Jonathan Rougier

Department of Mathematics
University of Bristol

APTS lecture notes on Statistical Inference

Cambridge, December 2013

1
Expectation and Probability Theory

The purpose of this chapter is to establish my notation, and to derive those results in probability theory that are most useful in statistical inference: the Law of Iterated Expectation, the Law of Total Probability, Bayes's Theorem, and so on. I have not covered independence and conditional independence. These are crucial for statistical modelling, but less so for inference, and they will be introduced in the chapters where they are needed.

What is a bit different about this chapter is that I have developed these results taking expectation, rather than probability, as primitive. Bruno de Finetti is my inspiration for this, notably de Finetti (1937, 1972, 1974/75) and the more recent books by Lad (1996) and Goldstein and Wooff (2007). Whittle (2000) is my source for many details, although my approach is quite different from his. Grimmett and Stirzaker (2001) is a standard orthodox text on probability theory. Bernardo and Smith (1994) is a standard Bayesian text on probability theory and statistics.

Why expectation as primitive? This is not the modern approach, where the starting point is a set, a sigma algebra on the set, and a non-negative normalised countably additive (probability) measure; see, for example, Billingsley (1995) or Williams (1991). This modern approach provides a formal basis for other less technical theories, such as ours, in the sense that if the two were found to be in conflict, then that would be alarming.

However, in the modern approach an uncertain quantity is a derived concept, and its expectation doubly so. But a statistician's objective is to reason sensibly in an uncertain world. For such a person (and I am one) the natural starting point is uncertain quantities, and the judgements one has about them. Thus uncertain quantities and their expectations are taken as primitive, and probability is defined in terms of expectation. There are no explicit sigma algebras and there is no measure theory, although the consistency of the two approaches indicates that they are implicitly present. However, they are not needed for day-to-day statistical inference.

The most daunting material in this introductory chapter comes in the section on conditional expectation (Section 1.6). This is because conditioning is a complicated operation, no matter what one's starting point. Most statistics textbooks fudge the issue of

conditioning on 'continuous' quantities. But I think this leads to confusion, and so I have presented a complete theory of conditioning, which, although a bit more complicated, is entirely consistent.

Some sections are starred—these can be skipped without loss of continuity.

1.0 *Conventions and notation*

This section must be read carefully, in order to understand the notation in the rest of this chapter.

A *proposition* is a statement which is either true or false. Thus, 'the moon is made of cheese' and '$x \leq 3$' are both propositions; the truth of the latter is contingent on the value for x. When a proposition p occurs in mathematical text, it must be read as 'it is true that p'. For example, 'Since $x \leq 3 \ldots$' must be read as 'Since it is true that $x \leq 3 \ldots$', and 'If $x \leq 3$, then $\ldots$' must be read as 'If it is true that $x \leq 3$, then $\ldots$'.

Unfortunately, there is ambiguity in the use of the symbol '$=$', which is used for both propositions and assignments. I will treat it as propositional, so that '$x = 3$' is either true or false. I will use '$:=$' to indicate the assignment of the righthand side to the label on the lefthand side, as in $f(x) := a + bx$. After assignment, $f(x) = a + bx$, interpreted as a proposition (i.e. 'it is true that $f(x) = a + bx$'). This important distinction is recognised in computing, for which the propositional use of '$=$' is represented by `.EQ.` or `==`, to give examples from FORTRAN and C. In computing, `=` usually indicates assignment.

R. In the statistical computing environment R (R Development Core Team, 2011), we have `==` for propositions, `=` for assignment of functional arguments, and `<-` or `=` for assignment. In the last case I prefer the former but logically they are equivalent, since `foo = 3` in a functional argument represents a promise to evaluate `foo <- 3` in the body of the function before `foo` is first used ('lazy evaluation' of arguments builds on this).

Occasionally I will want to restrict the value of a quantity. For example, if $x \leq 3$ then I might want to consider the particular case when $x = 2$. In this case I write $x \leftarrow 2$. Think of this as a 'local' assignment.

Brackets. I avoid using the same bracketing symbol contiguously. Where nesting of brackets is required I tend to use the ordering $[\{(\cdot)\}]$ with the following exceptions:

1. Parentheses (round brackets) are always used around propositions (Section 1.4);
2. I have a preference for parentheses around functional arguments; less so for operators such as E and Pr.
3. Sets are always denoted with $\{\ldots\}$, and ordered tuples (usually points in a subset of Euclidean space) with $(\ldots)$.

4. Intervals of the real numbers $\mathbb{R}$ are denoted using square brackets or parentheses, depending on whether the endpoints are closed or open.

Typeface conventions. Random quantities are denoted with capital roman letters, while specified arguments and constants are denoted with small roman letters. Collections of random quantities are denoted with bold letters[1] where it is necessary to emphasise that they are not scalar; otherwise they are plain. Sets are denoted with curly capital roman letters ('caligraphic' letters), plus the usual notation for the set of real numbers ($\mathbb{R}$). Statistical parameters (which do not occur until later chapters) are denoted with small greek letters. Operators and special functions are usually denoted with sans-serif roman letters; other functions are denoted with small roman letters. Expectation has no less than three symbols: $\mathcal{E}$, $\mathbb{E}$, and E.

[1] Or an underscore in handwritten material.

Definitions and equivalences. I use 'exactly when' to state definitions, and 'if and only if' to state equivalences in theorems. Proofs of equivalences such as 'A if and only if B' typically have an A-if-B branch ($\Leftarrow$) and an A-only-if-B branch ($\Rightarrow$).

1.1 Random quantities

My starting-point is a *random quantity*. For me, a random quantity is a set of instructions which, if followed, will yield a real value; this is an *operational definition*. Real-valued functions of random quantities are also random quantities.

random quantity

operational definition

It is conventional in statistics to represent random quantities using capital letters from the end of the alphabet, such as X, Y, and Z, and, where more quantities are required, using ornaments such as subscripts and primes (e.g. X_i, Y'). Representative values of random quantities are denoted with small letters. Thus $X = x$ states 'it is true that the operation X was performed and the value x was the result'.

The *realm* of a random quantity is the set of possible values it might take. I denote this with a curly capital letter, such as $\mathcal{X}$ for the realm of X, where $\mathcal{X}$ is always a subset of $\mathbb{R}$.[2] If the realm of a random quantity X is finite or countable, X is said to be a *discrete random quantity*, otherwise it is said to be a *continuous random quantity*. I tend not to use these terms because, conceptually, there is a larger difference between a finite and a countable realm than there is between a countable and a non-countable realm (Section 1.8.1).

realm

[2] I have taken the word 'realm' from Lad (1996); 'range' is also used.

discrete random quantity

continuous random quantity

A random quantity in which the realm contains only a single element is a *constant*, and typically denotes by a small letter from the start of the alphabet, such as a, b, or c.

constant

Below it will be necessary to make assertions about the realm of X and about the joint realm of X and Y. I introduce the following

notation for this. Let B be any binary relation, for example '$\leq$'. Then I write

$$\models\{X\ B\ Y\}$$

exactly when $x\ B\ y$ for every (x, y) in the joint realm of X and Y. So, for example, $\models\{X = 1\}$ asserts that the realm of X is $\{1\}$, and $\models\{X \leq Y\}$ asserts that X will never exceed Y. These assertions reflect the operational definitions represented by X and Y. Where there is no ambiguity, statements such as "$\models\{X\ B\ Y\}$ and $\models\{Y\ B\ Z\}$" will be chained together as $\models\{X\ B\ Y\ B\ Z\}$.

It is important to be clear that $\models\{X \leq Y\}$ and $X \leq Y$ are quite different. The first is an assertion about the joint realm of X and Y, while the second is a proposition which may be true or false.

1.2 *Expectations*

With each random quantity I associate a real scalar *expectation*; the expectation of X is denoted $\mathsf{E}(X)$. Expectations are not arbitrary, but are required to satisfy the following axioms.

expectation

Definition 1.1 (Axioms of expectation)**.**

1. *If* $\models\{X = 1\}$ *then* $\mathsf{E}(X) = 1$ *(normalisation),*
2. *If* $\models\{X > 0\}$ *then* $\mathsf{E}(X) > 0$ *(positivity),*
3. $\mathsf{E}(X + Y) = \mathsf{E}(X) + \mathsf{E}(Y)$ *(additivity).*

The simplest interpretation of expectation is that of a 'best guess'. Then it follows that these axioms are justified as being self-evident. For example, if X was the weight in ounces of a one-ounce weight, then I would be foolish indeed not to assert $\mathsf{E}(X) = 1$. Likewise, if X was the weight in ounces of the orange I am holding in my hand, then I would be foolish indeed not to assert $\mathsf{E}(X) > 0$. Likewise, if Y was the weight of a second orange, then I would be foolish indeed not to assert that $\mathsf{E}(X + Y) = \mathsf{E}(X) + \mathsf{E}(Y)$.

'Judgement'. We tend not to use 'best guess' in practice: the word 'guess' has negative connotations. Instead, the word *judgement* is used. Thus expectations represent my judgements about random quantities. The use of 'judgement' captures the essentially subjective nature of expectation. Expectations do not have any objective existence: they are a property of the mind, and will vary from one person to another. Any given person, however, would want their collection of expectations to satisfy the axioms, insofar as these axioms are self-evident. Thus, were we to point out to a person that his collection of expectations violated one of the axioms, we would expect him to thank us, and to adjust his expectations accordingly.

judgement

The axioms in Definition 1.1 have some immediate implications, which are also self-evident (or almost so). I will just pick out a few of the really useful ones.

Theorem 1.1 (Immediate implications)**.**

1. $\mathsf{E}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathsf{E}(X_i)$ *(finite additivity);*

2. *For any constant a, $\mathsf{E}(aX) = a\,\mathsf{E}(X)$ and $\mathsf{E}(a) = a$ (linearity);*
3. *If $\vDash\{X \geq 0\}$ then $\mathsf{E}(X) \geq 0$ (non-negativity);*
4. *If $\vDash\{X \geq Y\}$ then $\mathsf{E}(X) \geq \mathsf{E}(Y)$ (monotonicity);*
5. $\inf\{\mathcal{X}\} \leq \mathsf{E}(X) \leq \sup\{\mathcal{X}\}$ *(convexity);*
6. $|\mathsf{E}(X)| \leq \mathsf{E}(|X|)$ *(triangle inequality).*

For convenience, it is helpful to lump additivity and linearity together into

$$\mathsf{E}(aX + bY) = a\,\mathsf{E}(X) + b\,\mathsf{E}(Y)$$

which I will term 'linearity'. This result can be iterated to apply to any finite sum.

Proof.

1. Follows by iterating additivity, starting with $X := X_1$ and $Y := X_2 + \cdots + X_n$.
2. Let i and j be integers. Then $\mathsf{E}(iX) = i\,\mathsf{E}(X)$ by finite additivity. But then $\mathsf{E}(X) = \mathsf{E}(jX/j) = j\,\mathsf{E}(X/j)$, and hence $\mathsf{E}(X/j) = \mathsf{E}(X)/j$. So $\mathsf{E}\{(i/j)X\} = (i/j)\,\mathsf{E}(X)$. The result then follows by passing from the rationals to the reals.[3] This result and normalisation imply $\mathsf{E}(a) = \mathsf{E}(a1) = a\,\mathsf{E}(1) = a$ where a is any constant.
3. Let $\vDash\{X \geq 0\}$ and define $Y := X - \mathsf{E}(X)$. Suppose that $\mathsf{E}(X) < 0$. Then $\vDash\{Y > 0\}$ and $\mathsf{E}(Y) > 0$ by positivity. But, by additivity and linearity, $\mathsf{E}(Y) = \mathsf{E}(X) - \mathsf{E}(X) = 0$, a contradiction. Hence $\mathsf{E}(X) \geq 0$.
4. By non-negativity and linearity, since $\vDash\{X - Y \geq 0\}$.
5. By monotonicity and linearity, because $\vDash\{\inf\{\mathcal{X}\} \leq X \leq \sup\{\mathcal{X}\}\}$.
6. By monotonicity and linearity, because $\vDash\{-|X| \leq X \leq |X|\}$.

□

[3] Slightly subtle, see de Finetti (1974, footnote on p. 75).

A less immediate implication is *Schwarz's inequality*, which is extremely important and used several times below (often to prove things that are almost obvious).

Schwarz's inequality

Theorem 1.2 (Schwarz's inequality)**.**

$$\{\mathsf{E}(XY)\}^2 \leq \mathsf{E}(X^2)\,\mathsf{E}(Y^2).$$

Proof. For any constant a, $\mathsf{E}\{(aX + Y)^2\} \geq 0$, by non-negativity. Expanding out the square and using linearity,

$$\mathsf{E}\{(aX + Y)^2\} = a^2\,\mathsf{E}(X^2) + 2a\,\mathsf{E}(XY) + \mathsf{E}(Y^2).$$

This quadratic in a cannot have two distinct real roots, because that would violate non-negativity. Then it follows from the standard formula for the roots of a quadratic[4] that

[4] If $ax^2 + bx + c = 0$ then

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

$$\{2\,\mathsf{E}(XY)\}^2 - 4\,\mathsf{E}(X^2)\,\mathsf{E}(Y^2) \leq 0,$$

or $\{\mathsf{E}(XY)\}^2 \leq \mathsf{E}(X^2)\,\mathsf{E}(Y^2)$, as required. □

Note that there is another version of Schwarz's inequality, which comes from setting $X \leftarrow |X|$ and $Y \leftarrow |Y|$,

$$\{\mathsf{E}(|XY|)\}^2 \leq \mathsf{E}(X^2)\,\mathsf{E}(Y^2).$$

This is stronger because the triangle inequality implies that

$$\{\mathsf{E}(XY)\}^2 = |\mathsf{E}(XY)|^2 \leq \{\mathsf{E}(|XY|)\}^2.$$

1.3* *Do all random quantities have finite expectations?*

Operationally-defined random quantities always have finite and bounded realms and, from this point of view, there is no obligation to develop a theory of reasoning about uncertainty for the more general cases.[5] This is an important issue, because theories of reasoning with non-finite and unbounded realms are a lot more complicated. Debabrata Basu summarises a viewpoint held by many statisticians.

[5] A set is bounded if the distance between any two elements is never greater than some finite amount.

> "The author holds firmly to the view that this contingent and cognitive universe of ours is in reality only finite and, therefore, discrete. In this essay we steer clear of the logical quick sands of 'infinity' and the 'infinitesimal'. Infinite and continuous models will be used in the sequel, but they are to be looked upon as mere approximations to the finite realities." (Basu, 1975, footnote, p. 4)

Lad (1996) is developed entirely in terms of finite and bounded realms.

However, as Kadane (2011, ch. 3) discusses, it is convenient to be able to work with non-finite and unbounded realms, to avoid the need to make an explicit truncation. Likewise, it is convenient to work with infinite sequences rather than long but finite sequences: the realm of a countably infinite sum of random quantities with finite and bounded realms is uncountable and unbounded.[6] Finally, for the purposes of statistical modelling we often introduce auxiliary random quantities (statistical parameters) and these are conveniently represented with non-finite and unbounded realms.

[6] Think of the Central Limit Theorem.

Therefore I will outline a more general treatment, which does not insist that random quantities have bounded realms.[7] The issue of non-finite realms is discussed in Section 1.8.1.

[7] This material draws heavily on Billingsley (1995, ch. 3).

Let X be a random quantity with possibly unbounded realm. Define

$$X^+ := \begin{cases} 0 & X \leq 0 \\ X & X > 0 \end{cases} \quad \text{and} \quad X^- := \begin{cases} -X & X \leq 0 \\ 0 & X > 0 \end{cases}$$

which are both non-negative quantities, and for which $X = X^+ - X^-$. Then redefine the expectation as

$$\mathsf{E}(X) := \mathsf{E}(X^+) - \mathsf{E}(X^-).$$

This expectation should respect these self-evident rules:

$$\mathsf{E}(X) = \begin{cases} \text{finite} & \text{both } \mathsf{E}(X^+) \text{ and } \mathsf{E}(X^-) \text{ finite} \\ \infty & \mathsf{E}(X^+) = \infty \text{ and } \mathsf{E}(X^-) \text{ finite} \\ -\infty & \mathsf{E}(X^+) \text{ finite and } \mathsf{E}(X^-) = \infty \\ \text{undefined} & \text{both } \mathsf{E}(X^+) \text{ and } \mathsf{E}(X^-) \text{ infinite.} \end{cases}$$

Then a weakening of the additivity axiom in Definition 1.1 gives

3′. If X and Y are non-negative, then $\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y)$.

This includes $\mathsf{E}(X+Y) = \infty$ if either $\mathsf{E}(X) = \infty$ or $\mathsf{E}(Y) = \infty$. If $\mathsf{E}(X)$ is finite, X is said to be *integrable*. But because $|X| = X^+ + X^-$, where both terms are non-negative,

integrable

$$X \text{ is integrable} \iff \mathsf{E}(|X|) \text{ is finite.}$$

We then have, as an immediate implication, that if X and Y are both integrable, then $X+Y$ is integrable, and $\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y)$.

Proof. The integrability of $X+Y$ follows from

$$\mathsf{E}(|X+Y|) \leq \mathsf{E}(|X| + |Y|) = \mathsf{E}(|X|) + \mathsf{E}(|Y|) < \infty$$

by monotonicity and the non-negativity of $|X|$ and $|Y|$. Now let $Z := X+Y$. Then

$$Z^+ - Z^- = X + Y = X^+ - X^- + Y^+ - Y^-.$$

Rearrange this to give

$$Z^+ + X^- + Y^- = Z^- + X^+ + Y^+,$$

where all of the terms in each sum are non-negative. Taking expectations of both sides and rearranging (using new Axiom 3′) shows that

$$\mathsf{E}(Z^+) - \mathsf{E}(Z^-) = \mathsf{E}(X^+) - \mathsf{E}(X^-) + \mathsf{E}(Y^+) - \mathsf{E}(Y^-)$$

or $\mathsf{E}(Z) = \mathsf{E}(X) + \mathsf{E}(Y)$, completing the proof. □

Therefore, this generalisation includes the original Additivity axiom of Definition 1.1 as a special case which holds not just when all realms are bounded, but for any integrable random quantities. Reassuringly, it shows that the extension to random quantities with unbounded realms does not cause any real difficulties, and that infinities can be accommodated. However, I will let Bruno de Finetti have the last word:

> "… the unbounded X is a theoretical schematization substituted for simplicity in place of an actual X, which is in reality bounded, but whose bounds are very large and imprecisely known." (de Finetti, 1974, p. 132)

1.4 Probability

In the approach I am adopting, probability is defined in terms of expectation. Consider any proposition A, such as $X > a$, which is either false or true; the use of a capital letter indicates that its status is uncertain to me. I follow de Finetti (1974, chapters 1 and 2) in identifying false with zero and true with one, where propositions occur in mathematical expressions.[8] Often it is necessary to use parentheses to delimit a proposition in mathematical expressions, because many propositional relations have lower priority than other relations. For example, sums and integrals over restrictions of their domain can be represented as

$$\sum_{j \in J} a_j = \sum_j a_j \, (j \in J).$$

[8] This is also the convention in programming languages such as R.

As I identify propositions with their indicator functions, $A := (X > a)$ is a random quantity with realm $\{0, 1\}$. I refer to A as a *random proposition* to emphasise that it is just a special case of a random quantity. The effect of this convention is to turn logical statements into mathematical ones. Thus if A and B are random propositions,

random proposition

$$\left.\begin{array}{c} \text{not } A \\ A \text{ and } B \\ A \text{ or } B \\ A \Rightarrow B \end{array}\right\} \text{ becomes } \left\{\begin{array}{c} 1 - A \\ AB \\ 1 - (1 - A)(1 - B) \\ A \leq B \end{array}\right.$$

and so on.

The probability of a random proposition A is defined as

$$\mathsf{Pr}(A) := \mathsf{E}(A).$$

Thus there is no explicit reason to introduce another operator for probability, if expectation is taken as primitive. But it can be useful to do so, to remind the reader that the random quantity in the argument is a proposition.

It is easy to see that elementary logical results follow immediately from this definition, with $\mathsf{Pr}(A) = 0$ being synomymous with 'it is false that A' and $\mathsf{Pr}(A) = 1$ being synomymous with 'it is true that A'. Hence, by monotonicity, $\mathsf{Pr}(A) = 1$ and $A \Rightarrow B$ imply that $\mathsf{Pr}(B) = 1$, and $A \Rightarrow B$ and $\mathsf{Pr}(B) = 0$ imply that $\mathsf{Pr}(A) = 0$.

Some other simple results include $\mathsf{Pr}(AB) \leq \mathsf{Pr}(A)$, and if $A \Rightarrow B$ then $\mathsf{Pr}(A) \leq \mathsf{Pr}(B)$, both by monotonicity. And

$$\mathsf{Pr}(A \text{ or } B) = \mathsf{Pr}(A) + \mathsf{Pr}(B) - \mathsf{Pr}(AB),$$

by linearity. This result can be extended to the disjunction of any finite set of random propositions.

One useful notational convention is to write $\mathsf{Pr}(AB)$ as $\mathsf{Pr}(A, B)$. For example, if $A := (X = x)$ and $B := (Y = y)$ then

$$\mathsf{Pr}(AB) = \mathsf{Pr}\{(X = x)(Y = y)\} = \mathsf{Pr}(X = x, Y = y).$$

This is a good convention, because it saves on parentheses, since the comma has a lower priority than all binary relations. Just be clear, though, that commas only occur in probability statements, never in expectations. I will tend to write $\mathsf{Pr}(AB)$ for symbolic random propositions, and $\mathsf{Pr}(X = x, Y = y)$ for explicit random propositions involving binary relations.

An aside on the conventional definition. Probability is conventionally considered to be a measure on subsets of some set Ω. For any $A \subset \Omega$, $\mathsf{Pr}(A)$ is defined as $\mathsf{E}(\omega \in A)$, where ω is the 'true but unknown state of nature'. It is easy to verify that under this definition, 'Pr' satisfies the three axioms of probability, namely:

1. $\mathsf{Pr}(A) \geq 0$; by non-negativity, since $(\omega \in A) \in \{0, 1\}$.
2. $\mathsf{Pr}(\Omega) = 1$; by normalisation, since $(\omega \in \Omega) = 1$.
3. $\mathsf{Pr}(A \cup B) = \mathsf{Pr}(A) + \mathsf{Pr}(B)$ whenever A and B are disjoint; by additivity, since $(\omega \in A \cup B) = (\omega \in A) + (\omega \in B)$ in this case.

1.5 *The 'Fundamental Theorem of Prevision'*

Fundamental Theorem of Prevision (FTP)

The *Fundamental Theorem of Prevision (FTP)* is due to Bruno de Finetti (de Finetti, 1974, sec. 3.10).[9] It provides a complete characterisation of the set of expectations that are consistent with the axioms of expectation given in Definition 1.1.

[9] I am following Lad (1996, ch. 2) in using this particular name.

partition

First it is necessary to define a *partition*.

Definition 1.2 (Partition). *$\mathcal{D} := \{D_1, D_2, \ldots\}$ is a partition exactly when each of the D's is a random proposition, and $\sum_i D_i = 1$.*

A partition divides up the world into a set of mutually exclusive and exhaustive potential outcomes. The simplest partition is $\{A, 1 - A\}$ for any random proposition A. More helpful, though, is a *sufficiently fine partition*.

sufficiently fine partition

Definition 1.3 (Sufficiently fine partition). *A partition $\mathcal{D}$ is sufficiently fine for a collection of random quantities $\mathcal{S}$ exactly when all real-valued functions of the elements of $\mathcal{S}$ may be treated as deterministic functions of $\mathcal{D}$.*

If I judge $\mathcal{D} := \{D_1, D_2, \ldots\}$ to be a sufficiently fine partition for a collection of random quantities which includes X then I can write

$$X = \sum_i x_i D_i \quad \text{for known } x_1, x_2, \cdots \in \mathcal{X}.$$

This is the mathematical expression of "if outcome D_i occurs, then I know that X will be equal to x_i". Likewise, if, say, $\{X, Y, Z\} \subset \mathcal{S}$ then $g(X, Y, Z) = \sum_i g(x_i, y_i, z_i) D_i$ for any real-valued function g.

Here is the FTP. It asserts the equivalence of E being a valid expectation operator (i.e. satisfying the three axioms in Definition 1.1), and a representation of E in terms of a sufficiently fine partition.

Theorem 1.3 (FTP, finite case). *Let $\mathcal{D} := \{D_1, \ldots, D_m\}$ be a finite partition which is sufficiently fine for a collection of random quantities $\mathcal{S}$. Let X be any real-valued function of the elements of $\mathcal{S}$. Then the following two statements are equivalent.*

1. *E is a valid expectation operator.*

2. *there exists $(p_1, \ldots, p_m)$ with $p_i \geq 0$ and $\sum_i p_i = 1$ for which*

$$E(X) = \sum_{i=1}^{m} x_i \, p_i \tag{1.1}$$

for all X, and $p_i = Pr(D_i)$.

Proof. Note that in both branches of this proof, it is important that $\mathcal{D}$ be a *finite* partition. This is either because sums must have well-defined limits, or because the expectation is taken inside the sum.

(1. $\Leftarrow$ 2.) This is just a matter of checking that (1.1) satisfies the three axioms.

1. (Normalisation) If $\vDash\{X = 1\}$ then $x_i = 1$ for all i and $\mathsf{E}(X) = \sum_i p_i = 1$ as required.

2. (Positivity) If $\vDash\{X > 0\}$ then $x_i > 0$ for all i and $\mathsf{E}(X) > 0$ as required, since at least one of the p_i must be positive, and none can be negative.

3. (Additivity)

$$\mathsf{E}(X) + \mathsf{E}(Y) = \sum_i x_i \, p_i + \sum_i y_i \, p_i = \sum_i (x_i + y_i) p_i = \mathsf{E}(X + Y)$$

as required.

To show that $p_j = \mathsf{Pr}(D_j)$, write $D_j = \sum_i (i = j) D_i$, and then

$$\mathsf{Pr}(D_j) = \mathsf{E}(D_j) = \sum_i (i = j) p_i = p_j,$$

as required.

(1. $\Rightarrow$ 2.) Since $X = \sum_i x_i \, D_i$ so $\mathsf{E}(X) = \sum_i x_i \, \mathsf{Pr}(D_i)$ by linearity. Set $p_i := \mathsf{Pr}(D_i)$. Since $\mathcal{D}$ is a partition, $p_i \geq 0$ (non-negativity) and $\sum_i p_i = 1$ (normalisation), as required. □

Because it is 'if and only if', the FTP characterises every possible valid relationship that can exist between expectations (including probabilities). It will be used several times in the next few sections, and is crucial in Section 1.8. It does not hold when $\mathcal{D}$ is a non-finite partition: this is discussed and rectified in Section 1.8.1.

It might appear as though the FTP depends on the choice of sufficiently fine partition. That this is not true can be inferred from the following result, in which the actual choice of $\mathcal{D}$ does not matter.

Theorem 1.4. *Let X be any random quantity. If there is a finite sufficiently fine partition for X then*

$$E(X) = \sum_{x \in \mathcal{X}} x \, Pr(X = x).$$

Proof. Let $\mathcal{D} := \{D_1, \ldots, D_m\}$ be any finite sufficiently fine partition for X, which implies that

$$X = \sum_{i=1}^{m} x_i \, D_i$$

for some known $(x_1, \ldots, x_m) \in \mathcal{X}^m$. Now $(X = x)$ is a real-valued function of X for any given $x \in \mathcal{X}$, and thus, by the FTP,

$$\mathsf{Pr}(X = x) = \sum_i (x_i = x) p_i$$

where $p_i = \mathsf{Pr}(D_i)$. And by the FTP again, using the identity $x_i = \sum_{x \in \mathcal{X}} x(x_i = x)$,

$$\begin{aligned} \mathsf{E}(X) &= \sum_i x_i \, p_i \\ &= \sum_i \sum_{x \in \mathcal{X}} x(x_i = x) p_i \\ &= \sum_{x \in \mathcal{X}} x \sum_i (x_i = x) p_i \\ &= \sum_{x \in \mathcal{X}} x \, \mathsf{Pr}(X = x) \end{aligned}$$

as was to be shown. □

1.6 Conditional expectation

Anyone who has done a first course in probability knows the 'definition' of a conditional probability. If A and B are propositions, then

$$\mathsf{Pr}(A \mid B) = \frac{\mathsf{Pr}(AB)}{\mathsf{Pr}(B)} \quad \text{if } \mathsf{Pr}(B) > 0.$$

(Recollect that AB is the proposition 'A and B'.) The underlying definition of conditional expectation must be

$$\mathsf{E}(X \mid B) := \frac{\mathsf{E}(XB)}{\mathsf{Pr}(B)} \quad \text{if } \mathsf{Pr}(B) > 0,$$

from which the first expression follows after defining $\mathsf{Pr}(A \mid B) := \mathsf{E}(A \mid B)$. Note that XB is a well-defined random quantity, which takes the value zero when B is false, and X when B is true. In both cases '$\cdot \mid B$' is read as '$\cdot$ given B', and its meaning is 'conditional on B being true'.

The difficulty with this definition is that it does not accommodate a common situation, which is where $\mathsf{Pr}(B) = 0$. This might happen, for example, if Y was a random quantity with an uncountable realm, and $B := (Y = y)$, where y is some element of the realm of Y. It turns out to be very convenient to work with such

random quantities. This difficulty was resolved by the great Soviet mathematician Andrey Kolmogorov in his 1933 book *Foundations of the Theory of Probability*.[10] He provided a characterisation of the conditional expectation which worked in great generality, and which implied the standard definitions above. I will not follow his approach, but in Section 1.6.1 I will follow a very similar one.

[10] According to Grimmett and Stirzaker (2001, p. ???), Kolmogorov wrote this book to pay for the repairs to the roof of his *dacha*.

A proper definition of $\mathsf{E}(X \mid B)$, from which the definition of $\mathsf{Pr}(A \mid B)$ follows immediately, is given at the end of Section 1.6.1. The expression for $\mathsf{Pr}(A \mid B)$ is given and developed in Section 1.7.

1.6.0 Preliminary concepts

Although I maintain that all operationally-defined random quantities should have finite expectations, and likewise all real-valued functions of them, I have not insisted on finite expectations (see Section 1.3). But this section does require a restriction on the expectations of random quantities.

Definition 1.4 (Square integrable)**.** *The random quantity X is termed* square integrable *exactly when $\mathsf{E}(X^2)$ is finite.*

square integrable

Square integrability has implications for other expectations as well.

Theorem 1.5. *If X and Y are square integrable, then $\mathsf{E}(XY)$ is finite.*

Proof. Follows immediately from Schwarz's inequality (Theorem 1.2). □

As a special case, set $X \leftarrow |X|$ and $Y \leftarrow 1$ to infer that if X is square integrable, then $\mathsf{E}(|X|)$ and $\mathsf{E}(X)$ are finite.[11]

[11] Or, in the term used in Section 1.3, X is 'integrable'.

The second important preliminary concept is that random quantities can be effectively the same, even though they are not identical. Two random quantities X and Y are identical if the operations described in X and those described in Y cannot lead to different values. In this case I write $\vDash\{X = Y\}$. This assertion is stronger than a judgement.

More generally, however, it may be the case that while X and Y have different definitions, in my judgement they are *not materially different*. How might this be represented?

not materially different

Definition 1.5 (Not materially different)**.** *X and Y are not materially different exactly when*

$$\mathsf{E}\{g(X) \cdot Z\} = \mathsf{E}\{g(Y) \cdot Z\}$$

for all real-valued g and all Z.

Informally, X and Y are not materially different for me if I could substitute one for the other in an inference, and draw the same conclusions. This will turn out to be equivalent to the following property.

Definition 1.6 (Mean-square equivalent)**.** *X and Y are* mean-square equivalent*, written $X \overset{\text{ms}}{=} Y$, exactly when $\mathsf{E}\{(X-Y)^2\} = 0$.*

mean-square equivalent

There will be a lot of mean-square equivalence in the next few subsections, and so it is helpful to be able to interpret this as the more intuitive (for me) 'not materially different'.

Theorem 1.6. *Let X and Y be square integrable. Then X and Y are mean-square equivalent if and only if they are not materially different.*

Proof.

($\Leftarrow$). Set $g \leftarrow 1$ and $Z \leftarrow X - Y$ and it follows immediately that $X \overset{\text{ms}}{=} Y$.

($\Rightarrow$). In this branch I will assume the existence of a finite sufficiently fine partition. Let $X \overset{\text{ms}}{=} Y$. According to the FTP, if $\mathcal{D}$ is a finite sufficiently fine partition for $\{X, Y, Z\}$ then

$$\mathsf{E}\{(X-Y)^2\} = \sum_i (x_i - y_i)^2 \, p_i = 0.$$

Thus X and Y must take the same value on elements of $\mathcal{D}$ which have $p_i > 0$. So when we consider $\mathsf{E}\{(X-Y)Z\}$ we have, again by the FTP,

$$\mathsf{E}\{(X-Y)Z\} = \sum_i (x_i - y_i) z_i \, p_i = \sum_i (x_i - x_i) z_i \, p_i = 0.$$

And hence $\mathsf{E}(XZ) = \mathsf{E}(YZ)$. As Z was arbitrary, this holds for all Z, and generalises immediately to any real-valued g. □

1.6.1 Characterisation

Suppose that I wish to predict a random quantity X based on the values of a set of random quantities $\boldsymbol{Y} := (Y_1, \ldots, Y_n)$. Let $\mathcal{G}$ be the set of all real scalar functions of $\boldsymbol{y} := (y_1, \ldots, y_n)$. Note for later reference that $\mathcal{G}$ includes g's for which $g(\boldsymbol{Y})$ is square integrable, such as $g(\boldsymbol{y}) = a$ where a is any constant. I would like to find the 'best' g in $\mathcal{G}$, measured by the closeness of X to $g(\boldsymbol{Y})$.

Now suppose I define 'best' in the following way:

$$\psi := \underset{g \in \mathcal{G}}{\operatorname{argmin}} \; \mathsf{E}\left[\{X - g(\boldsymbol{Y})\}^2\right]. \tag{1.2}$$

In other words, the best choice of g minimises my expectation of the squared difference between X and the random quantity $g(\boldsymbol{Y})$. Why this objective function and not some other? The ideal choice of g would have $X \overset{\text{ms}}{=} g(\boldsymbol{Y})$, because in this case X and $g(\boldsymbol{Y})$ would be not materially different from each other, and I could use $g(\boldsymbol{Y})$ in place of X. So (1.2) is asserting that I would like my choice of g to make $g(\boldsymbol{Y})$ as close to 'not materially different from X' as possible.

Theorem 1.7. *The optimisation problem* (1.2) *is well-posed if and only if X is square integrable.*

Proof. Expanding out the objective function in (1.2),

$$\mathsf{E}\left[\{X - g(\boldsymbol{Y})\}^2\right] = \mathsf{E}\left[X^2 - 2Xg(\boldsymbol{Y}) + g(\boldsymbol{Y})^2\right].$$

The optimisation is well-posed exactly when all three terms on the righthand side have finite expectations for at least one element of $\mathcal{G}$.

($\Leftarrow$). There are elements of $\mathcal{G}$ for which $g(\boldsymbol{Y})$ is square integrable. Then if X is square integrable so is $\mathsf{E}\{Xg(\boldsymbol{Y})\}$, by Theorem 1.5, and hence all three terms are finite.

($\Rightarrow$). If X is not square integrable then clearly the righthand side is not finite for any $g \in \mathcal{G}$. □

So let X be square integrable. Without loss of generality redefine $\mathcal{G}$ to be

$$\mathcal{G} := \left\{g : \mathcal{Y} \to \mathbb{R}, \text{such that } g(\boldsymbol{Y}) \text{ is square integrable}\right\}$$

Now we derive a necessary condition for ψ to be a solution to (1.2).[12] Consider a small perturbation $\psi'(\boldsymbol{y}) := \psi(\boldsymbol{y}) + \varepsilon g(\boldsymbol{y})$ for arbitrary $g \in \mathcal{G}$. Then

[12] The material leading up to Theorem 1.8 draws heavily on Whittle (2000, sec. 5.3).

$$\mathsf{E}\left[\{X - \psi'(\boldsymbol{Y})\}^2\right] = \mathsf{E}\left[\{X - \psi(\boldsymbol{Y})\}^2\right] + 2\varepsilon\,\mathsf{E}\left[\{X - \psi(\boldsymbol{Y})\}g(\boldsymbol{Y})\right] + \varepsilon^2\,\mathsf{E}\{g(\boldsymbol{Y})^2\}$$

Hence

$$\mathsf{E}\left[\{X - \psi(\boldsymbol{Y})\}g(\boldsymbol{Y})\right] = 0 \quad \text{for all } g \in \mathcal{G} \tag{1.3}$$

is a necessary condition for ψ to be a minimum.

On the other hand, suppose that ψ satisfies (1.3). Set $g(\boldsymbol{y}) \leftarrow \psi(\boldsymbol{y}) - g(\boldsymbol{y})$ in (1.3) to deduce that

$$\mathsf{E}\left[\{X - \psi(\boldsymbol{Y})\}\{\psi(\boldsymbol{Y}) - g(\boldsymbol{Y})\}\right] = 0.$$

Now write $X - g(\boldsymbol{Y}) = X - \psi(\boldsymbol{Y}) + \psi(\boldsymbol{Y}) - g(\boldsymbol{Y})$ to deduce that

$$\mathsf{E}\left[\{X - g(\boldsymbol{Y})\}^2\right] = \mathsf{E}\left[\{X - \psi(\boldsymbol{Y})\}^2\right] + \mathsf{E}\left[\{\psi(\boldsymbol{Y}) - g(\boldsymbol{Y})\}^2\right] \tag{1.4}$$

as the cross-term is zero. This is minimised over g at $g = \psi$. Thus, (1.3) is sufficient for ψ to be a solution to (1.2).

Finally, assign $g \leftarrow \psi'$ in (1.4) to conclude that if ψ and ψ' both minimise $\mathsf{E}\left[\{X - g(\boldsymbol{Y})\}^2\right]$ then $\psi(\boldsymbol{Y}) \overset{\text{ms}}{=} \psi'(\boldsymbol{Y})$. We have proved the following theorem.

Theorem 1.8. *Let X be square integrable. Then ψ is a solution to* (1.2) *if and only if* (1.3) *holds; such a solution exists, $\psi(\boldsymbol{Y})$ is square integrable, and if ψ and ψ' are two solutions then $\psi(\boldsymbol{Y}) \overset{\text{ms}}{=} \psi'(\boldsymbol{Y})$.*

1.6.2 Notation and definitions

Notation for conditional expectation is important enough to have its own section! I will assume that all random quantities are square integrable.

First, we need a container for all solutions to (1.3). Therefore I write

$$\mathcal{E}(X \mid \boldsymbol{Y}) := \left\{\psi \in \mathcal{G} : \psi \text{ solves eq. (1.3)}\right\}. \tag{1.5}$$

Typical members of $\mathcal{E}(X \mid \boldsymbol{Y})$ will be denoted ψ, ψ' and so on. So Theorem 1.8 might have been written:

If X is square integrable, then $\mathcal{E}(X \mid \boldsymbol{Y})$ is non-empty. If $\psi, \psi' \in \mathcal{E}(X \mid \boldsymbol{Y})$, then $\psi(\boldsymbol{Y})$ is square integrable, and $\psi(\boldsymbol{Y}) \overset{\text{ms}}{=} \psi'(\boldsymbol{Y})$.

Now for the definition of *conditional expectation*.

conditional expectation

Definition 1.7 (Conditional expectation)**.** *Let X be square integrable. The conditional expectation of X given $\boldsymbol{Y}$ is*

$$\mathbb{E}(X \mid \boldsymbol{Y}) := \psi(\boldsymbol{Y})$$

where $\psi \in \mathcal{E}(X \mid \boldsymbol{Y})$.

Thus the conditional expectation is a random quantity, and it is mean-square unique.[13] It is the random quantity which best represents X using only $\boldsymbol{Y}$, according to the loss function in (1.2).

[13] $\psi(\boldsymbol{Y}), \psi'(\boldsymbol{Y}), \ldots$ are termed *versions* of the conditional expectation.

What about the conventional expectation, given at the start of this section? This is a fundamentally different object, because $\mathsf{E}(X \mid B)$ is a value, *not* a random quantity. Hence the need for two different notations, E and $\mathbb{E}$. But E is defined in terms of $\mathbb{E}$.

Definition 1.8 (Conventional conditional expectation)**.** *If B is a random proposition and $\mathsf{Pr}(B) > 0$, then*

$$\mathsf{E}(X \mid B) := \psi(1) \quad \textit{where } \psi \in \mathcal{E}(X \mid B).$$

This makes $\mathsf{E}(X \mid B)$ a value with the meaning 'the expectation of X conditional on B being true'. The definition might seem ambiguous, given that $\mathcal{E}(X \mid B)$ may contain many elements, but for the following result.

Theorem 1.9. *Let $\psi, \psi' \in \mathcal{E}(X \mid B)$. Then $\mathsf{Pr}(B) > 0$ is sufficient for $\psi(1) = \psi'(1)$.*

Proof. $\{\overline{B}, B\}$ is a finite sufficiently fine partition for B and for real-valued functions of B, where $\overline{B} := 1 - B$. Hence, by the FTP (Theorem 1.3)

$$\mathsf{E}[\{\psi(B) - \psi'(B)\}^2] = \{\psi(0) - \psi'(0)\}^2 \,\mathsf{Pr}(\overline{B}) + \{\psi(1) - \psi'(1)\}^2 \,\mathsf{Pr}(B) = 0,$$

since $\psi(B) \overset{\text{ms}}{=} \psi'(B)$. Therefore $\mathsf{Pr}(B) > 0$ implies that $\psi(1) = \psi'(1)$. □

Therefore, the condition $\mathsf{Pr}(B) > 0$ in the conventional definition ought to be recognised as the condition which ensures the uniqueness of $\mathsf{E}(X \mid B)$: it has nothing to do with 'dividing by zero'.

Here is a little table to keep track of the different E's:

$\mathcal{E}(X \mid \boldsymbol{Y})$: A set of functions of $\boldsymbol{y}$, defined in (1.5),
$\mathbb{E}(X \mid \boldsymbol{Y})$: A random quantity, defined in Definition 1.7,
$\mathsf{E}(X \mid B)$: A value, defined in Definition 1.8.

Remember that $\mathsf{E}(X)$ and $\mathsf{E}(X \mid B)$ are two completely different objects. The first is a 'primitive'—a reflection of my judgements. The second is a construction arising out of my judgements: it comes 'for free' once I have specified certain of my expectations (see Theorem 1.14).

1.6.3 *Properties of the conditional expectation*

Here are the most important properties of the conditional expectation, all inferred from (1.3) via Theorem 1.8. I will assume that all random quantities are square integrable, and for simplicity I will just use scalar Y's. The two main properties are that $\mathbb{E}$ is indeed an expectation (justifying its name and its symbol), and the Law of the Iterated Expectation. Then there are some useful special cases.

Theorem 1.10. *$\mathbb{E}(\cdot \mid Y)$ satisfies the axioms of expectation given in Definition 1.1, in mean-square.*

Proof. This is a matter of checking the three axioms one by one. In each case we find a $\psi \in \mathcal{E}(X \mid Y)$ for which $\psi(Y)$ has the required property in mean-square; and then $\mathbb{E}(X \mid Y)$ must have the required property in mean-square, according to Theorem 1.8.

1. Normalisation. Let $\vDash\{X = 1\}$. If $\psi(y) = 1$ then $\psi \in \mathcal{E}(X \mid Y)$ and the result follows immediately.

2. Positivity.[14] Let $\vDash\{X > 0\}$ and $\psi \in \mathcal{E}(X \mid Y)$. Let $g \leftarrow \psi^-$ in (1.3), where

$$\psi^-(y) := \begin{cases} \psi(y) & \psi(y) \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

   Then, from (1.3),

$$\mathsf{E}\{X\psi^-(Y)\} = \mathsf{E}\{\psi(Y)\psi^-(Y)\}.$$

   But if $\vDash\{X > 0\}$ the lefthand side is non-positive. The righthand side is non-negative, by construction. Hence $\mathsf{E}\{\psi(Y)\psi^-(Y)\} = 0$. But since $\psi(y)\psi^-(y) = \{\psi^-(y)\}^2$, so $\mathsf{E}[\{\psi^-(Y)\}^2] = 0$, or $\psi(Y) > 0$ in mean-square.

3. Additivity. If $\psi \in \mathcal{E}(X \mid Y)$ and $\psi' \in \mathcal{E}(X' \mid Y)$ then

$$\psi + \psi' \in \mathcal{E}(X + X' \mid Y),$$

   and the result follows immediately.

□

[14] This proof adapted from Whittle (2000, section 5.3).

Having established that $\mathbb{E}$ is indeed an expectation, we now turn to the very important *Law of Iterated Expectation (LIE)*. A simpler expression of the LIE is given below in (1.6), along with a brief discussion.

Law of Iterated Expectation (LIE)

Theorem 1.11 (Law of Iterated Expectation)**.**

$$\mathbb{E}(X \mid Z) \overset{\text{ms}}{=} \mathbb{E}\{\mathbb{E}(X \mid Y, Z) \mid Z\}.$$

Proof. (Whittle, 2000, section 5.3). Let

$$\psi_1 \in \mathcal{E}(X \mid Z) \quad \psi_2 \in \mathcal{E}(X \mid Y, Z) \quad \phi \in \mathcal{E}\big(\mathbb{E}(X \mid Y, Z) \mid Z\big).$$

Then three applications of (1.3) gives

$$\begin{aligned}
\mathsf{E}\left[\{X-\psi_1(Z)\}g_1(Z)\right] &= 0\\
\mathsf{E}\left[\{X-\psi_2(Y,Z)\}g_2(Y,Z)\right] &= 0\\
\mathsf{E}\left[\{\psi_2(Y,Z)-\phi(Z)\}g_3(Z)\right] &= 0.
\end{aligned}$$

Now set g_1, g_2, and g_3 to $\phi - \psi_1$, and subtract the second and third equations from the first, to get

$$\mathsf{E}\left[\{\phi(Z)-\psi_1(Z)\}^2\right] = 0$$

or $\mathbb{E}\{\mathbb{E}(X \mid Y,Z) \mid Z\} \overset{\text{ms}}{=} \mathbb{E}(X \mid Z)$, as was to be proved. □

The next result lists some useful special cases. The proofs are not given, being straightforward: in each case one simply verifies that there is a $\psi \in \mathcal{E}$ which satisfies the equality, and then mean-square equivalence of $\mathbb{E}$ follows from Theorem 1.8.

Theorem 1.12 (Conditional expectation, special cases)**.**

1. $\mathbb{E}(X \mid X) \overset{\text{ms}}{=} X$*, and, by extension,* $\mathbb{E}(X \mid X,Y) \overset{\text{ms}}{=} X$*.*
2. *If* $Y = a$*, some constant, then* $\mathbb{E}(X \mid Y) \overset{\text{ms}}{=} \mathsf{E}(X)$*.*
3. *If the elements of* $\mathcal{E}(X \mid Y,Z)$ *are invariant to* z*, then* $\mathbb{E}(X \mid Y) \overset{\text{ms}}{=} \mathbb{E}(X \mid Y,Z)$*.*
4. $\mathbb{E}\{g(Y)X \mid Y\} \overset{\text{ms}}{=} g(Y)\,\mathbb{E}(X \mid Y)$*.*

Parts (2) and (3) can be combined to provide a simpler representation of the LIE:

$$\mathsf{E}(X) = \mathsf{E}\{\mathbb{E}(X \mid Y)\} \tag{1.6}$$

(put $Z = a$, some constant). This states is that I can specify my expectation for X by thinking about my conditional expectation for X given Y, and then thinking about my expectation of this function of Y. It is often the case that breaking an expectation down into two or more steps is helpful, typically because one of the steps will be easier to assess than the others. Often $\mathbb{E}(X \mid Y)$ is quite easy (or uncontroversial) to assess, but Y itself is a random quantity about which I have limited judgements. In this case the convexity property of expectation asserts that I can bound my expectation for X by the smallest and largest values of the realm of $\mathbb{E}(X \mid Y)$.

1.6.4 The special case of a finite realm

Now consider the special case where the realm of Y is finite. In this case we can derive an explicit representation of $\psi \in \mathcal{E}(X \mid Y)$. I will continue to assume that all random quantities are square integrable, and, for simplicity, continue to use scalar Y's.

Initially, consider a random proposition B. Note that in the Theorem below I have written $(B = b_i)$ for clarity, where $b_i \in \{0,1\}$, but of course $(B = 0) = 1 - B$ and $(B = 1) = B$. Below I will define $\overline{B} := 1 - B$, where $\overline{B}$ denotes the random proposition 'B is false' (or 'not B' for short).

Theorem 1.13. *Let X be square integrable. If B is a random proposition and $\phi \in \mathcal{E}(X \mid B)$ then*

$$\phi(b_i) = \frac{E\{X(B = b_i)\}}{Pr(B = b_i)} \quad \text{if } Pr(B = b_i) > 0$$

and undefined otherwise, where $b_i \in \{0, 1\}$.

Proof. The realm of B is $\mathcal{B} := \{0, 1\}$. Let $\mathcal{G}$ be the set of all real-valued functions defined on $\mathcal{B}$: these can all be written as

$$g(b) = \alpha_0(b = 0) + \alpha_1(b = 1) \quad \text{for some } \alpha_0, \alpha_1 \in \mathbb{R}.$$

Let ϕ be written

$$\phi(b) = \beta_0(b = 0) + \beta_1(b = 1) \quad \text{for some } \beta_0, \beta_1 \in \mathbb{R}.$$

We need to find values of (β_0, β_1) for which (1.3) is true for all values of (α_0, α_1). Let $\overline{B} := 1 - B$. I will show that

$$\beta_0 = \frac{\mathsf{E}(X\overline{B})}{\mathsf{Pr}(\overline{B})} \quad \text{if } \mathsf{Pr}(\overline{B}) > 0$$
$$\text{and } \beta_1 = \frac{\mathsf{E}(XB)}{\mathsf{Pr}(B)} \quad \text{if } \mathsf{Pr}(B) > 0.$$

Starting from (1.3) and substituting for g and ϕ, (β_0, β_1) must satisfy

$$\mathsf{E}\left[\{X - (\beta_0\overline{B} + \beta_1 B)\}(\alpha_0\overline{B} + \alpha_1 B)\right] = 0 \quad \text{for all } \alpha_0, \alpha_1.$$

This is possible if and only if

$$\mathsf{E}\left[\{X - (\beta_0\overline{B} + \beta_1 B)\}\overline{B}\right] = 0$$
$$\text{and } \mathsf{E}\left[\{X - (\beta_0\overline{B} + \beta_1 B)\}B\right] = 0.$$

Multiplying out and taking expectations then gives

$$\mathsf{E}(X\overline{B}) - \beta_0 \mathsf{Pr}(\overline{B}) = 0$$
$$\text{and } \mathsf{E}(XB) - \beta_1 \mathsf{Pr}(B) = 0$$

because $\mathsf{E}(\overline{B}B) = 0$, and $\mathsf{E}(\overline{B}^2) = \mathsf{E}(\overline{B}) = \mathsf{Pr}(\overline{B})$, and the same for $\mathsf{E}(B^2)$. Focusing on β_0, if $\mathsf{Pr}(\overline{B}) > 0$ then there is a unique solution for β_0, as given above. Otherwise, if $\mathsf{Pr}(\overline{B}) = 0$ then Schwarz's inequality (Theorem 1.2) states that

$$\{\mathsf{E}(X\overline{B})\}^2 \leq \mathsf{E}(X^2)\,\mathsf{E}(\overline{B}^2) = \mathsf{E}(X^2)\,\mathsf{Pr}(\overline{B}) = 0$$

and so the equation has the form $0 - \beta_0 \times 0 = 0$. Hence β_0 is undefined in this case. An identical argument holds for β_1. □

Here is one immediate corollary of Theorem 1.13, which follows directly from the definition of $\mathsf{E}(X \mid B)$ given in Definition 1.8. This theorem is the basis for all of the standard results on conditional probability that are presented in Section 1.7.

Theorem 1.14. *Let X be square integrable. If B is a random proposition, then*

$$E(X \mid B) = \frac{E(XB)}{Pr(B)} \quad \text{if } Pr(B) > 0$$

and undefined otherwise.

Proof. From the definition, $\mathsf{E}(X \mid B) = \psi(1)$ where $\psi \in \mathcal{E}(X \mid B)$. Setting $b_i = 1$ in Theorem 1.13 gives the result. □

Finally, consider the more general case of conditioning on Y, a random quantity with a finite realm.

Theorem 1.15. *Let X be square integrable; let Y have a finite realm, $\mathcal{Y} := \{y_1, \ldots, y_m\}$; and let $\psi \in \mathcal{E}(X \mid Y)$. Then*

$$\psi(y_i) = E(X \mid Y = y_i) \quad \text{if } Pr(Y = y_i) > 0$$

and undefined otherwise.

This theorem states that $\psi \in \mathcal{E}(X \mid Y)$ takes the same value at y_i as $\phi_i(1)$, where $\phi_i \in \mathcal{E}(X \mid Y = y_i)$, thinking of $(Y = y_i)$ as a random proposition. Plugging in from Theorem 1.14 with $B := (Y = y_i)$,

$$\psi(y_i) = \mathsf{E}(X \mid Y = y_i) = \frac{\mathsf{E}\{X(Y = y_i)\}}{\mathsf{Pr}(Y = y_i)} \quad \text{if } \mathsf{Pr}(Y = y_i) > 0,$$

and undefined otherwise. This result is probably obvious (but reassuring!), but it can also be proved in the same way as Theorem 1.13.

1.7 Conditional probabilities

There is nothing new to say here! Conditional probabilities are just conditional expectations. But this section presents some of the standard results starting from Theorem 1.14 and the following definition.

Definition 1.9 (Conditional probability)**.** *Let A and B be random propositions. Then*

$$Pr(A \mid B) := E(A \mid B) \quad \text{if } Pr(B) > 0$$

and undefined otherwise.

Then by Theorem 1.14 we have

$$\mathsf{Pr}(A \mid B) = \frac{\mathsf{Pr}(AB)}{\mathsf{Pr}(B)} \quad \text{if } \mathsf{Pr}(B) > 0 \tag{1.7}$$

and undefined otherwise, where AB is the proposition 'A and B'.

Eq. (1.7) is often given as the definition of conditional probability, as stated at the start of Section 1.6. It is important to understand this is *not* the definition of conditional probability—it is a theorem. Go back and have another look at the end of Section 1.6.2 if this is not clear. As already explained there, the restriction to $\mathsf{Pr}(B) > 0$ may look like 'not dividing by zero', but in fact its real purpose is to make sure that $\mathsf{E}(A \mid B)$ is uniquely defined.

There are some useful relations between conditional probabilities, including 'unconditional' probability as a special case. The first one states that conditioning on a conjunction B_2B_1 gives the same result as conditioning on B_1 and then conditioning on B_2.

Theorem 1.16 (Sequential conditioning)**.**

$$\Pr(A \mid B_2B_1) = \frac{\Pr(AB_2 \mid B_1)}{\Pr(B_2 \mid B_1)} \quad \text{if } \Pr(B_2B_1) > 0$$

and undefined otherwise.

Proof. From Schwarz's inequality, $\Pr(B_2B_1) > 0$ implies that $\Pr(B_2) > 0$ and $\Pr(B_1) > 0$. Then

$$\Pr(A \mid B_2B_1) = \frac{\Pr(AB_2B_1)}{\Pr(B_2B_1)} = \frac{\Pr(AB_2B_1)}{\Pr(B_1)} \frac{\Pr(B_1)}{\Pr(B_2B_1)} = \frac{\Pr(AB_2 \mid B_1)}{\Pr(B_2 \mid B_1)}.$$

□

The following result is an immediate extension. Its purpose is constructive—often it is a lot easier to specify a single probability over a conjunction by specifying a marginal probability and one or more conditional probabilities.

Theorem 1.17 (Factorisation theorem)**.**

$$\Pr(B_1 \cdots B_n) = \Pr(B_1) \prod_{i=2}^{n} \Pr(B_i \mid B_1 \cdots B_{i-1}),$$

or zero if any of the terms in the product are undefined.

Proof. It suffices to set $n = 3$. Then

$$\Pr(B_1B_2B_3) = \Pr(B_1) \Pr(B_2B_3 \mid B_1) \quad \text{if } \Pr(B_1) > 0$$

and zero otherwise, from (1.7). But by sequential conditioning (Theorem 1.16),

$$\Pr(B_2B_3 \mid B_1) = \Pr(B_2 \mid B_1) \Pr(B_3 \mid B_1, B_2) \quad \text{if } \Pr(B_2 \mid B_1) > 0$$

and zero otherwise. Combining these gives the result. For $n > 3$, the result can be iterated. □

Then there is the very useful *Law of Total Probability (LTP)*, also known as the *Partition Theorem*.

Law of Total Probability (LTP)
Partition Theorem

Theorem 1.18 (Law of Total Probability)**.** *Let $\mathcal{D} := \{D_1, \ldots, D_m\}$ be any finite partition. Then*

$$\Pr(A) = \sum_{i=1}^{m} \Pr(A \mid D_i) \Pr(D_i),$$

where terms with $\Pr(D_i) = 0$ are dropped.

Proof. As $\sum_i D_i = 1$, we have

$$A = A\Big(\sum_{i=1}^{m} D_i\Big) \text{ and } \Pr(A) = \sum_{i=1}^{m} \Pr(AD_i).$$

If $\Pr(D_i) = 0$ then $\Pr(AD_i) = 0$ by Schwarz's inequality (Theorem 1.2), and so such terms can be dropped. For the other terms, $\Pr(AD_i) = \Pr(A \mid D_i) \Pr(D_i)$ by (1.7), and the result follows. □

The LTP plays the same role as the LIE (Theorem 1.11). In particular, in situations where it is hard to assess $\Pr(A)$ directly, it is possible to bound $\Pr(A)$ using the lower and upper limits of $\Pr(A \mid D_i)$ taken over all $D_i \in \mathcal{D}$.

Finally, there is the celebrated *Bayes's Theorem*.

Bayes's Theorem

Theorem 1.19 (Bayes's Theorem)**.** *If* $Pr(B) > 0$ *then*

$$Pr(A \mid B) = \begin{cases} 0 & Pr(A) = 0 \\ \dfrac{Pr(B \mid A)\, Pr(A)}{Pr(B)} & \text{otherwise.} \end{cases}$$

Proof. Two cases are required, $\Pr(A) = 0$ and $\Pr(A) > 0$, because $\Pr(B \mid A)$ is only defined in the latter case.

First, $\Pr(A) = 0$ implies $\Pr(AB) = 0$, by Schwarz's inequality. Then (1.7) shows that $\Pr(AB) = 0$ implies $\Pr(A \mid B) = 0$ when $\Pr(B) > 0$.

Now consider the case where $\Pr(A) > 0$. But then both $\Pr(A)$ and $\Pr(B)$ are non-zero, and

$$\Pr(AB) = \Pr(A \mid B)\Pr(B) = \Pr(B \mid A)Pr(A)$$

from (1.7). Rearranging gives the result. □

The first branch of Theorem 1.19 is important enough to be given its own name, due to Dennis Lindley (see, e.g., Lindley, 1985).

Theorem 1.20 (Cromwell's Rule)**.** *Probabilities can never be raised from zero by conditioning.*

Thus if you choose to model the learning process as probabilistic conditioning, then you should only use zero probabilities for propositions that are impossible, because if $\Pr(A) = 0$ then no amount of new information 'B is true' can change your judgement about A.

There are several other versions of Bayes's Theorem. For example, there is a sequential Bayes's Theorem:

$$\Pr(A \mid B_2B_1) = \frac{\Pr(B_2 \mid A, B_1)}{\Pr(B_2 \mid B_1)}\ \Pr(A \mid B_1)$$

if $\Pr(A) > 0$ and $\Pr(B_2B_1) > 0$. And there is Bayes's theorem for a finite partition, $\mathcal{D} := \{D_1, \ldots, D_m\}$:

$$\Pr(D_i \mid B) = \frac{\Pr(B \mid D_i)\Pr(D_i)}{\sum_j \Pr(B \mid D_j)\Pr(D_j)} \quad i = 1, \ldots, m$$

if $\Pr(A) > 0$ and $\Pr(B) > 0$, assuming for simplicity that $\Pr(D_i) > 0$ for all i. And there is Bayes's Theorem in odds form,

$$\frac{\Pr(D_i \mid B)}{\Pr(D_j \mid B)} = \frac{\Pr(B \mid D_i)}{\Pr(B \mid D_j)}\ \frac{\Pr(D_i)}{\Pr(D_j)}$$

if $\Pr(D_i), \Pr(D_j) > 0$, and $\Pr(B) > 0$.

The logic of implication. It is reassuring that conditional probability obeys the logic of implication. B implies A exactly when $B \leq A$. But if B

implies A then $AB = B$, and so $\mathsf{Pr}(A \mid B) = \mathsf{Pr}(B)/\mathsf{Pr}(B) = 1$. Likewise, if B implies 'not A' then $\mathsf{Pr}(A \mid B) = 0$.

Bayes's Theorem also gives an interesting result. Let A imply B. Then

$$\mathsf{Pr}(A \mid B) = \frac{\mathsf{Pr}(B \mid A)\,\mathsf{Pr}(A)}{\mathsf{Pr}(B)} = \frac{\mathsf{Pr}(A)}{\mathsf{Pr}(B)} \geq \mathsf{Pr}(A).$$

So if $\mathsf{Pr}(B)$ is small, then $\mathsf{Pr}(A \mid B) \gg \mathsf{Pr}(A)$. Typically A is a theory which implies a hypothesis B, and we are interested in the degree to which confirming the hypothesis (i.e. 'B is true') strengthens our belief in the theory. Naturally, this depends on how many other theories also imply B. If $\mathsf{Pr}(B)$ is small, then not many other theories imply B and hence observing B is strongly confirmatory for theory A. These kinds of results are discussed in *Bayesian Confirmation Theory*; see, e.g., Howson and Urbach (2006).

Bayesian Confirmation Theory

1.8 Probability mass functions

Consider a set of random quantities $\{X, Y, Z\}$.[15] Suppose initially that all three random quantities are discrete; i.e. that their realms are finite or countable. One way to summarise my judgements about these quantities is as a *probability mass function (PMF)*.

[15] The extension of the results in this section to any finite number of random quantities is immediate.

probability mass function (PMF)

Definition 1.10 (Probability mass function, PMF)**.** *The function $f_{X,Y,Z}$ is a probability mass function for discrete random quantities $\{X, Y, Z\}$ exactly when*

$$f_{X,Y,Z}(x, y, z) := \mathsf{Pr}\{X = x, Y = y, Z = z\}.$$

Conditional PMFs are defined below in (1.10)*. The* support *of $f_{X,Y,Z}$ is the set $\{x, y, z : f_{X,Y,Z}(x, y, z) > 0\}$.*

support

This notation, which includes both the labels of the random quantities and their arguments, is a bit cumbersome. Many statisticians would write $f(x, y, z)$, in which the labels are inferred from the symbols used for the arguments. But this can be ambiguous and I prefer to play it safe. Less ink is used when the set of random quantities is written as $\boldsymbol{X} := \{X_1, \ldots, X_n\}$, for which $f_{\boldsymbol{X}}(\boldsymbol{x})$ is a compact way of writing $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$.

In the case where all random quantities have finite realms, *the FTP completely specifies the properties of the PMF*. To see this, define the random propositions

$$D_{xyz} := (X = x)(Y = y)(Z = z) \quad \text{for all } x \in \mathcal{X},\ y \in \mathcal{Y},\ z \in \mathcal{Z}.$$

Then

$$\mathcal{D} := \bigcup_{x,y,z} \{D_{xyz}\}$$

is a finite sufficiently fine partition for $\{X, Y, Z\}$, because it represents the outer product of the individual realms. Note that $f_{X,Y,Z}(x, y, z) = \mathsf{Pr}(D_{xyz})$. The FTP states that E is a valid expectation operator if and only if

$$f_{X,Y,Z}(x, y, z) \geq 0, \quad \sum_{x,y,z} f_{X,Y,Z}(x, y, z) = 1 \tag{1.8}$$

and

$$\mathsf{E}\{g(X,Y,Z)\} = \sum_{x,y,z} g(x,y,z)\, f_{X,Y,Z}(x,y,z) \qquad (1.9)$$

for all real-valued g.

The FTP implies that the *marginal PMFs* of subsets of $\{X,Y,Z\}$ are deduced from $f_{X,Y,Z}$. For example, by setting $g(x,y,z) \leftarrow (x = x')(y = y')$ in (1.9), it is easily seen that

marginal PMFs

$$f_{X,Y}(x',y') = \sum_z f_{X,Y,Z}(x',y',z)$$

which is the standard rule for marginalising out Z.

For conditional probabilities, the definition is

$$f_{X,Y|Z}(x,y \mid z) := \mathsf{Pr}(X = x, Y = y \mid Z = z) \qquad (1.10)$$

and then Theorem 1.14 and (1.7) imply that

$$f_{X,Y|Z}(x,y \mid z) = \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)} \quad \text{if } f_Z(z) > 0 \qquad (1.11)$$

and undefined otherwise. Because conditional expectations satisfy the same axioms as expectations, conditional PMFs must have the same properties as PMFs, for all z for which $f_Z(z) > 0$. Thus they must be non-negative, sum to one, marginalise, and condition. All of these properties can be inferred from (1.11).

Modern statistical practice. This almost invariably starts by specifying marginal and conditional PMFs, constructs a joint PMF using (1.11), and then defines all expectations using (1.9). Typically the set of random quantities is augmented by additional random quantities termed *statistical parameters*. The specification of marginal and conditional PMFs over an appropriately chosen set of random quantities and statistical parameters is the subject of *statistical modelling*. This can be treated as distinct from statistical inference, even though in practice the two are tightly related due to the ease with which some types of inference can be applied to some types of model.

statistical parameters

statistical modelling

1.8.1 Non-finite realms*

The difficulty with non-finite realms is that the axiom of finite additivity is not strong enough to prove the ($\Rightarrow$) branch of the FTP, in situations where the number of terms in the partition is infinite. In this subsection I outline a generalisation for this case. But it is worth stressing, once again, that operationally defined random quantities have finite realms, and the decision to treat X as a random quantity with a non-finite realm is made for our convenience. Therefore it should *not* introduce pathologies which would not be present were X to be represented more realistically.

Note that non-finite realms may be unbounded, so that expectations may be infinite or undefined. This can be addressed using the generalisation outlined in Section 1.3, and so I will not worry about it here.

Consider the case where the realm of X is non-finite but countable,

$$\mathcal{X} := \{x_1, x_2, \ldots\}$$

where the x_i's are ordered. This X can be approximated by a truncated version

$$X_n := X(X \leq x_n) + x_n(X > x_n)$$

and hence

$$\mathcal{D}_n := \{(X = x_1), (X = x_2), \ldots, (X = x_n), (X > x_n)\}$$

is a finite sufficiently fine partition for X_n. By construction $X_1, X_2, \ldots$ is a non-decreasing sequence of random quantities for which $\lim_n X_n = X$. The idea is to approach $\mathsf{E}(X)$ through the $\mathsf{E}(X_n)$'s, but there is nothing in our axioms to ensure that this is valid. Therefore we need an additional restriction on the properties of E, which extends finite additivity to *countable additivity*.

countable additivity

Definition 1.11 (Countable additivity)**.** *Let $X_1, X_2, \ldots$ be a non-decreasing sequence of random quantities with limit X. Let E be a valid expectation operator. Then E is countably additive exactly when*

$$\mathsf{E}(X) = \lim_n \mathsf{E}(X_n).$$

If I accept countable additivity as a reasonable property of my expectations, then the ($\Rightarrow$) branch of the FTP becomes

$$\begin{aligned}
\mathsf{E}(X) &= \lim_n \mathsf{E}(X_n) \\
&= \lim_n \mathsf{E}\Big(\sum_{i=1}^{n} x_i(X = x_i) + x_n(X > x_n)\Big) \\
&= \lim_n \Big(\sum_{i=1}^{n} x_i\, f_X(x_i) + x_n \,\mathsf{Pr}(X > x_n)\Big) \\
&= \sum_{i=1}^{\infty} x_i\, f_X(x_i) \qquad (1.12)
\end{aligned}$$

as might be anticipated.[16] Eq. (1.12) gives the FTP for countable realms.

[16] The final term in the third line may have the form $\infty \cdot 0$ in the limit, but the appropriate convention in this case is $\infty \cdot 0 = 0$.

Thus there is an FTP for random quantities with non-finite countable realms, but only if the three axioms given in Definition 1.1 are augmented with a fourth axiom of countable additivity. However, this fourth axiom has a very different character.[17] It is a lot less self-evident, because we have no practical experience of reasoning about infinite sequences of random quantities—only our intuition. But if we trusted our intuition on these matters, we would not need the axioms of expectation and all their implications in the first place. In fact, countable additivity is philosophically controversial. But it is almost universally accepted as a pragmatic bridge to pass over into the convenient world of random quantities with non-finite realms.

[17] Countable additivity implies finite additivity, but it is best to keep it as a separate axiom, due to its different character.

Once countable additivity is accepted, all of the finite realm results of the previous sections hold for countable realms as well.

Now consider the case where the realm of X is finite or countable, but where the operations used to determine X are more precise than my judgements can discern. Define

$$F_X(x) := \mathsf{Pr}(X \leq x)$$

termed the *distribution function* of X. Suppose that I specify a function f_X with the property that

distribution function

$$\int_{-\infty}^{x_i} f_X(x)\,\mathrm{d}x = F_X(x_i) \quad \text{for all } x_i \in \mathcal{X}.$$

Then, setting $x_0 := -\infty$, and using countable additivity,

$$\begin{aligned}
\mathsf{E}(X) &= \sum_{i=1}^{\infty} x_i \left\{F_X(x_i) - F_X(x_{i-1})\right\} \\
&= \sum_{i=1}^{\infty} x_i \int_{x_{i-1}}^{x_i} f_X(x)\,\mathrm{d}x \\
&\approx \sum_{i=1}^{\infty} \int_{x_{i-1}}^{x_i} x\, f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x. \qquad (1.13)
\end{aligned}$$

The Riemann integral is approximating a sum over a realm with a huge number of very finely spaced points, and the reason I accept this approximation as valid is that my specified f_X respects my judgements on the countable $\mathcal{X}$, and interpolates them smoothly between the points in $\mathcal{X}$. Effectively I am approximating $\mathcal{X}$ with a a convex subset of $\mathbb{R}$. Eq. (1.13) gives the FTP for an uncountable realm.

Formally, the definition of f_X is

$$f_X(x)\,\mathrm{d}x := \mathsf{Pr}\left\{X \in [x, x + \mathrm{d}x)\right\}$$

termed the *probability density function* of X, where $\mathrm{d}x$ is a differential element. Going back to $\{X, Y, Z\}$, and defining $f_{X,Y,Z}$ in the obvious way, the FTP states that E is a valid expectation operator if and only if

probability density function

$$f_{X,Y,Z}(x,y,z) \geq 0 \quad \text{and} \quad \iiint f_{X,Y,Z}(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = 1.$$

The marginalisation result is

$$f_{X,Y}(x',y')\,\mathrm{d}x\,\mathrm{d}y = \Big[\int f_{X,Y,Z}(x',y',z)\,\mathrm{d}z\Big]\mathrm{d}x\,\mathrm{d}y.$$

And the conditioning result is

$$\begin{aligned}
f_{X,Y|Z}(x,y \mid z)\,\mathrm{d}x\,\mathrm{d}y &= \frac{f_{X,Y,Z}(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z}{f_Z(z)\,\mathrm{d}z} \\
&= \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)}\,\mathrm{d}x\,\mathrm{d}y \quad \text{if } f_Z(z) > 0
\end{aligned}$$

and undefined otherwise.

Thus valid PMFs and PDFs follow the same rules, with the only difference being the replacement of sums with integrals, and the inclusion of the differential elements $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ where appropriate.[18] This justifies the use of the same notation in both cases (even though the units are different). There is a more formal justification for the use of the same notation within the unifying treatment of Measure Theory, but this is part of the formal mathematical theory of probability, rather than the practical statistical theory of expectation and probability.

[18] The extension to a mixture of discrete and continuous random quantities is straightforward.

Hybrid random quantities. Just occasionally it is useful to specify a random quantity X with an uncountable realm but with an atom of probability of size p_a at some location x_a. Such a random quantity is not discrete, but it does not have a continuous PDF. The usual way to represent X in this case is

$$X = Ax_a + (1 - A)Y$$

where A is a random proposition, $\mathsf{Pr}(A) = p_a$, Y is a continuous random quantity, and $\mathsf{E}(AY) = \mathsf{E}(A)\,\mathsf{E}(Y)$. This construction can be extended to a countable set of atoms, using a partition.

2
Bibliography

D. Basu, 1975. Statistical information and likelihood. *Sankhyā*, **37**(1), 1–71. With discussion. 12

J.M. Bernardo and A.F.M. Smith, 1994. *Bayesian Theory*. Chichester, UK: Wiley. 7

P. Billingsley, 1995. *Probability and Measure*. John Wiley & Sons, Inc., New York NY, USA, third edition. 7, 12

B. de Finetti, 1937. la prévision, ses lois logiques, ses sources subjectives. *Annals de L'Institute Henri Poincaré*, **7**, 1–68. See de Finetti (1964). 7

B. de Finetti, 1964. Foresight, its logical laws, its subjective sources. In H. Kyburg and H. Smokler, editors, *Studies in Subjective Probability*, pages 93–158. New York: Wiley. 2nd ed., New York: Krieger, 1980. 33

B. de Finetti, 1972. *Probability, Induction and Statistics*. London: John Wiley & Sons. 7

B. de Finetti, 1974. *Theory of Probability*, volume 1. London: Wiley. 11, 13, 14, 15

B. de Finetti, 1974/75. *Theory of Probability*. London: Wiley. Two volumes (2nd vol. 1975); A.F.M. Smith and A. Machi (trs.). 7

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. John Wiley & Sons, Chichester, UK. 7

G.R. Grimmett and D.R. Stirzaker, 2001. *Probability and Random Processes*. Oxford, UK: Oxford University Press, 3rd edition. 7, 18

C. Howson and P. Urbach, 2006. *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court Publishing Co., 3rd edition. 28

J.B. Kadane, 2011. *Principles of Uncertainty*. Chapman & Hall/CRC Press, Boca Raton FL, USA. 12

F. Lad, 1996. *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 7, 9, 12, 15

D.V. Lindley, 1985. *Making Decisions*. London: John Wiley & Sons, 2nd edition. 27

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. 8

P. Whittle, 2000. *Probability via Expectation*. New York: Springer, 4th edition. 7, 20, 22

D. Williams, 1991. *Probability With Martingales*. Cambridge University Press, Cambridge, UK. 7