Chapter 1
Motivation

SECTION 1 offers some reasons for why anyone who uses probability should know about the measure theoretic approach.

SECTION 2 describes some of the added complications, and some of the compensating benefits that come with the rigorous treatment of probabilities as measures.

SECTION 3 argues that there are advantages in approaching the study of probability theory via expectations, interpreted as linear functionals, as the basic concept.

SECTION 4 describes the de Finetti convention of identifying a set with its indicator function, and of using the same symbol for a probability measure and its corresponding expectation.

*SECTION *5 presents a fair-price interpretation of probability, which emphasizes the linearity properties of expectations. The interpretation is sometimes a useful guide to intuition.*

1. Why bother with measure theory?

Following the appearance of the little book by Kolmogorov (1933), which set forth a measure theoretic foundation for probability theory, it has been widely accepted that probabilities should be studied as special sorts of measures. (More or less true—see the Notes to the Chapter.) Anyone who wants to understand modern probability theory will have to learn something about measures and integrals, but it takes surprisingly little to get started.

For a rigorous treatment of probability, the measure theoretic approach is a vast improvement over the arguments usually presented in undergraduate courses. Let me remind you of some difficulties with the typical introduction to probability.

Independence

There are various elementary definitions of independence for random variables. For example, one can require factorization of distribution functions,

$$\mathbb{P}\{X \le x, Y \le y\} = \mathbb{P}\{X \le x\}\,\mathbb{P}\{Y \le y\} \qquad \text{for all real } x, y.$$

The problem with this definition is that one needs to be able to calculate distribution functions, which can make it impossible to establish rigorously some desirable

properties of independence. For example, suppose $X_1, \ldots, X_4$ are independent random variables. How would you show that

$$Y = X_1 X_2 \left[\log \left(\frac{X_1^2 + X_2^2}{|X_1| + |X_2|} \right) + \frac{|X_1|^3 + X_2^3}{X_1^4 + X_2^4} \right]$$

is independent of

$$Z = \sin \left[X_3 + X_3^2 + X_3 X_4 + X_4^2 + \sqrt{X_3^4 + X_4^4} \right],$$

by means of distribution functions? Somehow you would need to express events $\{Y \le y, Z \le z\}$ in terms of the events $\{X_i \le x_i\}$, which is not an easy task. (If you did figure out how to do it, I could easily make up more taxing examples.)

You might also try to define independence via factorization of joint density functions, but I could invent further examples to make your life miserable, such as problems where the joint distribution of the random variables are not even given by densities. And if you could grind out the joint densities, probably by means of horrible calculations with Jacobians, you might end up with the mistaken impression that independence had something to do with the smoothness of the transformations.

The difficulty disappears in a measure theoretic treatment, as you will see in Chapter 4. Facts about independence correspond to facts about product measures.

Discrete versus continuous

Most introductory texts offer proofs of the Tchebychev inequality,

$$\mathbb{P}\{|X - \mu| \ge \epsilon\} \le \text{var}(X)/\epsilon^2,$$

where μ denotes the expected value of X. Many texts even offer two proofs, one for the discrete case and another for the continuous case. Indeed, introductory courses tend to split into at least two segments. First one establishes all manner of results for discrete random variables and then one reproves almost the same results for random variables with densities.

Unnecessary distinctions between discrete and continuous distributions disappear in a measure theoretic treatment, as you will see in Chapter 3.

Univariate versus multivariate

The unnecessary repetition does not stop with the discrete/continuous dichotomy. After one masters formulae for functions of a single random variable, the whole process starts over for several random variables. The univariate definitions acquire a prefix *joint*, leading to a whole host of new exercises in multivariate calculus: joint densities, Jacobians, multiple integrals, joint moment generating functions, and so on.

Again the distinctions largely disappear in a measure theoretic treatment. Distributions are just image measures; joint distributions are just image measures for maps into product spaces; the same definitions and theorems apply in both cases. One saves a huge amount of unnecessary repetition by recognizing the role of image

measures (described in Chapter 2) and recognizing joint distributions as measures on product spaces (described in Chapter 4).

Approximation of distributions

Roughly speaking, the central limit theorem asserts:

> *If $\xi_1, \ldots, \xi_n$ are independent random variables with zero expected values and variances summing to one, and if none of the ξ_i makes too large a contribution to their sum, then $\xi_1 + \ldots + \xi_n$ is approximately $N(0, 1)$ distributed.*

What exactly does that mean? How can something with a discrete distribution, such as a standardized Binomial, be approximated by a smooth normal distribution? The traditional answer (which is sometimes presented explicitly in introductory texts) involves pointwise convergence of distribution functions of random variables; but the central limit theorem is seldom established (even in introductory texts) by checking convergence of distribution functions. Instead, when proofs are given, they typically involve checking of pointwise convergence for some sort of generating function. The proof of the equivalence between convergence in distribution and pointwise convergence of generating functions is usually omitted. The treatment of convergence in distribution for random vectors is even murkier.

As you will see in Chapter 7, it is far cleaner to start from a definition involving convergence of expectations of "smooth functions" of the random variables, an approach that covers convergence in distribution for random variables, random vectors, and even random elements of metric spaces, all within a single framework.

***

In the long run the measure theoretic approach will save you much work and help you avoid wasted effort with unnecessary distinctions.

2. The cost and benefit of rigor

In traditional terminology, probabilities are numbers in the range $[0, 1]$ attached to events, that is, to subsets of a sample space Ω. They satisfy the rules

> *(i)* $\mathbb{P}\emptyset = 0$ *and* $\mathbb{P}\Omega = 1$
>
> *(ii) for disjoint events $A_1, A_2, \ldots$, the probability of their union, $\mathbb{P}(\cup_i A_i)$, is equal to $\sum_i \mathbb{P}A_i$, the sum of the probabilities of the individual events.*

When teaching introductory courses, I find that it pays to be a little vague about the meaning of the dots in (ii), explaining only that it lets us calculate the probability of an event by breaking it into disjoint pieces whose probabilities are summed. Probabilities add up in the same way as lengths, areas, volumes, and masses. The fact that we sometimes need a countable infinity of pieces (as in calculations involving potentially infinite sequences of coin tosses, for example) is best passed off as an obvious extension of the method for an arbitrarily large, finite number of pieces.

In fact the extension is not at all obvious, mathematically speaking. As explained by Hawkins (1979), the possibility of having the additivity property (ii)

hold for countable collections of disjoint events, a property known officially as ***countable additivity***, is one of the great discoveries of modern mathematics. In his 1902 doctoral dissertation, Henri Lebesgue invented a method for defining lengths of complicated subsets of the real line, in a countably additive way. The definition has the subtle feature that not every subset has a length. Indeed, under the usual axioms of set theory, it is impossible to extend the concept of length to *all* subsets of the real line while preserving countable additivity.

The same subtlety carries over to probability theory. In general, the collection of events to which countably additive probabilities are assigned cannot include all subsets of the sample space. The domain of the set function $\mathbb{P}$ (the ***probability measure***) is usually just a ***sigma-field***, a collection of subsets of Ω with properties that will be defined in Chapter 2.

Many probabilistic ideas are greatly simplified by reformulation as properties of sigma-fields. For example, the unhelpful multitude of possible definitions for independence coalesce nicely into a single concept of independence for sigma-fields.

The sigma-field limitation turns out to be less of a disadvantage than might be feared. In fact, it has positive advantages when we wish to prove some probabilistic fact about all events in some sigma-field, $\mathcal{A}$. The obvious line of attack—first find an explicit representation for the typical member of $\mathcal{A}$, then check the desired property directly—usually fails. Instead, as you will see in Chapter 2, an indirect approach often succeeds.

(a) Show directly that the desired property holds for all events in some subclass $\mathcal{E}$ of "simpler sets" from $\mathcal{A}$.

(b) Show that $\mathcal{A}$ is the smallest sigma-field for which $\mathcal{A} \supseteq \mathcal{E}$.

(c) Show that the desired property is preserved under various set theoretic operations. For example, it might be possible to show that if two events have the property then so does their union.

(d) Deduce from (c) that the collection $\mathcal{B}$ of all events with the property forms a sigma-field of subsets of Ω. That is, $\mathcal{B}$ is a sigma-field, which, by (a), has the property $\mathcal{B} \supseteq \mathcal{E}$.

(e) Conclude from (b) and (d) that $\mathcal{B} \supseteq \mathcal{A}$. That is, the property holds for all members of $\mathcal{A}$.

REMARK. Don't worry about the details for the moment. I include the outline in this Chapter just to give the flavor of a typical measure theoretic proof. I have found that some students have trouble adapting to this style of argument.

The indirect argument might seem complicated, but, with the help of a few key theorems, it actually becomes routine. In the literature, it is not unusual to see applications abbreviated to a remark like "a simple generating class argument shows ...," with the reader left to fill in the routine details.

Lebesgue applied his definition of length (now known as Lebesgue measure) to the construction of an integral, extending and improving on the Riemann integral. Subsequent generalizations of Lebesgue's concept of measure (as in the 1913 paper of Radon and other developments described in the Epilogue to

Hawkins 1979) eventually opened the way for Kolmogorov to identify probabilities with measures on sigma-fields of events on general sample spaces. From the Preface to Kolmogorov (1933), in the 1950 translation by Morrison:

> The purpose of this monograph is to give an axiomatic foundation for the theory of probability. The author set himself the task of putting in their natural place, among the general notions of modern mathematics, the basic concepts of probability theory—concepts which until recently were considered to be quite peculiar.
>
> This task would have been a rather hopeless one before the introduction of Lebesgue's theories of measure and integration. However, after Lebesgue's publication of his investigations, the analogies between measure of a set and probability of an event, and between integral of a function and mathematical expectation of a random variable, became apparent. These analogies allowed of further extensions; thus, for example, various properties of independent random variables were seen to be in complete analogy with the corresponding properties of orthogonal functions. But if probability theory was to be based on the above analogies, it still was necessary to make the theories of measure and integration independent of the geometric elements which were in the foreground with Lebesgue. This has been done by Fréchet.
>
> While a conception of probability theory based on the above general viewpoints has been current for some time among certain mathematicians, there was lacking a complete exposition of the whole system, free of extraneous complications. (Cf., however, the book by Fréchet ...)

Kolmogorov identified random variables with a class of real-valued functions (the ***measurable functions***) possessing properties allowing them to coexist comfortably with the sigma-field. Thereby he was also able to identify the expectation operation as a special case of integration with respect to a measure. For the newly restricted class of random variables, in addition to the traditional properties

(i) $\mathbb{E}(c_1X_1 + c_2X_2) = c_1\mathbb{E}(X_1) + c_2\mathbb{E}(X_2)$, for constants c_1 and c_2,

(ii) $\mathbb{E}(X) \geq \mathbb{E}(Y)$ if $X \geq Y$,

he could benefit from further properties implied by the countable additivity of the probability measure.

As with the sigma-field requirement for events, the measurability restriction on the random variables came with benefits. In modern terminology, no longer was $\mathbb{E}$ just an ***increasing linear functional*** on the space of real random variables (with some restrictions to avoid problems with infinities), but also it had acquired some continuity properties, making possible a rigorous treatment of limiting operations in probability theory.

3. Where to start: probabilities or expectations?

From the example set by Lebesgue and Kolmogorov, it would seem natural to start with probabilities of events, then extend, via the operation of integration, to the study of expectations of random variables. Indeed, in many parts of the mathematical world that is the way it goes: probabilities are the basic quantities, from which expectations of random variables are derived by various approximation arguments.

The apparently natural approach is by no means the only possibility, as anyone brought up on the works of the fictitious French author Bourbaki could affirm. (The treatment of measure theory, culminating with Bourbaki 1969, started from integrals defined as linear functionals on appropriate spaces of functions.) Moreover, historically speaking, expectation has a strong claim to being the preferred starting point for a theory of probability. For instance, in his discussion of the 1657 book *Calculating in Games of Chance* by Christian Huygens, Hacking (1978, page 97) commented:

> The fair prices worked out by Huygens are just what we would call the expectations of the corresponding gambles. His approach made expectation a more basic concept than probability, and this remained so for about a century.

The fair price interpretation is sketched in Section 5.

The measure theoretic history of integrals as linear functionals also extends back to the early years of the twentieth century, starting with Daniell (1918), who developed a general theory of integration via extension of linear functionals from small spaces of functions to larger spaces. It is also significant that, in one of the greatest triumphs of measure theory, Wiener (1923, Section 10) defined what is now known as Wiener measure (thereby providing a rigorous basis for the mathematical theory of Brownian motion) as an averaging operation for functionals defined on Brownian motion paths, citing Daniell (1919) for the basic extension theorem.

There are even better reasons than historical precedent for working with expectations as the basic concept. Whittle (1992), in the *Preface* to an elegant, intermediate level treatment of *Probability via Expectations*, presented some arguments:

(i) To begin with, people probably have a better intuition for what is meant by an 'average value' than for what is meant by a 'probability.'

(ii) Certain important topics, such as optimization and approximation problems, can be introduced and treated very quickly, just because they are phrased in terms of expectations.

(iii) Most elementary treatments are bedeviled by the apparent need to ring the changes of a particular proof or discussion for all the special cases of continuous or discrete distribution, scalar or vector variables, etc. In the expectations approach these are indeed seen as special cases, which can be treated with uniformity and economy.

His list continued. I would add that:

(a) It is often easier to work with the linearity properties of integrals than with the additivity properties of measures. For example, many useful probability inequalities are but thinly disguised consequences of pointwise inequalities, translated into probability form by the linearity and increasing properties of expectations.

(b) The linear functional approach, via expectations, can save needless repetition of arguments. Some theorems about probability measures, as set functions, are just special cases of more general results about expectations.

(c) When constructing new probability measures, we save work by defining the integral of measurable functions directly, rather than passing through the preliminary step of building the set function then establishing theorems about the corresponding integrals. As you will see repeatedly, definitions and theorems sometimes collapse into a single operation when expressed directly in terms of expectations, or integrals.

***

I will explain the essentials of measure theory in Chapter 2, starting from the traditional set-function approach but working as quickly as I can towards systematic use of expectations.

4. The de Finetti notation

The advantages of treating expectation as the basic concept are accentuated by the use of an elegant notation strongly advocated by de Finetti (1972, 1974). Knowing that many traditionally trained probabilists and statisticians find the notation shocking, I will introduce it slowly, in an effort to explain why it is worth at least a consideration. (Immediate enthusiastic acceptance is more than I could hope for.)

Ordinary algebra is easier than Boolean algebra. The correspondence $A \leftrightarrow \mathbb{I}_A$ between subsets A of a fixed set $\mathcal{X}$ and their indicator functions,

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \in A^c, \end{cases}$$

transforms Boolean algebra into ordinary pointwise algebra with functions. I claim that probability theory becomes easier if one works systematically with expectations of indicator functions, $\mathbb{E}\mathbb{I}_A$, rather than with the corresponding probabilities of events.

Let me start with the assertions about algebra and Boolean algebra. The operations of union and intersection correspond to pointwise maxima (denoted by max or the symbol $\vee$) and pointwise minima (denoted by min or the symbol $\wedge$), or pointwise products:

$$\mathbb{I}_{\cup_i A_i}(x) = \bigvee_i \mathbb{I}_{A_i}(x) \qquad \text{and} \qquad \mathbb{I}_{\cap_i A_i}(x) = \bigwedge_i \mathbb{I}_{A_i}(x) = \prod_i \mathbb{I}_{A_i}(x).$$

Complements correspond to subtraction from one: $\mathbb{I}_{A^c}(x) = 1 - \mathbb{I}_A(x)$. Derived operations, such as the set theoretic difference $A \backslash B := A \cap B^c$ and the symmetric difference, $A \Delta B := (A \backslash B) \cup (B \backslash A)$, also have simple algebraic counterparts:

$$\mathbb{I}_{A \backslash B}(x) = (\mathbb{I}_A(x) - \mathbb{I}_B(x))^+ := \max\left(0, \mathbb{I}_A(x) - \mathbb{I}_B(x)\right),$$
$$\mathbb{I}_{A \Delta B}(x) = |\mathbb{I}_A(x) - \mathbb{I}_B(x)|.$$

To check these identities, just note that the functions take only the values 0 and 1, then determine which combinations of indicator values give a 1. For example, $|\mathbb{I}_A(x) - \mathbb{I}_B(x)|$ takes the value 1 when exactly one of $\mathbb{I}_A(x)$ and $\mathbb{I}_B(x)$ equals 1.

The algebra looks a little cleaner if we omit the argument x. For example, the horrendous set theoretic relationship

$$\left(\cap_{i=1}^n A_i\right) \Delta \left(\cap_{i=1}^n B_i\right) \subseteq \cup_{i=1}^n \left(A_i \Delta B_i\right)$$

corresponds to the pointwise inequality

$$\left|\prod_i \mathbb{I}_{A_i} - \prod_i \mathbb{I}_{B_i}\right| \le \max_i \left|\mathbb{I}_{A_i} - \mathbb{I}_{B_i}\right|,$$

whose verification is easy: when the right-hand side takes the value 1 the inequality is trivial, because the left-hand side can take only the values 0 or 1; and when right-hand side takes the value 0, we have $\mathbb{I}_{A_i} = \mathbb{I}_{B_i}$ for all i, which makes the left-hand side zero.

<1> **Example.** One could establish an identity such as

$$(A\Delta B)\Delta(C\Delta D) = A\Delta\left(B\Delta(C\Delta D)\right)$$

by expanding both sides into a union of many terms. It is easier to note the pattern for indicator functions. The set $A\Delta B$ is the region where $\mathbb{I}_A + \mathbb{I}_B$ takes an odd value (that is, the value 1); and $(A\Delta B)\Delta C$ is the region where $(\mathbb{I}_A + \mathbb{I}_B) + \mathbb{I}_C$ takes an odd value. And so on. In fact both sides of the set theoretic identity equal the region where $\mathbb{I}_A + \mathbb{I}_B + \mathbb{I}_C + \mathbb{I}_D$ takes an odd value. Associativity of set theoretic differences is a consequence of associativity of pointwise addition. □

<2> **Example.** The lim sup of a sequence of sets $\{A_n : n \in \mathbb{N}\}$ is defined as

$$\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{i \ge n} A_i.$$

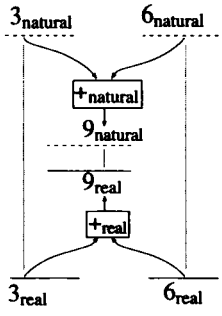
That is, the lim sup consists of those x for which, to each n there exists an $i \ge n$ such that $x \in A_i$. Equivalently, it consists of those x for which $x \in A_i$ for infinitely many i. In other words,

$$\mathbb{I}_{\limsup_n A_n} = \limsup_n \mathbb{I}_{A_n}.$$

Do you really need to learn the new concept of the lim sup of a sequence of sets? Theorems that work for lim sups of sequences of functions automatically carry over to theorems about sets. There is no need to prove everything twice. The correspondence between sets and their indicators saves us from unnecessary work. □

After some repetition, it becomes tiresome to have to keep writing the $\mathbb{I}$ for the indicator function. It would be much easier to write something like $\tilde{A}$ in place of $\mathbb{I}_A$. The indicator of the lim sup of a sequence of sets would then be written $\limsup_n \tilde{A}_n$, with only the tilde to remind us that we are referring to functions. But why do we need reminding? As the example showed, the concept for the lim sup of sets is really just a special case of the concept for sequences of functions. Why preserve a distinction that hardly matters?

There is a well established tradition in Mathematics for choosing notation that eliminates inessential distinctions. For example, we use the same symbol 3 for the natural number and the real number, writing $3 + 6 = 9$ as an assertion both about addition of natural numbers and about addition of real numbers.



It does not matter if we cannot tell immediately which interpretation is intended, because we know there is a one-to-one correspondence between natural numbers and a subset of the real numbers, which preserves all the properties of interest. Formally, there is a map $\psi : \mathbb{N} \to \mathbb{R}$ for which

$$\psi(x +_{natural} y) = \psi(x) +_{real} \psi(y) \qquad \text{for all } x, y \text{ in } \mathbb{N},$$

with analogous equalities for other operations. (Notice that I even took care to distinguish between addition as a function from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ and as a function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$.) The map ψ is an isomorphism between $\mathbb{N}$ and a subset of $\mathbb{R}$.

REMARK. Of course there are some situations where we need to distinguish between a natural number and its real counterpart. For example, it would be highly confusing to use indistinguishable symbols when first developing the properties of the real number system from the properties of the natural numbers. Also, some computer languages get very upset when a function that expects a floating point argument is fed an integer variable; some languages even insist on an explicit conversion between types.

We are faced with a similar overabundance of notation in the correspondence between sets and their indicator functions. Formally, and traditionally, we have a map $A \mapsto \mathbb{I}_A$ from sets into a subset of the nonnegative real functions. The map preserves the important operations. It is firmly in the Mathematical tradition that we should follow de Finetti's suggestion and **use the same symbol for a set and its indicator function**.

REMARK. A very similar convention has been advocated by the renowned computer scientist, Donald Knuth, in an expository article (Knuth 1992). He attributed the idea to Kenneth Iversen, the inventor of the programming language APL.

In de Finetti's notation the assertion from Example <2> becomes

$$\limsup A_n = \limsup A_n,$$

a fact that is quite easy to remember. The theorem about lim sups of sequences of sets has become incorporated into the notation; we have one less theorem to remember.

The second piece of de Finetti notation is suggested by the same logic that encourages us to replace $+_{natural}$ and $+_{real}$ by the single addition symbol: use the same symbol when extending the domain of definition of a function. For example, the symbol "sin" denotes both the function defined on the real line and its extension to the complex domain. More generally, if we have a function g with domain G_0, which can be identified with a subset $\widetilde{G}_0$ of some $\widetilde{G}$ via a correspondence $x \leftrightarrow \tilde{x}$, and if $\tilde{g}$ is a function on $\widetilde{G}$ for which $\tilde{g}(\tilde{x}) = g(x)$ for x in G_0, then why not write g instead of $\tilde{g}$ for the function with the larger domain?

With probability theory we often use $\mathbb{P}$ to denote a probability measure, as a map from a class $\mathcal{A}$ (a sigma-field) of subsets of some Ω into the subinterval $[0, 1]$ of the real line. The correspondence $A \leftrightarrow \tilde{A} := \mathbb{I}_A$, between a set A and its indicator function $\tilde{A}$, establishes a correspondence between $\mathcal{A}$ and a subset of the collection of

random variables on Ω. The expectation maps random variables into real numbers, in such a way that $\mathbb{E}(\tilde{A}) = \mathbb{P}(A)$. This line of thinking leads us to de Finetti's second suggestion: **use the same symbol for expectation and probability measure**, writing $\mathbb{P}X$ instead of $\mathbb{E}X$, and so on.

The de Finetti notation has an immediate advantage when we deal with several probability measures, $\mathbb{P}$, $\mathbb{Q}$, ... simultaneously. Instead of having to invent new symbols $\mathbb{E}_{\mathbb{P}}$, $\mathbb{E}_{\mathbb{Q}}$, ..., we reuse $\mathbb{P}$ for the expectation corresponding to $\mathbb{P}$, and so on.

REMARK. You might have the concern that you will not be able to tell whether $\mathbb{P}A$ refers to the probability of an event or the expected value of the corresponding indicator function. The ambiguity should not matter. Both interpretations give the same number; you will never be faced with a choice between two different values when choosing an interpretation. If this ambivalence worries you, I would suggest going systematically with the expectation/indicator function interpretation. It will never lead you astray.

<3> **Example.** For a finite collection of events $A_1, \ldots, A_n$, the so-called ***method of inclusion and exclusion*** asserts that the probability of the union $\cup_{i\le n} A_i$ equals

$$\sum_i \mathbb{P}A_i - \sum_{i\neq j} \mathbb{P}(A_i \cap A_j) + \sum_{i,j,k} \{i, j, k \text{ distinct}\}\mathbb{P}(A_i \cap A_j \cap A_k) - \ldots \pm \mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n).$$

The equality comes by taking expectations on both sides of an identity for (indicator) functions,

$$\cup_{i\le n} A_i = \sum_i A_i - \sum_{i\neq j} A_i A_j + \sum_{i,j,k} \{i, j, k \text{ distinct}\} A_i A_j A_k - \ldots \pm A_1 A_2 \ldots A_n.$$

The right-hand side of this identity is just the expanded version of $1 - \prod_{i\le n}(1 - A_i)$. The identity is equivalent to

$$1 - \cup_{i\le n} A_i = \prod_{i\le n} (1 - A_i),$$

which presents two ways of expressing the indicator function of $\cap_{i\le n} A_i^c$. See Problem [1] for a generalization. □

<4> **Example.** Consider Tchebychev's inequality, $\mathbb{P}\{|X - \mu| \ge \epsilon\} \le \text{var}(X)/\epsilon^2$, for each $\epsilon > 0$, and each random variable X with expected value $\mu := \mathbb{P}X$ and finite variance, $\text{var}(X) := \mathbb{P}(X - \mu)^2$. On the left-hand side of the inequality we have the probability of an event. Or is it the expectation of an indicator function? Either interpretation is correct, but the second is more helpful. The inequality is a consequence of the increasing property for expectations invoked for a pair of functions, $\{|X - \mu| \ge \epsilon\} \le (X - \mu)^2/\epsilon^2$. The indicator function on the left-hand side takes only the values 0 and 1. The quadratic function on the right-hand side is nonnegative, and is ≥ 1 whenever the left-hand side equals 1. □

***

For the remainder of the book, I will be using the same symbol for a set and its indicator function, and writing $\mathbb{P}$ instead of $\mathbb{E}$ for expectation.

REMARK. For me, the most compelling reason to adopt the de Finetti notation, and work with $\mathbb{P}$ as a linear functional defined for random variables, was not that I would save on symbols, nor any of the other good reasons listed at the end of