal{C}$, i.e. they know whether or not $\omega \in A$ for all $A \in \mathcal{C}$, then all they know is which $\mathcal{C}$-equivalence class ω lies in. But by the lemma, this class is the same as the $\sigma(\mathcal{C})$-equivalence class of ω, so again passing from $\mathcal{C}$ to $\sigma(\mathcal{C})$ yields no new information.

Thus it really does not seem like σ-algebras can do any more work than the algebras that generate them. And since we are comfortable with algebras carrying some kind of information, maybe σ-algebras do too.

To show that we cannot indiscriminately think of σ-algebras as information, we give an example of a case in which we seem to have both no information and a lot of information. This is Example 4.10 in Billingsley (1995).

Example 4.16. Consider the probability space $([0,1], \mathcal{F}, \lambda)$, where $\mathcal{F}$ is the Borel algebra $\mathcal{B}([0,1])$ and λ is the Lebesgue measure restricted to $[0,1]$. Furthermore, let $\mathcal{G}$ be the sub-σ-algebra of $\mathcal{F}$ consisting of countable and cocount-

able sets. Then the measure of each element in $\mathcal{G}$ is either 0 or 1, so $\mathcal{F}$ and $\mathcal{G}$ are independent. Given some event $E \in \mathcal{F}$,

(a) $\mathcal{F}$ contains *no* information about E, in the sense that E is independent of $\mathcal{F}$.

On the other hand, the $\mathcal{G}$-equivalence classes are singletons, so

(b) $\mathcal{F}$ contains *all* the information about E, for given $\mathcal{F}$ an observer knows precisely which ω was drawn, hence whether E occurred or not.

These are in apparent contradiction, so it must not be the case that we can always interpret σ-algebras as information.

Of course this example is rather artificial (indeed $\mathcal{G}$ is not even countably generated, and λ restricted to $\mathcal{G}$ is also almost trivial), and it should not be seen as prohibiting the interpretation of σ-algebras as information entirely.⌟

4.5 • Kolmogorov's 0-1 law

Let $\mathcal{F}$ be a σ-algebra, and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a sequence of σ-algebras contained in $\mathcal{F}$. For $n \in \mathbb{N}$ define σ-algebras

$$\mathcal{T}_n = \bigvee_{i\le n} \mathcal{F}_i = \sigma\Big(\bigcup_{i\le n} \mathcal{F}_i\Big) \quad \text{and} \quad \mathcal{T}^n = \bigvee_{n<i} \mathcal{F}_i = \sigma\Big(\bigcup_{n<i} \mathcal{F}_i\Big),$$

and further define

$$\mathcal{T}_\infty = \bigvee_{n\in\mathbb{N}} \mathcal{T}_n = \sigma\Big(\bigcup_{n\in\mathbb{N}} \mathcal{T}_n\Big) \quad \text{and} \quad \mathcal{T}^\infty = \bigwedge_{n\in\mathbb{N}} \mathcal{T}^n = \bigcap_{n\in\mathbb{N}} \mathcal{T}^n.$$

We call $\mathcal{T}_n$ the ***past of $\mathcal{F}_n$*** and $\mathcal{T}^n$ the ***future of $\mathcal{F}_n$***, and we furthermore call $\mathcal{F}^\infty$ the ***total history*** and $\mathcal{T}^\infty$ the ***ultimative future*** or the ***tail-σ-algebra*** of the sequence $(\mathcal{F}_n)$. Notice that $\mathcal{T}^\infty \subseteq \mathcal{T}^1 \subseteq \mathcal{T}_\infty$, where the last inclusion follows since $\mathcal{F}_i \subseteq \mathcal{T}_{i+1} \subseteq \mathcal{T}_\infty$ for all $i \in \mathbb{N}$.

Lemma 4.17

In the notation above, if the sequence $(\mathcal{F}_n)$ is independent, then for each $n \in \mathbb{N}\cup\{\infty\}$ the σ-algebras $\mathcal{T}_n$ and $\mathcal{T}^n$ are independent.

Proof. This follows from Proposition 4.13 for $n \in \mathbb{N}$, and hence $\mathcal{T}_n$ and $\mathcal{T}^\infty$ are also independent. Now notice that $\tilde{\mathcal{T}} = \bigcup_{n\in\mathbb{N}} \mathcal{T}_n$ is $\cap$-stable: For if $A, B \in \tilde{\mathcal{T}}$, then since the sequence $(\mathcal{T}_n)_{n\in\mathbb{N}}$ is increasing, A and B lie in a common $\mathcal{T}_n$, and this is obviously $\cap$-stable. But since $\tilde{\mathcal{T}}$ and $\mathcal{T}^\infty$ are independent, it follows from Proposition 4.12 that $\mathcal{T}_\infty = \sigma(\tilde{\mathcal{T}})$ and $\mathcal{T}^\infty$ are also independent. □

Theorem 4.18: Kolmogorov's 0-1 law

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be an independent sequence of σ-algebras $\mathcal{F}_n \subseteq \mathcal{F}$. Then $P(A) \in \{0,1\}$ for all $A \in \mathcal{T}^\infty$.

Proof. Recall that $\mathcal{T}^\infty \subseteq \mathcal{T}_\infty$. Then $A \in \mathcal{T}_\infty$, so Lemma 4.17 implies that A is independent of itself. The claim follows. □

4.6 • Random variables

There are various ways of motivating the definition of measurability of functions in general measure theory: In order to make possible integration, by analogy with continuous maps between topological spaces, or heuristically by appealing to an intuitive notion of measurability.

For random variables we do not have this luxury. Why should random variables be integrable? Why should random variables have anything to do with continuity? Hence we focus on the third point and try to explore what measurability means for random variables.

We begin by fixing terminology and notation: If $(X, \mathcal{E})$ and $(Y, \mathcal{F})$ are measure spaces, a map $f : X \to Y$ is said to be ***$(\mathcal{E}, \mathcal{F})$-measurable*** if $f^{-1}(B) \in \mathcal{E}$ for all $B \in \mathcal{F}$. Denote by $\mathcal{M}(\mathcal{E})$ the space of functions $X \to \mathbb{R}$ that are $(\mathcal{E}, \mathcal{B}(X))$-measurable.

A ***random variable*** on a probability space $(\Omega, \mathcal{F}, P)$ is a function in $\mathcal{M}(\mathcal{F})$. If a random variable $\mathsf{X}\colon \Omega \to \mathbb{R}$ is measurable, then in particular

$$\{\mathsf{X} = x\} := \{\omega \in \Omega \mid \mathsf{X}(\omega) = x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. Thus the information in $\mathcal{F}$ at least needs to determine on which sets X takes on which values for X to be measurable. The converse, however, is not the case:

Example 4.19. Let $\mathcal{E}$ be the countable, cocountable σ-algebra on $\mathbb{R}$, and let $\mathsf{X}\colon (\mathbb{R}, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(X))$ be the identity function. Then $\{\mathsf{X} = x\} = \{x\} \in \mathcal{E}$, but $\mathsf{X}^{-1}([0,\infty)) = [0,\infty)$, which is neither countable nor cocountable. Hence X has measurable fibres despite not being measurable. ⌟

Is measurability then too strong a condition to impose on a random variable? Let us imagine that some outcome $\omega \in \Omega$ has obtained, and that there is enough information in $\mathcal{F}$ to determine $\mathsf{X}(\omega)$. Then certainly we should be able to answer questions on the form 'is $\mathsf{X}(\omega) < a$?' for $a \in \mathbb{Q}$. That is, we require that $\mathsf{X}^{-1}((-\infty, a)) \in \mathcal{F}$. But since sets on the form $(-\infty, a)$ for $a \in \mathbb{Q}$ generate $\mathcal{B}(\mathbb{R})$, this implies that X is measurable. Thus measurability seems to be a consequence of our claim that $\mathcal{F}$ should determine the values that X takes.

Let us again say that some outcome $\omega \in \Omega$ has obtained, and that we are not able to simply observe the value $\mathsf{X}(\omega)$. How do we then determine $\mathsf{X}(\omega)$ using only the information in $\mathcal{F}$? We have interpreted 'information' in this sense as being able to answer questions on the form 'is $\omega \in A$?' for $A \in \mathcal{F}$. We will try to construct a procedure by which we can determine, or at least approximate, $\mathsf{X}(\omega)$ only by asking questions on this form.

Of course we cannot ask whether $\omega \in \mathsf{X}^{-1}(\mathsf{X}(\omega))$, since we don't know $\mathsf{X}(\omega)$. Furthermore, it seems unreasonable that we should be able to ask whether $\omega \in A$ for *all* $A \in \mathcal{F}$. Rather, it seems like we need a general way of calculating the value some random variable $\Omega \to \mathbb{R}$ takes, given that ω has occurred.

Suppose for definiteness that we have drawn an element $\omega \in \Omega$ such that $x = \mathsf{X}(\omega) > 0$, and that we wish to determine x within some error $\varepsilon > 0$. First, check if x lies in the interval $(-\infty, n)$ for $n \in \mathbb{N}$, starting with $n = 1$. After a finite number of checks we find that $x \in I_0 \coloneqq [n_0, n_0 + 1)$ for some $n_0 \in \mathbb{N}$. Next, we construct a sequence $(I_n)_{n\in\mathbb{N}}$ as follows: For $n \in \mathbb{N}_0$, let $a_n \in \mathbb{Q}$ be the midpoint of I_n, and check if $x \in (-\infty, a_n)$. If this is the case, let $I_{n+1} = I_n \cap (-\infty, a_n)$, and otherwise let $I_{n+1} = I_n \cap [a_n, \infty)$. Since the interval length is halved in each step, then for some $N \in \mathbb{N}$ the length of I_N is less than ε. Hence we obtain in finitely many steps an approximation $a_N \in \mathbb{Q}$ of x such that $|x - a_N| < \varepsilon$.

Notice that we only used intervals on the form $(-\infty, a)$ with $a \in \mathbb{Q}$ during this process. Thus only knowledge of the events $\{\mathsf{X} \in (-\infty, a)\}$ is necessary to determine $\mathsf{X}(\omega)$, at least approximately. And since the sets $(-\infty, a)$ generate the Borel algebra $\mathcal{B}(\mathbb{R})$, this requirement is the same as requiring X to be $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable.

In total, we arrive at the interpretation that a random variable $\mathsf{X}\colon \Omega \to \mathbb{R}$ is $\mathcal{F}$-measurable if and only if $\mathcal{F}$ contains enough information to determine the values that X takes. The minimal information required is then $\sigma(\mathsf{X})$, the σ-algebra on Ω generated by X.

The factorisation lemma

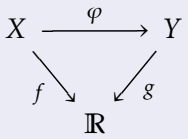
We prove a fundamental lemma that helps shed further light on the interpretation of measurability of random variables.

Proposition 4.20: The factorisation lemma

Let X be a set, let $(Y, \mathcal{F})$ be a measurable space, and let $\varphi\colon X \to Y$ be a map. Equip X with the σ-algebra $\mathcal{E} = \sigma(\varphi)$ generated by φ. Then

$$\mathcal{M}(\mathcal{E}) = \{g \circ \varphi \mid g \in \mathcal{M}(\mathcal{F})\}. \tag{4.2}$$

That is, for every $f \in \mathcal{M}(\mathcal{E})$ there exists a $g \in \mathcal{M}(\mathcal{F})$ such that the diagram



commutes.

In other words, a function $f\colon X \to \mathbb{R}$ is $\sigma(\varphi)$-measurable if and only if it factors through φ.

Proof. Denote the set on the right-hand side of (4.2) by W. It is clear that W is a subspace of $\mathcal{M}(\mathcal{E})$. If $A \in \mathcal{E}$, then $A = \varphi^{-1}(B)$ for some $B \in \mathcal{F}$. Notice then that

$$\mathbf{1}_A = \mathbf{1}_{\varphi^{-1}(B)} = \mathbf{1}_B \circ \varphi \in \mathcal{F},$$

so $\mathbf{1}_A \in W$. Hence W contains all simple $\mathcal{E}$-measurable functions.

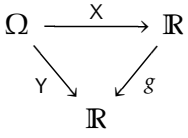
Now let $f \in \mathcal{M}(\mathcal{E})^+$ be a non-negative function. Then there exists an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of simple functions, i.e. functions in W, such that f is the pointwise limit of (f_n). For each $n \in \mathbb{N}$ there is a function $g_n \in \mathcal{M}(\mathcal{F})$ such that $f_n = g_n \circ \varphi$. Hence we have for each $x \in X$,

$$f(x) = \sup_{n\in\mathbb{N}} f_n(x) = \sup_{n\in\mathbb{N}} g_n(\varphi(x)) = g(\varphi(x)),$$

where $g\colon Y \to \mathbb{R}$ is given by $g = \sup_{n\in\mathbb{N}} g_n$. This is $\mathcal{F}$-measurable as a function into the extended real line $\overline{\mathbb{R}}$. Next let $B = \{y \in Y \mid g(y) = \infty\}$. Then $\tilde{g} = g\mathbf{1}_{B^c} \in \mathcal{M}(\mathcal{F})$, and since $\varphi(X) \subseteq B^c$ we have $f = \tilde{g} \circ \varphi$ as desired.

Finally let $f \in \mathcal{M}(\mathcal{E})$ be an arbitrary measurable function, and write $f = f^+ - f^-$ where $f^+ = f \vee 0$ and $f^- = -(f \wedge 0)$. Applying the above to f^+ and f^- yields functions $g^+, g^- \in \mathcal{M}(\mathcal{F})$ such that $f^\pm = g^\pm \circ \varphi$. Letting $g = g^+ - g^-$ we obtain $f = g \circ \varphi$, and the theorem is proved. □

Now let X and Y be random variables on $(\Omega, \mathcal{F}, P)$. Then the factorisation lemma says that, if Y is in fact $\sigma(\mathsf{X})$-measurable, then there exists a measurable function $g\colon \mathbb{R} \to \mathbb{R}$ such that $\mathsf{Y} = g(\mathsf{X})$, i.e. such that the diagram



commutes. This is also clearly sufficient for Y to be $\sigma(\mathsf{X})$-measurable.

Let us try to understand this in terms of the information interpretation of measurability: Knowing $\sigma(\mathsf{X})$ is the same as knowing the value $\mathsf{X}(\omega)$ that X

takes at each $\omega \in \Omega$, or at least being able to approximate it. And an ability to, for all $\omega \in \Omega$, determine $\mathsf{Y}(\omega)$ given the value $\mathsf{X}(\omega)$ just means that there is a function $g\colon \mathbb{R} \to \mathbb{R}$ such that $\mathsf{Y} = g(\mathsf{X})$. Hence there is such a function g just in case Y is determined by – that is, measurable with respect to – $\sigma(\mathsf{X})$.

4.7 • Sub-σ-algebras and conditional expectation

If $(\Omega, \mathcal{F}, P)$ is a probability space, then we have seen that we, at least to some degree, can think of $\mathcal{F}$ as the amount of information we have available. More precisely, given any event $A \in \mathcal{F}$ we are able to decide whether A has occurred or not. Furthermore, a map $\mathsf{X}\colon \Omega \to \mathbb{R}$ being $\mathcal{F}$-measurable means that, given $\omega \in \Omega$, we are able to approximate $\mathsf{X}(\omega)$ to arbitrary precision in finitely many steps.

But what happens if we do not have access to all the information in $\mathcal{F}$, but only to the information in some sub-σ-algebra $\mathcal{B} \subseteq \mathcal{F}$? If X is not $\mathcal{B}$-measurable, then there is some $a \in \mathbb{Q}$ (since the intervals with rational endpoints generate $\mathcal{B}(\mathbb{R})$) such that we cannot even tell whether $\mathsf{X} < a$ or not. We wish to construct a random variable that in some sense is the best approximation of X, using only the information in $\mathcal{B}$.

First let $B \in \mathcal{B}$ be an event with $P(B) > 0$. We define the ***conditional expectation of*** X ***given*** B by

$$\mathbb{E}[\mathsf{X} \mid B] = \frac{\mathbb{E}[\mathsf{X}\mathbf{1}_B]}{P(B)} = \frac{1}{P(B)} \int_B \mathsf{X}\,\mathrm{d}P.$$

In other words, $\mathbb{E}[\mathsf{X} \mid B]$ is the mean of X with respect to the probability measure $A \mapsto P(A \mid B)$. If instead $P(B) = 0$, then we let $\mathbb{E}[\mathsf{X} \mid B] = 0$ for simplicity (an arbitrary real number would work). If the mean $\mathbb{E}[X]$ is the best approximation of X given no further information, we interpret $\mathbb{E}[\mathsf{X} \mid B]$ as the best approximation of X given that B has occurred. Since we *know* that B has occurred there is nothing random about $\mathbb{E}[\mathsf{X} \mid B]$. The above of course requires that the mean of X and $\mathsf{X}\mathbf{1}_\mathrm{B}$ exist; to ensure this we will assume that $\mathsf{X} \in \mathcal{L}^1(P)$.

Next let $\mathcal{B}$ be the sub-σ-algebra of $\mathcal{F}$ generated by B, i.e. $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$. We would then expect the conditional expectation of X given $\mathcal{B}$ to be the random variable $\mathbb{E}[\mathsf{X} \mid \mathcal{B}]$ given by

$$\mathbb{E}[\mathsf{X} \mid \mathcal{B}](\omega) = \begin{cases} \mathbb{E}[\mathsf{X} \mid B], & \omega \in B, \\ \mathbb{E}[\mathsf{X} \mid B^c], & \omega \in B^c. \end{cases} \tag{4.3}$$

That is, if we know that $\omega \in B$, then the best approximation of X must be the conditional probability of X given B, and similarly if $\omega \in B^c$. If more generally $\mathcal{B}$ is finitely generated by events $B_1, \ldots, B_n$, then each $\omega \in \Omega$ lies in

precisely one of B_i and B_i^c for $i = 1, \dots, n$. That is, Ω can be partitioned into sets $B_1^* \cap \cdots \cap B_n^*$, where each B_i^* is either B_i or B_i^c. We would then have

$$\mathbb{E}[\mathsf{X} \mid \mathcal{B}](\omega) = \mathbb{E}[\mathsf{X} \mid B_1^* \cap \cdots \cap B_n^*], \quad \text{for} \quad \omega \in B_1^* \cap \cdots \cap B_n^*.$$

Clearly this reduces to (4.3) when $n = 1$. Even more generally, let $\mathcal{F}$ be a partition σ-algebra, say $\mathcal{F} = \sigma(\{B_i \mid i \in I\})$ where $(B_i)_{i \in I}$ is a partition of Ω. In this case we would expect that

$$\mathbb{E}[\mathsf{X} \mid \mathcal{B}] = \sum_{i \in I} \mathbb{E}[\mathsf{X} \mid B_i] \mathbf{1}_{B_i}.$$

Notice that the sum is actually finite at any given $\omega \in \Omega$ (indeed, ω lies in precisely one set B_i, so the ith term is the only one that is nonzero).

If $\mathcal{B}$ is not a partition σ-algebra, it seems unlikely that we should be able to write down $\mathbb{E}[\mathsf{X} \mid \mathcal{B}]$ explicitly in this case. Instead we will try to find properties that $\mathbb{E}[\mathsf{X} \mid \mathcal{B}]$ must have for it to properly be called an approximation of X.

If nothing else, its mean should certainly agree with that of X. Furthermore, for it to be a proper approximation of X it should probably also resemble X locally. But we should be careful here: If we zoom in too far and let $\mathbb{E}[\mathsf{X} \mid \mathcal{B}]$ *equal* X at each point, then we have gotten nowhere. Hence we should 'zoom in' as far as the information in $\mathcal{B}$ allows us to, but no further.

Furthermore, it is an easy theorem (see e.g. Thorbjørnsen 2014, Sætning 10.2.1), whose proof we will not reproduce here, that integrable functions whose integrals on any measurable set agree are equal almost everywhere. In probabilistic terms, $\mathcal{F}$-measurable random variables $\mathsf{X}, \mathsf{Y} \in \mathcal{L}^1(P)$ are equal almost surely in the case that $\mathbb{E}[\mathsf{X} \mid B] = \mathbb{E}[\mathsf{Y} \mid B]$ for all $B \in \mathcal{F}$. But in our case we do not have access to every event in $\mathcal{F}$, only those that lie in $\mathcal{B}$. This motivates the following definition, which is easily seen to be a generalisation of the ones we have considered so far:

Definition 4.21: Conditional expectations

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{F}$, and let $\mathsf{X} \in \mathcal{L}^1(P)$ be a random variable.

A ***conditional expectation of*** X ***given*** $\mathcal{B}$ is a random variable U on $(\Omega, \mathcal{F}, P)$ such that

(i) $\mathsf{U} \in \mathcal{L}^1(P)$,

(ii) U is $\mathcal{B}$-measurable, and

(iii) $\mathbb{E}[\mathsf{U} \mid B] = \mathbb{E}[\mathsf{X} \mid B]$ for all $B \in \mathcal{B}$.

Notice that condition (iii) is equivalent to the requirement that

$$\int_B \mathsf{U} \, \mathrm{d}P = \int_B \mathsf{X} \, \mathrm{d}P$$

for all $B \in \mathcal{B}$. The Radon–Nikodym theorem ensures the existence of a conditional expectation of any $X \in \mathcal{L}^1(P)$, and the arguments above show that it is uniquely determined P-almost surely. We denote any conditional expectation of X given $\mathcal{B}$ by $\mathbb{E}[X \mid \mathcal{B}]$.

We cite without proof a few results that can help us judge whether Definition 4.21 gives us the approximation of X that we wanted. The results can be found in any text on probability theory. Only the final property causes any difficulty: its proof uses the dominated convergence theorem for conditional expectations.

Proposition 4.22

Let $X \in \mathcal{L}^1(P)$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{F}$.

(i) *If $\mathcal{B} = \{\emptyset, \Omega\}$ then $\mathbb{E}[X \mid \mathcal{B}] = \mathbb{E}[X]$.*

(ii) *If X is $\mathcal{B}$-measurable, then $\mathbb{E}[X \mid \mathcal{B}] = X$ P-a.s.*

(iii) *If U is a $\mathcal{B}$-measurable random variable such that $UX \in \mathcal{L}^1(P)$, then*

$$\mathbb{E}[UX \mid \mathcal{B}] = U\mathbb{E}[X \mid \mathcal{B}] \quad P\text{-a.s.}$$

The trivial σ-algebra $\mathcal{B} = \{\emptyset, \Omega\}$ is supposed to model the situation where we have no information: We only know that $\emptyset$ has not occurred and that Ω has. Hence $\mathbb{E}[X \mid \mathcal{B}]$ cannot depend on which events have occurred, so it must be constant. And the best constant approximation of X is $\mathbb{E}[X]$, hence (i).

In contrast, if X is actually $\mathcal{B}$-measurable as in (ii), then $\mathcal{B}$ contains all the information necessary to determine X, and the best approximation of X that only depends on $\mathcal{B}$ is just X itself.

Finally consider the situation in (iii). One sometimes hears that $\mathcal{B}$-measurable variables are 'constant' when calculating conditional expectations with respect to $\mathcal{B}$. This is true, as the result above shows, in the sense that we can 'pull out' such variables from the conditional expectation. Another way to understand this is in relation to (ii): Since U is completely determined by the information in $\mathcal{B}$, the best approximation is just U itself. This result then says that, in this special case, the best approximation of the product UX is the product of the best approximations of U and X respectively.

The tower principle

Say that we are given a probability space $(\Omega, \mathcal{F}, P)$ and *two* sub-σ-algebras $\mathcal{B}$ and $\mathcal{B}_1$ of $\mathcal{F}$. If neither of $\mathcal{B}$ and $\mathcal{B}_1$ is contained in the other, the information

contained in them is in some sense incompatible. But if they are nested, we have the following result:

Proposition 4.23: The tower principle

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $X \in \mathcal{L}^1(P)$, and let $\mathcal{B}$ and $\mathcal{B}_1$ be sub-σ-algebras of $\mathcal{F}$. If $\mathcal{B}_1 \subseteq \mathcal{B}$, then

$$\mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{B}] \,\Big|\, \mathcal{B}_1\Big] = \mathbb{E}[X \mid \mathcal{B}_1] = \mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{B}_1] \,\Big|\, \mathcal{B}\Big] \quad P\text{-a.s.}$$

Proof. The first equality follows easily from the definition, and the second is a consequence of Proposition 4.22(ii). □

This result tells us that approximating X using the information in $\mathcal{B}$ and then afterwards approximating further using $\mathcal{B}_1$ is the same as just approximating X using $\mathcal{B}_1$ directly.

Another way to understand this result is in terms of projections. Recall that we for $p \in (0, \infty)$ denote by $L^p(P)$ the space of equivalence classes of random variables in $\mathcal{L}^p(P)$, where X and Y are equivalent if $X = Y$ P-a.s. In the case that $X, Y \in \mathcal{L}^1(P)$ it is easy to show that if $X = Y$ P-a.s., then also $\mathbb{E}[X \mid \mathcal{B}] = \mathbb{E}[Y \mid \mathcal{B}]$ P-a.s. for any sub-σ-algebra $\mathcal{B}$ of $\mathcal{F}$. Hence $\mathcal{B}$ induces a well-defined linear map

$$\begin{aligned} P_{\mathcal{B}} \colon L^1(P) &\to L^1(P), \\ X &\mapsto \mathbb{E}[X \mid \mathcal{B}]. \end{aligned}$$

Here we have suppressed the distinction between elements of $\mathcal{L}^1(P)$ and of $L^1(P)$. The content of Proposition 4.23 can then be phrased simply as

$$P_{\mathcal{B}_1} \circ P_{\mathcal{B}} = P_{\mathcal{B}_1} = P_{\mathcal{B}} \circ P_{\mathcal{B}_1},$$

making the comparison with projection operators very tempting.

In fact, we can make this analogy more explicit. Let $\mathcal{L}^p(\mathcal{B}, P)$ denote the subspace of $\mathcal{L}^p(P)$ consisting of $\mathcal{B}$-measurable functions. Let $P_{\mathcal{B}}$ be the restriction of P to $\mathcal{B}$. It is then not difficult to show that integrals of $\mathcal{B}$-measurable functions with respect to P and $P_{\mathcal{B}}$ coincide, and hence that $\mathcal{L}^p(\mathcal{B}, P) = \mathcal{L}^p(P_{\mathcal{B}})$. Taking equivalence classes we find that also[2] $L^p(\mathcal{B}, P) = L^p(P_{\mathcal{B}})$.

Proposition 4.24

The map $P_{\mathcal{B}}$ restricted to $L^2(P)$ is the orthogonal projection onto $L^2(\mathcal{B}, P)$.

[2] If $X, Y \in \mathcal{L}^p(\mathcal{B}, P)$ are two random variables, then the set on which they differ is $\mathcal{B}$-measurable. Hence two $\mathcal{B}$-measurable variables that belong to the same equivalence class differ on a $\mathcal{B}$-measurable set.

Proof. First notice that $L^2(\mathcal{B}, P) = L^2(P_{\mathcal{B}})$ is complete, so it is a closed subspace of $L^2(P)$. Now let $\mathsf{X} \in \mathcal{L}^2(P)$, and let $\mathcal{X}$ be the orthogonal projection of $[\mathsf{X}]$ onto $L^2(\mathcal{B}, P)$. If $\mathsf{U} \in \mathcal{X}$ then $[\mathsf{X}] - [\mathsf{U}] \in L^2(\mathcal{B}, P)^\perp$, so in particular $[\mathsf{X}] - [\mathsf{U}] \perp \mathbf{1}_B$ for all $B \in \mathcal{B}$. But this implies that

$$\int_B \mathsf{X} \, \mathrm{d}P = \int_B \mathsf{U} \, \mathrm{d}P,$$

so U is indeed a conditional expectation of X, i.e. $P_{\mathcal{B}}([\mathsf{X}]) = [\mathsf{U}]$. □

In this setting Proposition 4.23 becomes obvious. Furthermore, in makes our desire that $\mathbb{E}[\mathsf{X} \mid \mathcal{B}]$ be the *best* approximation of X that is $\mathcal{B}$-measurable precise, as long as we interpret 'best' to mean closest in the L^2-norm.

Independence

Recall that we in Subsection 4.3 defined what it means for a collection of σ-algebras to be independent. We now extend this definition to random variables.

Definition 4.25: Independence III

Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $(\mathcal{C}_i)_{i \in I}$ is a family of sets $\mathcal{C}_i$ of events and $(\mathsf{X}_j)_{j \in J}$ is a collection of random variables on $(\Omega, \mathcal{F}, P)$, then we say that the collection

$$\{\mathcal{C}_i \mid i \in I\} \cup \{\mathsf{X}_j \mid j \in J\}$$

is ***independent*** if the corresponding collection

$$\{\mathcal{C}_i \mid i \in I\} \cup \{\sigma(\mathsf{X}_j) \mid j \in J\}$$

is independent in the sense of Definition 4.5.

In particular, if X is a random variable and $\mathcal{B}$ a sub-σ-algebra of $\mathcal{F}$, then $\{\mathsf{X}, \mathcal{B}\}$ is independent if

$$P(\{\mathsf{X} \in A\} \cap B) = P(\mathsf{X} \in A)P(B)$$

for all $A \in \mathcal{B}(\mathbb{R})$ and $B \in \mathcal{B}$. The interpretation of this independence is quite natural: We need the information in $\sigma(\mathsf{X})$ to determine the value X assumes given some outcome. Say that this information is unknown to us. Then the best approximation of X that we can make is just $\mathbb{E}[\mathsf{X}]$. Does knowing the information in the σ-algebra $\mathcal{B}$ help us better approximate X? If $\sigma(\mathsf{X})$ and $\mathcal{B}$ are independent, then knowing whether the events in $\mathcal{B}$ have occurred does not change our knowledge the events in $\sigma(\mathsf{X})$, and hence does not enable us to make a better approximation. Hence we have have the following result:

Proposition 4.26

Let $X \in \mathcal{L}^1(P)$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{F}$. If X and $\mathcal{B}$ are independent, then

$$\mathbb{E}[X \mid \mathcal{B}] = \mathbb{E}[X].$$

Proof. Since X and $\mathcal{B}$ are independent, so are X and $\mathbf{1}_B$ for all $B \in \mathcal{B}$. It follows that

$$\int_B \mathbb{E}[X]\,\mathrm{d}P = \mathbb{E}[X]P(B) = \mathbb{E}[X]\mathbb{E}[\mathbf{1}_B] = \mathbb{E}[X\mathbf{1}_B] = \int_B X\,\mathrm{d}P,$$

proving the claim. □

This result in particular implies Proposition 4.22(i), since $\mathcal{B} = \{\emptyset, \Omega\}$ is independent of any random variable.

⁂ Bibliography

Bauer, Heinz (1995). *Probability Theory*. 1st ed. de Gruyter. 523 pp. ISBN: 3-11-013935-9.

— (2001). *Measure and Integration Theory*. 1st ed. de Gruyter. 230 pp. ISBN: 3-11-016719-0.

Billingsley, Patrick (1995). *Probability and Measure*. 3rd ed. Wiley. 593 pp. ISBN: 0-471-00710-2.

Cohn, Donald L. (2001). *Measure Theory*. 2nd ed. Birkhäuser. 457 pp. ISBN: 978-1-4614-6955-1. DOI: 10.1007/978-1-4614-6956-8.

Davey, B. A. and H. A. Priestley (2002). *Introduction to Lattices and Order*. 2nd ed. Cambridge University Press. 298 pp. ISBN: 978-0-521-78451-1.

Folland, Gerald B. (2007). *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. Wiley. 386 pp. ISBN: 0-471-31716-0.

Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability*. 2nd ed. Chelsea Publishing Company. 84 pp.

Kolmogorov, A. N. and Richard Jeffrey (1995). 'Complete Metric Boolean Algebras'. In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 77.1, pp. 57–66. ISSN: 00318116, 15730883. URL: http://www.jstor.org/stable/4320553.

Thorbjørnsen, Steen (2014). *Grundlæggende mål- og integralteori*. Aarhus Universitetsforlag. 425 pp. ISBN: 978-87-7124-508-0.

Vickers, Steven (1989). *Topology via Logic*. 1st ed. Cambridge University Press. 200 pp. ISBN: 0-521-36062-5.

Willard, Stephen (1970). *General Topology*. Addison-Wesley Publishing Company, Inc. 369 pp.