1
Expectation and statistical inference

The ostensible 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 Chapter 5.

What is unusual 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 some details, although my approach is quite different from his. For standard textbooks, I recommend Grimmett and Stirzaker (2001) on probability theory, Schervish (1995) on the theory of statistics, and either Bernardo and Smith (1994) or Robert (2007) on Bayesian 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). 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 beliefs[1] one has about them. Thus uncertain quantities and their expectations are taken as primitive, and probability is defined in terms of expectation.

[1] See footnote 4 on p. 7.

As will be demonstrated in this chapter, this change of perspective radically alters the way we think about statistical inference, most notably by clarifying our objectives in the light of our (human) limitations; although the theorems are all the same. It gives us a naturalistic viewpoint from which to appraise modern statistical practice. Chapter 2 discusses modern practice in more detail.

1.1 Random quantities and their realms

My starting-point is a *random quantity*. A random quantity is a set of instructions which, if followed, will yield a real value; this is an *operational definition*. Experience suggests that thinking about random quantities is already hard enough, without having to factor in ambiguitities of definition—hence my insistence on operational definitions. Real-valued functions of random quantities are also random quantities.

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'). Thus XY represents the random quantity that arises when the instructions X and Y are both performed, and the resulting two values are multiplied together. Representative values of random quantities are denoted with small letters. I will write '$X \to x$' to represent 'instructions X were performed and the value x was the result'.

The *realm* of a random quantity is the set of possible values it might take; this is implicit in the instructions. 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] I write a collection of random quantities as $\boldsymbol{X} := (X_1, \ldots, X_m)$, and their joint realm as $\boldsymbol{\mathcal{X}}$, where $\boldsymbol{x} := (x_1, \ldots, x_m)$ is an element of $\boldsymbol{\mathcal{X}}$, and

$$\boldsymbol{\mathcal{X}} \subset \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \subset \mathbb{R}^m.$$

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

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

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

> 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)

For similar sentiments, see, e.g., Berger and Wolpert (1984, sec. 3.4), or Cox (2006, sec. 1.6). This is not just statistical parochialism. David Hilbert, one of the great mathematicians and a huge admirer of Cantor's work on non-finite sets, stated

> If we pay close attention, we find that the literature of mathematics is replete with absurdities and inanities, which can usually be blamed on the infinite.

and later in the same essay,

> [T]he infinite is not to be found anywhere in reality, no matter what experiences and observations or what kind of science we may adduce. Could it be, then, that thinking about objects is so unlike the events involving objects and that it proceeds so differently, so apart from reality? (Hilbert, 1926, p. 370 and p. 376 in the English translation)

The complications and paradoxes of the infinite are well-summarised in Vilenkin (1995).[3] I reckon the task of the statistician is hard enough, without having to grapple with an abstraction which has so consistently perplexed and bamboozled.

[3] Wallace (2003) is also worth a look. David Foster Wallace was a tremendous writer of fiction and essays, but this book displays the limitations of his literary style when writing about highly technical matters—also one has to acknowledge that he did not have sufficient mastery of his material.

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. Finally, for the purposes of statistical modelling we often introduce auxiliary random variables (e.g. statistical parameters) and these are conveniently represented with non-finite and unbounded realms.

So I will presume the following principle:

Definition 1.1 (Principle of Excluding Pathologies, PEP)**.**

Extensions to non-finite realms are made for the convenience of the statistician; it is the statistician's responsibility to ensure that such extensions do not introduce pathologies that are not present in the finite realm.

These notes consider random quantities with finite realms. But I have taken care to ensure that the results also apply, with minor amendments, in the more convenient (but less realistic) case of non-finite and even non-countable realms.

1.2 *Introduction to expectation*

Let X be a random quantity—under what conditions might I be said to 'know' X? Philosophers have developed a working definition for knowledge: knowledge is 'justified true belief' (Ladyman, 2002, pp. 5–6). So I would know X if I had carried out the instructions specified by X myself, or if they had been carried out by someone I trusted. In other circumstances—for example instructions that take place in the future—I have belief, but not knowledge.[4] Expectations and the expectations calculus are a way of quantifying and organising these beliefs, so that they hold together sensibly.

[4] I will consistently use 'belief' in favour of the of the more sober-sounding 'judgement', to honour this working definition of knowledge.

For concreteness, let X be sea-level rise by 2100, suitably operationalised. This is a random quantity about which no one currently has knowledge, and about which beliefs vary widely from person to person. When I consider my own beliefs about sea-level rise, I find I do not have a single value in mind. Instead, I have values, more or less nebulous, for quantities that I consider to be related to sea-level rise. So I believe, for example, that sea-level rise over the last century is of the order of 10's of centimetres. That the Greenland icesheet and the Western Antarctic icesheet each contain enough ice to raise sea-level globally by between 6 and 7 metres.

But that simulations suggest that they will not melt substantially by 2100. But I am cautious about the value of simulations of complex environmental systems. And a lot more things too: about people I know who work in this field, the degree of group-think in the field as a whole, the pressures of doing science in a field related to the effects of climate change, and so on. I do not have well-formed beliefs about sea-level rise, but it turns out that I have lots of ill-formed beliefs about things related to sea-level rise.

And if I wanted to I could easily acquire more beliefs: for example I could ask a glaciologist for her opinion. But once this was given, this would simply represent more related beliefs (my beliefs about her beliefs) to incorporate into my beliefs. And she will be facing exactly the same challenge as me, albeit with a richer set of beliefs about things related to sea-level rise.

I do not think there is any formal way to model the mental processes by which this collection of ill-formed beliefs about things related to sea-level rise get turned into a quantitative expression of my beliefs about sea-level rise. Ultimately, though, I can often come up with some values, even though I cannot describe their provenence. For sea-level rise by 2100, 80 cm from today seems about right to me. I could go further, and provide a range: unlikely to be less than 40 cm, or more than 350 cm, perhaps. These are unashamedly guesses, representing my ill-defined synthesis of my beliefs about things related to sea-level rise.[5] If you were the Mayor of London, you would be well-advised to consult someone who knows more about sea-level rise than I do. But you should not think that he has a better method for turning his beliefs into quantities than I do. Rather, he starts with a richer set of beliefs.

[5] And also representing more general aspects of my personality, such as risk aversion and optimism.

These considerations lead me to my first informal definition of an expectation.

Definition 1.2 (Expectation, informal)**.**
Let X be a random quantity. My expectation for X, denoted $\mathrm{E}(X)$, is a sensible guess for X which is likely to be wrong.

We will need to define 'sensible' in a way that is generally acceptable, in order for you to understand the conditions under which my expectation is formed (Sec. 1.3). I am using 'guess' to describe my ill-defined synthesis of my beliefs related to X. And I am stressing that it is common knowledge that my guess is likely to be wrong. I think this last point is important, because experts (e.g. glaciologists) may be reluctant to provide wrong guesses, preferring to say nothing at all. So let's get the wrongness out in the open. As the Mayor of London, I would much rather have the wrong guess of a glaciologist than the wrong guess of a statistician.

Now I am able to provide an informal definition of statistical inference. This definition is in the same vein as L.J. Savage's definition of 'statistics': "quantitative thinking about uncertainty as it affects scientific and other investigations" (Savage, 1960, p. 541), although adapted to the use of expectation as primitive, and to the

limitations of our beliefs.

Definition 1.3 (Statistical inference, informal)**.**
Statistical inference is checking that my current set of expectations is sensible, and extending this set to expectations of other random quantities.

Checking and extending are largely mechanical tasks. But there is also a reflexive element. I may well discover that if Y is some other random quantity, then my $\mathrm{E}(Y)$ based on my current set of expectations is not constrained to a single value, but may be an interval of possible values: I would say I was 'undecided' about $\mathrm{E}(Y)$. If this interval is intolerably wide (in the context for which I would like to know Y), then I must go back and reconsider my current set of expectations: could I refine them further, or augment them?

Statistical inference is discussed in more detail in Sec. 1.6 and Chapter 2. First, I clarify what I mean by 'sensible', and some of the properties that follow from it.

1.3 *Definition and simple implications*

The axioms given below (in Def. 1.4) are the standard axioms of expectation. In this respect they are the 'what' rather than the 'why'. For the 'why' I refer back to the previous section, and the informal definition of expectation in Def. 1.2. I interpret these axioms as a minimal characterisation of 'sensible'.

Definition 1.4 (Axioms of expectation)**.**
Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random quantities with finite realms. Then the expectations of X and Y must satisfy the following properties:

0. $\mathrm{E}(X) \in \mathbb{R}$ *exists and is unique,* *(existence)*
1. $\mathrm{E}(X) \geq \min \mathcal{X}$, *(lower boundedness)*
2. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$. *(additivity)*

You can see that this sets the bar on 'sensible' quite low—it is continuing a source of amazement to me that we can do so much with such simple beginnings. The 'existence' axiom does not insist that I know my expectation for every random quantity, but only that I acknowedge that it exists as a (real-valued) number and is unique. I use the word *undecided* to describe expectations that I am not currently able to quantify.

'Lower-boundedness' is an extremely weak condition, given that $\mathcal{X}$ ought to be inferrable from X itself, and have nothing to do with my particular beliefs about things related to X. For example, if X is the weight of this orange, then $\min \mathcal{X}$ must be $0\,\mathrm{g}$, to represent the physical impossibility of an orange with negative weight. I might believe that the weight cannot be less than $50\,\mathrm{g}$, but lower boundedness only requires that my $\mathrm{E}(X)$ is non-negative.

'Additivity' is a bit more subtle. I think we would all agree that if X and Y were the weights of two oranges, then anything other

than $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ would be not-sensible. But there are more interesting situations. Consider the following example, following Ellenberg (2014, ch. 11).[6] A man has seven children, and is planning to leave his £1m fortune to exactly one, the choice to be decided by the day of the week on which he dies. Let X_i be the amount in £m received by the ith child. The most likely outcome for each child is $X_i \to 0$. And yet $X_1 + \cdots + X_7 \to 1$ with certainty. And so to interpret $\mathrm{E}(X_i)$ as 'most likely' will not satisfy the additivity axiom. Most people in this case would take $\mathrm{E}(X_i) \leftarrow 1/7$ for each i, using a symmetry argument, and this would satisfy all three axioms. Mind you, $\mathrm{E}(X_1) \leftarrow 1$ and $\mathrm{E}(X_2) \leftarrow \cdots \leftarrow \mathrm{E}(X_7) \leftarrow 0$ would also satisfy all three axioms.

[6] This book is highly recommended, and would make an excellent Christmas present.

The asymmetric expectations in the seven children example illustrates the aforementioned point that the bar on 'sensible' is quite low. There is a strong case for introducing another word to mean precisely that the axioms are satisfied, so that 'sensible' does not seem misapplied. The standard choice among Bayesian statisticians is *coherent*, following de Finetti (1974/75). From now on I will use 'coherent' to describe a set of expectations satisfying Def. 1.4. In public discourse, when my expectations matter to people other than myself, I would use *defensible* to mean something more than simply coherent, although I hesitate to characterise this further, since it depends so much on context.

* * *

The axioms in Def. 1.4 have many implications. There are several reasons for considering these implications explicitly:

1. They give us confidence in the axioms, if they seem consistent with our interpretation of expectation.
2. They prevent us from making egregious specifications for expectations.
3. They provide a quick source of results when we assume that our beliefs are coherent.

Here I will just pick out a few of the basic implications, which are important enough to have names.

Theorem 1.1 (Implied by additivity alone)**.**

1. $\mathrm{E}(0) = 0$ *and* $\mathrm{E}(-X) = -\mathrm{E}(X)$,
2. $\mathrm{E}(X_1 + \cdots + X_k) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_k)$. *(finite additivity)*
3. $\mathrm{E}(aX) = a\,\mathrm{E}(X)$. *(linearity)*

Proof.

1. Since $0 = 0 + 0$, we have $\mathrm{E}(0) = 2\,\mathrm{E}(0)$ from which the result follows. The second result follows from $0 = X + (-X)$.

2. Follows iteratively from $X_1 + \cdots + X_k = X_1 + (X_2 + \cdots + X_k)$.

3. Here is the proof for rational a. If i is a non-negative integer, then $\mathrm{E}(iX) = i\,\mathrm{E}(X)$ by the previous result. And if j is a positive integer, then $\mathrm{E}(X) = \mathrm{E}(jX/j) = j\,\mathrm{E}(X/j)$ from which $\mathrm{E}(X/j) = \mathrm{E}(X)/j$. Hence $\mathrm{E}(aX) = a\,\mathrm{E}(X)$ whenever a is a non-negative rational number. Extend to $a < 0$ using $aX = |a|(-X)$.

The extension of the final part to real numbers is slightly subtle; see de Finetti (1974, footnote on p. 75). □

The *linearity* property is usually taken to subsume finite additivity, giving

$$\mathrm{E}(a_1X_1 + \cdots + a_kX_k) = a_1\,\mathrm{E}(X_1) + \cdots + a_k\,\mathrm{E}(X_k). \qquad \textit{(linearity)}$$

This is the property that must be strengthened in the case where there are a non-finite number of random quantities, or, which comes to the same thing, the realm of a random quantity is non-finite. The stronger *countable additivity* axiom extends finite additivity and (finite) linearity to countably-infinite sequences. This stronger axiom is almost universally accepted, as it ought to be according to the PEP (Def. 1.1).[7]

[7] The deep and mysterious book by Dubins and Savage (1965) is a notable exception.

Here are some further implications, using both additivity and lower-boundedness.

Theorem 1.2.

1. $\mathrm{E}(a) = a$, *(normalisation)*
2. *If* $X \leq Y$, *then* $\mathrm{E}(X) \leq \mathrm{E}(Y)$, *(montonicity)*
3. $\min \mathcal{X} \leq \mathrm{E}(X) \leq \max \mathcal{X}$ *(convexity)*
4. $|\mathrm{E}(X)| \leq \mathrm{E}(|X|)$. *(triangle inequality)*

Proof.

1. $a \geq a$, so $\mathrm{E}(a) \geq a$. And $-a \geq -a$, so $\mathrm{E}(-a) \geq -a$, and then $\mathrm{E}(-a) = -\mathrm{E}(a)$ implies that $\mathrm{E}(a) \leq a$; hence $\mathrm{E}(a) = a$.

2. The minimum of the realm of $Y - X$ is non-negative, hence $\mathrm{E}(Y - X) \geq 0$ which implies that $\mathrm{E}(X) \leq \mathrm{E}(Y)$.

3. Same argument as above, as X is never greater than $\max \mathcal{X}$, and $\mathrm{E}(\max \mathcal{X}) = \max \mathcal{X}$.

4. Same argument as above, as $-|X|$ is never greater than X, and X is never greater than $|X|$. Together these imply that $\mathrm{E}(X) \leq \mathrm{E}(|X|)$ and $-\mathrm{E}(X) \leq \mathrm{E}(|X|)$, as required.

□

Finally in this section, we have *Schwarz's inequality*, which is proved using linearity and monotonicity.

Theorem 1.3 (Schwarz's inequality)**.**

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

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

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

This quadratic in a cannot have two distinct real roots, because that would indicate a negative value for the expectation, violating monotonicity. Then it follows from the standard formula for the roots of a quadratic[8] that

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

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

[8] If $ax^2 + bx + c = 0$ then $x = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

Another similarly useful result is Jensen's inequality, which concerns the expectation of convex functions of random quantities. This result can also be proved at this stage using linearity and monotonicity, but only if we accept the Separating Hyperplane Theorem. Instead, I will defer Jensen's inequality until Sec. 1.5.2, at which point I will be able to give a self-contained proof.

1.3.1 Quantities related to expectation*

Here is a brief summary of other quantities that are defined in terms of expections, and their properties. These properties follow immediately from the axioms and are not proved.

If X is a random quantity with expectation μ, then the *variance* of X is defined as

$$\mathrm{Var}(X) := \mathrm{E}\{(X-\mu)^2\},$$

and often denoted σ^2; clearly $\sigma^2 \geq 0$ by monotonicity. Expanding out shows that

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2.$$

The square root of $\mathrm{Var}(X)$ is termed the *standard deviation*; I denote it as $\mathrm{Sd}(X)$. It has the same units as X, and is often denoted as σ. $\mathrm{Var}(a+bX) = b^2\,\mathrm{Var}(X)$, and $\mathrm{Sd}(a+bX) = b\,\mathrm{Sd}(X)$.

If X and Y are two random quantities with expectations μ and ν then the covariance of X and Y is defined as

$$\mathrm{Cov}(X,Y) := \mathrm{E}\{(X-\mu)(Y-\nu)\}.$$

Hence $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ and $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$. Expanding out shows that

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mu\nu.$$

$\mathrm{Cov}(a+bX, c+dY) = bd\,\mathrm{Cov}(X,Y)$, $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$, and, by iteration,

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j).$$

If $\mathrm{Cov}(X, Y) = 0$ then X and Y are *uncorrelated*. If $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$ then $(X_1, \ldots, X_n)$ are *mutually uncorrelated*. In this case

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_i \mathrm{Var}(X_i).$$

Hence, unlike expectation, variance is only additive for mutually uncorrelated random quantities. Schwartz's inequality implies that

$$\mathrm{Cov}(X, Y)^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y).$$

When both $\mathrm{Sd}(X)$ and $\mathrm{Sd}(Y)$ are positive, the *correlation* between X and Y is defined as

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\mathrm{Sd}(X)\,\mathrm{Sd}(Y)}.$$

It is unitless, and invariant to linear transformations of X and Y, i.e.

$$\mathrm{Corr}(X, Y) = \mathrm{Corr}(a + bX, c + dY),$$

and is often denoted ρ. Schwartz's inequality implies that

$$-1 \leq \mathrm{Corr}(X, Y) \leq 1,$$

with equality if and only if $Y = a + bX$.[9]

[9] Technically, this '=' should be interpreted as 'mean square equivalent', see Sec. 1.7.1.

1.4 *Probability*

If expectation is primitive, then probability is just a special type of expectation. In a nutshell, a probability is the expectation of the indicator function of a random proposition.

You may want to consult the material on first order logic in Sec. 1.A: in particular, the definition of a first order sentence on p. 36. This is the basis for the following definition.

Definition 1.5 (Random proposition)**.**
A random proposition is a first order sentence in which one or more constants have been replaced by random quantities.

In the simplest case, if x and y are constants then $x \doteq y$ is a first order sentence.[10] If X and Y are random quantities, then $X \doteq x$ and $X \doteq Y$ are random propositions. The truth value of a first order sentence is known, but the truth value of a random proposition is uncertain, because it contains random quantities instead of constants.

[10] The need to distinguish the symbol '$\doteq$' from '$=$' is explained in Sec. 1.A.

The *indicator function* of a first order sentence ψ is the function $\mathbb{1}_\psi$ for which

$$\mathbb{1}_\psi := \begin{cases} 0 & \psi \text{ is false} \\ 1 & \psi \text{ is true.} \end{cases}$$

In other words, the indicator function turns false into *zero* and true into *one*.[11] Note that the indicator function of a conjunction of sentences is the product of the indicator functions:

$$\mathbb{1}_{\psi \wedge \phi} = \mathbb{1}_\psi \cdot \mathbb{1}_\phi.$$

[11] I will also write $\mathbb{1}(\cdot)$ for more complicated random propositions.

The indicator function is used to define a probability.

Definition 1.6 (Probability)**.**
Let Q be a random propostion. Then $\Pr(Q) := \mathrm{E}(\mathbb{1}_Q)$.

So, continuing the example for the simplest case given above, $\Pr(X \doteq x) := \mathrm{E}(\mathbb{1}_{X \doteq x})$ and $\Pr(X \doteq Y) := \mathrm{E}(\mathbb{1}_{X \doteq Y})$. These probabilities are expectations of specified functions of the random quantities X and Y.

This definition of probability might seem strange to people used to treating probability as primitive. And so it is worth taking a moment to check that the usual axioms of probability are satisfied. Thus, if P and Q are random propositions:

1. $\Pr(P) \geq 0$, by lower-boundedness.
2. If P is a tautology, then $\mathbb{1}_P = 1$ and $\Pr(P) = 1$ by normalisation.
3. If P and Q are incompatible, i.e. $\mathbb{1}_{P \wedge Q} = 0$, then $\mathbb{1}_{P \vee Q} = \mathbb{1}_P + \mathbb{1}_Q$, and $\Pr(P \vee Q) = \Pr(P) + \Pr(Q)$, by linearity.

Thus all of the usual probability results apply; I will not give them here.

One very useful convention helps us to express probabilities of conjunctions efficiently. If $\{A_1, \ldots, A_k\}$ is a collection of random propositions, then define

$$\Pr(A_1, \ldots, A_k) := \Pr(A_1 \wedge \cdots \wedge A_k).$$

In other words, commas between random propositions represent conjunctions. I will return to this convention in Sec. 1.8.3.

1.4.1 Simple inequalities*

There are some simple inequalities linking expectations and probabilities, and these can be useful for providing bounds on probabilities, or for specifying beliefs about a random quantity that includes both probabilities of logical propositions about X and expectations of functions of X. The starting-point for many of these is *Markov's inequality*.

Theorem 1.4 (Markov's inequality)**.**
If X is non-negative and $a > 0$ then

$$\Pr(X \dot{\geq} a) \leq \frac{\mathrm{E}(X)}{a}.$$

Proof. Follows from monotonicity and linearity, because

$$a\, \mathbb{1}_{X \dot{\geq} a} \leq X,$$

see Figure 1.1. Taking expectations of both sides and rearranging gives the result. □

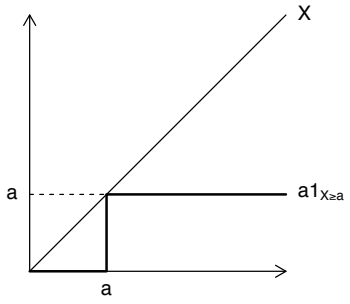

Figure 1.1: Markov's inequality.

One immediate generalisation of Markov's inequality is

$$\Pr(X \dot{\geq} a) \leq \frac{\mathrm{E}\{g(X)\}}{g(a)}$$

whenever g is a non-negative increasing function: this follows because $g(X)$ is non-negative and because $X \dot{\geq} a \iff g(X) \dot{\geq} g(a)$. A useful application of this generalisation is

$$\Pr(|X| \dot{\geq} a) \leq \min_{r>0} \frac{\mathrm{E}\{|X|^r\}}{|a|^r}$$

which follows because $|x|^r$ is a non-negative increasing function of $|x|$ for every positive r. A special case is *Chebyshev's inequality*. This is usually expressed in terms of $\mu := \mathrm{E}(X)$ and $\sigma^2 := \mathrm{E}\{(X-\mu)^2\}$ (see Sec. 1.3.1). Setting $r \leftarrow 2$ then gives

$$\Pr(|X-\mu| \dot{\geq} a) \leq \frac{\sigma^2}{a^2} \tag{1.1}$$

for $a > 0$.

1.5 The Fundamental Theorem of Prevision

The Fundamental Theorem of Prevision (FTP) is due to Bruno de Finetti (see de Finetti, 1974, sec. 3.10).[12] Its epithet 'fundamental' is well-deserved, because it provides a complete characterisation of the set of expectations that are consistent with the axioms of expectation given in Def. 1.4.

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

The following theorem uses the $(s-1)$-dimensional *unit simplex*, defined as

$$\mathbb{S}^{s-1} := \left\{ \boldsymbol{p} \in \mathbb{R}^s : p_j \geq 0 \text{ and } \textstyle\sum_j p_j = 1 \right\}. \tag{1.2}$$

Theorem 1.5 (Fundamental Theorem of Prevision, FTP)**.**
Let $\boldsymbol{X} := (X_1, \ldots, X_m)$ be any finite collection of random quantities (with finite realms) and let

$$\mathcal{X} := \left\{ \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(s)} \right\} \qquad \boldsymbol{x}^{(j)} \in \mathbb{R}^m,$$

be their joint realm. Then E *is a valid expectation if and only if there is a $\boldsymbol{p} \in \mathbb{S}^{s-1}$ for which*

$$\mathrm{E}\{g(\boldsymbol{X})\} = \sum_{j=1}^{s} g(\boldsymbol{x}^{(j)}) \cdot p_j \tag{†}$$

for all $g : \mathbb{R}^m \to \mathbb{R}$. In this case, $p_j = \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)})$.

Proof.

($\Leftarrow$). This is just a matter of checking that (†) satisfies the axioms in Def. 1.4. The zeroth axiom is obviously satisfied. Lower-boundedness follows from

$$\begin{aligned} \mathrm{E}\{g(\boldsymbol{X})\} &= \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot p_j \\ &\geq \min_{\boldsymbol{p} \in \mathbb{S}^{s-1}} \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot p_j = \min_j g(\boldsymbol{x}^{(j)}), \end{aligned}$$

as required. Additivity follows immediately from the linearity of (†). Let $g(\boldsymbol{x}) \leftarrow \mathbb{1}_{\boldsymbol{x} \doteq \boldsymbol{x}^{(i)}}$. Then

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(i)}) = \textstyle\sum_j \mathbb{1}_{\boldsymbol{x}^{(j)} \doteq \boldsymbol{x}^{(i)}} \cdot p_j = p_i,$$

as required.

(⇒). Note that

$$1 = \sum_{j=1}^{s} \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}, \qquad (\ddagger)$$

where $\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}$ denotes the conjunction $X_1 \doteq x_1^{(j)} \wedge \cdots \wedge X_m \doteq x_m^{(j)}$. Hence

$$\begin{aligned} \mathrm{E}\{g(\boldsymbol{X})\} &= \mathrm{E}\left\{g(\boldsymbol{X}) \textstyle\sum_j \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\right\} \\ &= \mathrm{E}\left\{ \textstyle\sum_j g(\boldsymbol{X}) \cdot \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\right\} \\ &= \mathrm{E}\left\{ \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\right\} \\ &= \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot \mathrm{E}\{\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\} \qquad \text{by linearity.} \end{aligned}$$

The result then follows on setting $p_j := \mathrm{E}\{\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\}$, as $p_j \geq 0$ by lower-boundedness, and $\sum_j p_j = 1$ by linearity and normalisation, from (‡). Hence $\boldsymbol{p} \in \mathbb{S}^{s-1}$.

□

Eq. (†) is familiar as the definition of an expectation in the case where probability is taken as primitive. In contrast, the FTP states that it is an inevitable consequence of the axioms of expectation that probabilities $\boldsymbol{p} \in \mathbb{S}^{s-1}$ must exist, satisfying (†).

1.5.1 Marginalisation

One immediate application of the FTP is in marginalisation, which is 'collapsing' a probability assessment onto a subset of random quantities.

Theorem 1.6 (Marginalisation)**.** *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two collections of random quantities. Then*

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \Pr(\boldsymbol{X} \doteq \boldsymbol{x}, \boldsymbol{Y} \doteq \boldsymbol{y})$$

where $\mathcal{Y}$ is the realm of $\boldsymbol{Y}$.

Removing $\boldsymbol{Y}$ in this way is termed *marginalising out* $\boldsymbol{Y}$.

Proof. Take $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ to be scalars, without loss of generality, and write

$$\mathcal{X} \times \mathcal{Y} = \left\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(1)}) \ldots, (x^{(s)}, y^{(t)})\right\},$$

where $s := \dim \mathcal{X}$ and $t := \dim \mathcal{Y}$. This product space may be a superset of the realm of (X, Y), but we can set $\Pr(X \doteq x, Y \doteq y) \leftarrow 0$ if (x, y) is not in the realm of (X, Y). From the FTP,

$$\mathrm{E}\{g(X, Y)\} = \sum_{i=1}^{s} \sum_{j=1}^{t} g(x^{(i)}, y^{(j)}) \cdot \Pr(X \doteq x^{(i)}, Y \doteq y^{(j)}),$$

for all g. Now set $g(x, y) \leftarrow \mathbb{1}_{x \doteq x'}$, and then

$$\begin{aligned}
\Pr(X \doteq x') &= \sum_i \sum_j \mathbb{1}_{x^{(i)} \doteq x'} \cdot \Pr(X \doteq x^{(i)}, Y \doteq y^{(j)}) \\
&= \sum_j \Big(\sum_i \mathbb{1}_{x^{(i)} \doteq x'} \cdot \Pr(X \doteq x^{(i)}, Y \doteq y^{(j)}) \Big) \\
&= \sum_j \Pr(X \doteq x', Y \doteq y^{(j)}) \\
&= \sum_{y \in \mathcal{Y}} \Pr(X \doteq x', Y \doteq y),
\end{aligned}$$

as required. □

1.5.2 *Jensen's inequality*

Jensen's inequality concerns the expectation of convex functions of $\boldsymbol{X}$. Recollect that a function $g : \mathbb{R}^k \to \mathbb{R}$ is a *convex function* exactly when

$$g\big(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{x}'\big) \leq \alpha g(\boldsymbol{x}) + (1 - \alpha) g(\boldsymbol{x}')$$

for all $0 \leq \alpha \leq 1$. Informally, the chord between any two points on $g(\boldsymbol{x})$ never goes below $g(\boldsymbol{x})$.

Theorem 1.7 (Jensen's inequality)**.**

Let $\boldsymbol{X} := (X_1, \ldots, X_m)$. If g is a convex function of $\boldsymbol{x}$, then $\mathrm{E}\{g(\boldsymbol{X})\} \geq g(\mathrm{E}\{\boldsymbol{X}\})$.

Proof. There is a conventional proof using the Separating Hyperplane Theorem, but I like the following proof based on the FTP and induction on s, the size of the realm of $\boldsymbol{X}$.

Denote the realm of $\boldsymbol{X}$ as $\mathcal{X} := \{\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(s)}\}$. According to the FTP, for each s

$$\mathrm{E}\{g(\boldsymbol{X})\} = \sum_{j=1}^{s} p_j^{(s)} \, g(\boldsymbol{x}^{(j)})$$

for some $\boldsymbol{p}^{(s)} := (p_1^{(s)}, \ldots, p_s^{(s)}) \in \mathbb{S}^{s-1}$. I'll drop the superscript on $\boldsymbol{p}$ to avoid clutter.

Now if g is convex and $s = 2$, then

$$\begin{aligned}
g(\mathrm{E}\{\boldsymbol{X}\}) &= g(p_1 \boldsymbol{x}_1 + p_2 \boldsymbol{x}_2) && \text{by the FTP} \\
&\leq p_1 \, g(\boldsymbol{x}_1) + p_2 \, g(\boldsymbol{x}_2) && \text{by convexity of } g \\
&= \mathrm{E}\{g(\boldsymbol{X})\} && \text{FTP again.}
\end{aligned}$$

This proves Jensen's inequality for the case $s = 2$.

Now suppose that Jensen's inequality is true for s, and consider the case $s + 1$. At least one of the p_j in $(p_1, \ldots, p_{s+1})$ must be positive, take it to be p_1. If $p_1 = 1$ then $g(\mathrm{E}\{\boldsymbol{X}\}) = g(x^{(1)}) = \mathrm{E}\{g(\boldsymbol{X})\}$

satisfying the theorem, so take $p_1 < 1$. Then

$$\begin{aligned}
g(\mathrm{E}\{\boldsymbol{X}\}) &= g\Big(\sum_{j=1}^{s+1} p_j \boldsymbol{x}_j\Big) && \text{by the FTP} \\
&= g\Big(p_1 \boldsymbol{x}_1 + (1-p_1)\sum_{j=2}^{s+1} q_j \boldsymbol{x}_j\Big) && \text{where } q_j := p_j/(1-p_1) \\
&\leq p_1\, g(\boldsymbol{x}_1) + (1-p_1)\, g\Big(\sum_{j=2}^{s+1} q_j \boldsymbol{x}_j\Big) && \text{by convexity of } g \\
&\leq p_1\, g(\boldsymbol{x}_1) + (1-p_1)\sum_{j=2}^{s+1} q_j\, g(\boldsymbol{x}_j) && \text{Jensen's inequality holds for } s \\
&= \sum_{j=1}^{s+1} p_j\, g(\boldsymbol{x}_j) = \mathrm{E}\{g(\boldsymbol{X})\} && \text{FTP again,}
\end{aligned}$$

where the Jensen's inequality line uses $(q_2, \ldots, q_{s+1}) \in \mathbb{S}^{s-1}$. □

Jensen's inequality is the basis for the very powerful *Gibbs's inequality*. This will appear in Sec. 4.7.

Theorem 1.8 (Gibbs's inequality)**.**
Let $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{S}^{k-1}$. *Then*

$$\sum_{j=1}^{k} p_j \log(p_j/q_j) \geq 0,$$

and is zero if and only if $\boldsymbol{p} = \boldsymbol{q}$.

Proof.

$$\begin{aligned}
\textstyle\sum_j p_j \log(p_j/q_j) &= \textstyle\sum_j p_j(-\log(q_j/p_j)) \\
&\geq -\log\big(\textstyle\sum_j p_j \cdot q_j/p_j\big) && \text{Jensen's inequality} \\
&= -\log\big(\textstyle\sum_j q_j\big) \\
&= 0,
\end{aligned}$$

where Jensen's inequality applies because the sum over j is an expectation according to the FTP, and $-\log(x)$ is convex. For 'only if', fix $\boldsymbol{q}$, and then note that $\sum_{j=1}^{k} p_j \log(p_j/q_j)$ is strictly convex in $\boldsymbol{p}$, and hence the minimum is unique. □

1.6 Coherence and extension

Now I review the definition of statistical inference, stated informally in Def. 1.3 as covering *coherence* and *extension*. For any particular application, I identify a relevant set of random quantities, $\boldsymbol{X} := (X_1, \ldots, X_m)$. I have beliefs about these quantities, encoded as expectations of a set of functions of $\boldsymbol{X}$. I would like to check that this set of expectations is coherent. Then I would like to use these expectations to constrain my expectations of another set of functions of $\boldsymbol{X}$. In other words, I want to extend the expectations I

have to expectations about which I am currently undecided. So if I am currently undecided about my $\mathrm{E}\{h(\boldsymbol{X})\}$, I would like to know the subset of $\mathbb{R}$ that represents values of $\mathrm{E}\{h(\boldsymbol{X})\}$ that are coherent with my current expectations. I can also ask more general questions about collections of expectations.

1.6.1 The FTP again

The results in this section are immediate applications of the FTP (Sec. 1.5). The first result concerns the coherence of my current set of expectations. Recollect that $\mathbb{S}^{s-1}$ is the $(s-1)$-dimensional unit simplex, defined in (1.2).

Theorem 1.9 (Coherence of expectations)**.**
Let $\boldsymbol{X} := (X_1, \ldots, X_m)$ and let

$$\mathfrak{X} := \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(s)}\right\} \qquad \boldsymbol{x}^{(j)} \in \mathbb{R}^m$$

be their joint realm. Let $\{g_1, \ldots, g_k\}$ be a set of real-valued functions, and let G be the $(k \times s)$ matrix with $G_{ij} := g_i(\boldsymbol{x}^{(j)})$. Then the set of expectations

$$\mathrm{E}\{g_1(\boldsymbol{X})\} \leftarrow v_1, \ldots, \mathrm{E}\{g_k(\boldsymbol{X})\} \leftarrow v_k$$

is coherent if and only if the linear equations

$$G\boldsymbol{p} = \boldsymbol{v}$$

have a solution $\boldsymbol{p} \in \mathbb{S}^{s-1}$, where $\boldsymbol{v} := (v_1, \ldots, v_k)$.

Proof. This is just the matrix expression for the FTP as stated in Thm 1.5, with each row representing the equality

$$\mathrm{E}\{g_i(\boldsymbol{X})\} = \textstyle\sum_j g_i(\boldsymbol{x}^{(j)}) \cdot p_j = v_i.$$

The FTP must hold for all $g : \mathbb{R}^m \to \mathbb{R}$, and is if-and-only-if. □

The second result concerns the set of values for my expectations of other functions of $\boldsymbol{X}$ that is coherent with my current set of expectations.

Theorem 1.10 (Extension of expectations)**.**
Let $\{h_1, \ldots, h_n\}$ be another set of real-valued functions of $\boldsymbol{x}$, and denote $H_{ij} := h_i(\boldsymbol{x}^{(j)})$. Then the set of coherent values for my expectations of $h_1(\boldsymbol{X}), \ldots, h_n(\boldsymbol{X})$ is

$$\mathcal{H} := \left\{\boldsymbol{h} \in \mathbb{R}^n : \boldsymbol{h} = H\boldsymbol{p} \text{ for some } \boldsymbol{p} \in \mathbb{S}^{s-1} \text{ satisfying } G\boldsymbol{p} = \boldsymbol{v}\right\}.$$

Proof. Because, again by the FTP, $[\boldsymbol{v}, \boldsymbol{h}]$ is a valid set of expectations if and only if

$$\begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p} = \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{h} \end{bmatrix}$$

for some $\boldsymbol{p} \in \mathbb{S}^{s-1}$. □

The coherence and extension steps can be combined, because $\mathcal{H}$ will be non-empty if and only if $[G, \boldsymbol{v}]$ is coherent.

Here is one very important property of the set of coherent extensions. Informally, it states that the coherent set for any undecided expectation is an interval; i.e. completely defined by lower and upper bounds. Recollect that a set S is convex exactly when $s, s' \in S$ implies that $\alpha s + (1-\alpha)s' \in S$ for all $0 \leq \alpha \leq 1$.

Theorem 1.11. *The set $\mathcal{H}$ is convex.*

Proof. Empty sets are convex, so let $\mathcal{H}$ be non-empty, and let $\boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}$. Now consider the new point

$$\boldsymbol{h}'' := \alpha \boldsymbol{h} + (1-\alpha)\boldsymbol{h}'$$

for some $\alpha \in (0,1)$. Then

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{h}'' \end{bmatrix} &= \alpha \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{h} \end{bmatrix} + (1-\alpha) \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{h}' \end{bmatrix} \\
&= \alpha \begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p} + (1-\alpha) \begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p} \quad \text{for some } \boldsymbol{p} \in \mathbb{S}^{(s-1)}, \text{ because } \boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H} \\
&= \begin{bmatrix} G \\ H \end{bmatrix} \left(\alpha \boldsymbol{p} + (1-\alpha)\boldsymbol{p}\right) \\
&= \begin{bmatrix} G \\ H \end{bmatrix} \boldsymbol{p},
\end{aligned}$$

showing that $\boldsymbol{h}'' \in \mathcal{H}$. □

1.6.2 *Representing beliefs*

Suppose I am satisfied that my beliefs $[G, \boldsymbol{v}]$ are coherent, and I am now considering their extension to some new random quantity $h(\boldsymbol{X})$. The best possible outcome is to find that my set of coherent values for $\mathrm{E}\{h(\boldsymbol{X})\}$ is constrained to a single point; in other words, my expectation of $h(\boldsymbol{X})$ is completely constrained by my expectations for $g_1(\boldsymbol{X}), \ldots, g_k(\boldsymbol{X})$. This can arise in the obvious way: for example, where $g_1(\boldsymbol{x}) := x_1$, $g_2(\boldsymbol{x}) := x_2$, and $h(\boldsymbol{x}) := x_1 + x_2$. But it can also arise in much less obvious ways, involving the interplay of the more subtle constraints that are represented by the theorems of the expectations calculus. Because these theorems follow directly from the axioms, they are automatically enacted in the FTP. Thus $\mathcal{H}$ must respect Schwartz's inequality, Jensen's inequality, Markov's inequality, and so on. Expectations for a rich set of g_i's will have many more implications for the nature of $\mathcal{H}$ than I can easily envisage, and computation is the only method I have to infer them all. Computation is briefly discussed in Sec. 1.6.3.

In general, however, we must accept that many of my expectations will not be constrained to a point, i.e. I will remain undecided about my $\mathrm{E}\{h(\boldsymbol{X})\}$. Thm 1.11 states that my set of coherent expectations for $\mathrm{E}\{h(\boldsymbol{X})\}$ can be represented by an interval, and defined

in terms of lower and upper bounds. It is important to present this clearly. For example, to state "My expectation for $X_1 + 2\log X_2$ is *undecided* but lies in the interval $[3.2, 5.5]$." This is because there are advocates of a more general calculus of expection, who propose that my beliefs about the $g_i(\boldsymbol{X})$'s may themselves be expressed in terms of intervals (see, e.g., Walley, 1991; Troffaes and de Cooman, 2014). So I would like the word 'undecided' to indicate a technical meaning associated with a purely mechanical derivation from a coherent set of specified expectations.

A wide range of beliefs can be encoded as expectations, and we should look beyond obvious beliefs such as $\mathrm{E}(X_1) \leftarrow v_1$. As discussed in Sec. 1.4, probabilities are also expectations, so each probability I specify constitutes a row of $[G, \boldsymbol{v}]$. For example, suppose that $q(\boldsymbol{x})$ is a first-order sentence, so that $Q := q(\boldsymbol{X})$ is a random proposition. If I think that Q has probability p_q then this is represented by a row of $[G, \boldsymbol{v}]$ with

$$G_{ij} \leftarrow \mathbb{1}_{q(\boldsymbol{x}^{(j)})} \quad \text{and} \quad v_i \leftarrow p_q.$$

Certainty is a special case: a random proposition to which I assign probability 1. If I am certain that Q is true, i.e. $p_q \leftarrow 1$, then this has the effect of zeroing those p_j for which $q(\boldsymbol{x}^{(j)})$ is false. So the same effect could be achieved by removing from $\mathfrak{X}$ all of the elements for which $q(\boldsymbol{x}^{(j)})$ is false.

Beyond certainty, there are a number of ways I could represent my belief that X_1 is close to w. Perhaps the simplest of these is to add the row

$$\mathrm{E}\{(X_1 - w)^2\} \leftarrow v,$$

where both w and v must specified. Then $v \leftarrow 0$ is another way to implement the special case of certainty about X_1, and a positive value of v indicates uncertainty. If I also add $\mathrm{E}(X_1) \leftarrow w$ then the value v is my variance for X_1, and Chebyshev's inequality (eq. 1.1) can be used to think about my uncertainty about X_1 in terms of probabilities, if this is helpful.

A variant on this approach can be used to implement measurement error. For example, suppose that X_2 is a measurement on X_1 which is known to be accurate to within $\pm v$. This can be implemented by adding the row

$$\Pr(|X_1 - X_2| \leq v) \leftarrow 1. \tag{†}$$

If I then learn the value of the observation, i.e. $X_2 \rightarrow w$, this becomes another additional row for $\mathrm{E}\{(X_2 - w)^2\} \leftarrow 0$; or else the realm of $\boldsymbol{X}$ is thinned, as described above. If I am uncertain about the accuracy of the measurement, then this too can be represented by a random quantity, say X_3, which would replace v in (†). X_3 might appear in many rows of $[G, \boldsymbol{v}]$, if the same type of instrument was being used to take many measurements. In this way, the values of the measurements will also constrain my expectation for functions of X_3, such as the standard deviation of X_3 (see Sec. 1.3.1).

In summary, expectations provide a rich framework for representing such beliefs about $\boldsymbol{X}$ as I feel able to specify. But there are computational difficulties, as discussed in the next subsection.

1.6.3 *Computation*

Consider the case of finding the lower bound on $\mathrm{E}\{h(\boldsymbol{X})\}$ for some specified function h, based on beliefs $[G, \boldsymbol{v}]$. We must solve

$$\min_{\boldsymbol{p}\in\mathbb{R}^s} \boldsymbol{h}^T\boldsymbol{p} \quad \text{subject to} \quad \begin{cases} G\boldsymbol{p} = \boldsymbol{v} \\ \sum_j p_j = 1 \\ p_j \geq 0 \quad j = 1, \ldots, s \end{cases}$$

where $\boldsymbol{h} := (h^{(1)}, \ldots, h^{(s)})$ and $h^{(j)} := h(\boldsymbol{x}^{(j)})$. This is a *linear programming (LP)* problem. LP represents one of the pinnacles of computer-based optimisation, discussed in Nocedal and Wright (2006, chh. 13 and 14).

Unfortunately, however, even modern linear programming methods will grind to a halt if s, the size of the joint realm of $\boldsymbol{X}$, is too large. And because s is exponential in the number of random quantities, it only takes a few random quantities before this happens. This is a tragedy for statistical inference as I have presented it here, because our inability to do the computations forces us down another route which provides a very different framework for specifying beliefs, one in which almost all of our limitations as uncertainty assessors is suppressed. This alternative framework is discussed in detail in Chapter 2.

But I believe it is valuable to explore how we *ought* to do statistical inference, and then to encounter the practical difficulties, in order to understand better why in practice we do statistical inference the way we do. I hazard that most people who work with uncertainty are not aware that there is a rich calculus of expectation that allows me to specify just as many beliefs as I feel able, and represents the results in terms of 'undecided' intervals for those expectations that I am unable to specify. It is true that in many applications these unaware people are not disadvantaged, because the implementation of such a calculus is computationally impractical. But even then it is important to know that there is a substantial gulf between what one ought to do, and what one ends up doing.

1.7 *Conditional expectation*

Conditional expectations allow me to access another type of belief, different from expectations but which can nonetheless be expressed in terms of expectations. In the terms of Sec. 1.6, they allow me to add new rows to $[G, \boldsymbol{v}]$. This section and Sec. 1.8 present the definition and properties of conditional expectation and conditional probability.

The practically important new concept in this section (and the next) is a *hypothetical expectation*, written $\mathrm{E}(X \mid Q)$. This is my

expectation of a random quantity X 'supposing Q to be true', where Q is a random proposition which might be either true or false. Hypothetical expectations give us a much wider palette for specifying our beliefs, allowing us to exercise our imagination and to play out different scenarios.[13] In scientific modelling, they allow us to incorporate notions of cause and effect. In a simulation of future sea level, for example, $Q_1, Q_2, \ldots$ might be different boundary conditions, representing different scenarios for future greenhouse gas emissions.

[13] Although we should be aware of what Kahneman (2011) calls the 'narrative fallacy'. This is another highly recommended book.

This section is about the 'plumbing' that gets us to $\mathrm{E}(X \mid Q)$, and to other useful quantities besides. But the big picture is this. We develop an intuitive understanding of $\mathrm{E}(X \mid Q)$ which allows me to assign it a value on the basis of my beliefs about things relevant to X and Q, say $\mathrm{E}(X \mid Q) \leftarrow w$. But we also prove that

$$\mathrm{E}(X\mathbb{1}_Q) = \mathrm{E}(X \mid Q)\,\mathrm{Pr}(Q). \tag{1.3}$$

Together, my w and this formula provide a new row for my $[G, v]$ as follows. Use the FTP to write out the expectation on the left of (1.3) and the probability on the right, to give

$$\textstyle\sum_j x^{(j)}\, \mathbb{1}_{q(x^{(j)})} \cdot p_j = w \sum_j \mathbb{1}_{q(x^{(j)})} \cdot p_j$$

where I am simplifying by assuming just one random quantity (without loss of generality), and where $q(x)$ is a first order sentence, and $Q := q(X)$. Then rearrange to give

$$\textstyle\sum_j \left(x^{(j)} - w\right) \mathbb{1}_{q(x^{(j)})} \cdot p_j = 0$$

which is a row of $[G, v]$ with

$$G_{ij} \leftarrow (x^{(j)} - w)\, \mathbb{1}_{q(x^{(j)})} \quad \text{and} \quad v_i \leftarrow 0.$$

This is the key thing to appreciate: conditional expectations allows me to make another type of belief assessment, which can be used to constrain my expectations of other random quantities.

1.7.1 Types of equivalence*

This subsection is a detour to motivate a particular choice of loss function in Sec. 1.7.2.

Two random quantities X and Y can be equivalent: we inspect their operational definitions and conclude that the value which results is always the same. But there are also weaker forms of equivalence, where the operational definitions may be different, but not practically different. One way to capture this notion is in the following definition.

Definition 1.7 (Effectively equivalent)**.**
Random quantities X and Y are effectively equivalent *exactly when*

$$\mathrm{E}\{g(X, Z)\} = \mathrm{E}\{g(Y, Z)\}$$

for all g and all Z.

In this case, any conceivable inference involving X would give the same result if X was replaced by Y, and *vice versa*.[14]

Here is another way to capture the notion that X and Y are not practically different: this way is mathematically much more tractable.

Definition 1.8 (Mean-square equivalent)**.**

Random quantities X and Y are mean-square equivalent*, written $X \overset{\text{ms}}{=} Y$, exactly when*

$$\mathrm{E}\{(X-Y)^2\} = 0.$$

What is perhaps surprising is that these two definitions are equivalent.

Theorem 1.12. *X and Y are effectively equivalent if and only if they are mean-square equivalent.*

Proof. This proof passes through the FTP. First, if X and Y are effectively equivalent then they are mean-square equivalent, as can be seen by setting $g(x,z) \leftarrow xz$ and setting $z \leftarrow (x-y)$.

Now suppose that X and Y are mean-square equivalent. The FTP implies that

$$\mathrm{E}\{(X-Y)^2\} = \sum_{i,j,k} (x^{(i)} - y^{(j)})^2 \cdot p_{ijk}$$

where $p_{ijk} = \Pr(X \doteq x^{(i)}, Y \doteq y^{(j)}, Z \doteq z^{(k)})$. Since this expectation must equal zero, it follows that

$$p_{ijk} = 0 \qquad \text{whenever } x^{(i)} \neq y^{(j)}.$$

Hence, for arbitrary g,

$$\begin{aligned}\mathrm{E}\{g(X,Z)\} &= \sum_{i,j,k} g(x^{(i)}, z^{(k)}) \cdot p_{ijk} \\ &= \sum_{i,j,k} g(y^{(j)}, z^{(k)}) \cdot p_{ijk} \\ &= \mathrm{E}\{g(Y,Z)\},\end{aligned}$$

i.e. X and Y are effectively equivalent. □

The characterisation of conditional expectation in the next subsection is based on mean-square equivalence, but it is important to appreciate that this mathematically tractable property is equivalent to the more intuitive property that X and Y are not practically different.

1.7.2 *Characterisation and definition*

This is an advanced treatment of conditional expectation, along the lines laid down by the great Soviet mathematician Andrey Kolmogorov in the 1930s.[15] A simpler treatment would be possible in this chapter, in which I treat all realms as finite (see Sec. 1.1); but this does not generalise easily to the cases we often use in

[14] This definition and the following results generalise immediately to the case where $\boldsymbol{Z}$ is any finite collection of random quantities.

[15] In his 1933 book *Foundations of the Theory of Probability*. According to Grimmett and Stirzaker (2001, p. 571), Kolmogorov wrote this book to help pay for the repairs to the roof of his *dacha*.

practice, in which realms may be non-finite and even uncountably infinite. It is important to appreciate that conditioning on random quantities which have uncountable realms is well-defined, despite the (elementary) textbook prohibition on conditioning on random propositions which have probability zero.

In this section and the next I will assume that all random quantities are *mean-square integrable*, i.e. $\mathrm{E}(X^2)$ is finite for any random quantity X. This will always be the case when realms are finite, and so I am entitled to make this assumption according to the PEP (Def. 1.1). In fact, with more advanced tools we can relax this condition to *absolutely integrable*, i.e. $\mathrm{E}(|X|)$ is finite. But there is much more intuition in the former case.

A conditional expectation addresses the question "How might I represent some random quantity X in terms of some other random quantities $\boldsymbol{Y} := (Y_1, \ldots, Y_n)$?" The idea is for me to make a new random quantity based on $\boldsymbol{Y}$ that I believe is as close as possible to X. My representation will be some function $g : \mathcal{Y} \to \mathbb{R}$, and the very best that I can hope for is that X and $g(\boldsymbol{Y})$ are not materially different, or that

$$\mathrm{E}\left[\{X - g(\boldsymbol{Y})\}^2\right] = 0,$$

according to Thm 1.12. It would be unusual if I could find a g that achieves this lower bound, but I can aim for it; which suggests that when representing X in terms of $\boldsymbol{Y}$ I should envisage a function ψ which minimises the expected squared difference.

This optimality property characterises the function ψ, but the following equivalence result provides a much more tractable representation.

Theorem 1.13 (Projection theorem)**.**
The following two statements are equivalent:

A. $\mathrm{E}\left[\{X - \psi(\boldsymbol{Y})\}^2\right] \leq \mathrm{E}\left[\{X - g(\boldsymbol{Y})\}^2\right]$ *for all* g*,*

B. $\mathrm{E}\left[\{X - \psi(\boldsymbol{Y})\}\, g(\boldsymbol{Y})\right] = 0$ *for all* g*.*

Furthermore, if ψ *and* ψ' *are two solutions to (A) or (B), then* $\psi(\boldsymbol{Y}) \overset{\text{ms}}{=} \psi'(\boldsymbol{Y})$*.*

Proof.
(A) $\Rightarrow$ (B). Suppose that ψ satisfies (A) and let g be a perturbation away from ψ, i.e. $g(\boldsymbol{y}) \leftarrow \psi(\boldsymbol{y}) + \varepsilon\, h(\boldsymbol{y})$ for arbitrary ε and h. Then

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

But as ψ is a minimum no matter what sign ε has, we must have

$$\mathrm{E}\left[\{X - \psi(\boldsymbol{Y})\}\, h(\boldsymbol{Y})\right] = 0 \quad \text{for all } h, \tag{1.4}$$

which is (B).

(B) $\Rightarrow$ (A). We have

$$\begin{aligned} \mathrm{E}\left[\{X - g(\boldsymbol{Y})\}^2\right] &= \mathrm{E}\left[\{X - \psi(\boldsymbol{Y}) + \psi(\boldsymbol{Y}) - g(\boldsymbol{Y})\}^2\right] \\ &= \mathrm{E}\left[\{X - \psi(\boldsymbol{Y})\}^2\right] + \mathrm{E}\left[\{\psi(\boldsymbol{Y}) - g(\boldsymbol{Y})\}^2\right] \qquad (\dagger) \\ &\geq \mathrm{E}\left[\{X - \psi(\boldsymbol{Y})\}^2\right] \end{aligned}$$

where the cross-product terms in (†) are zero if (B) is true.

For the final statement, set $g \leftarrow \psi'$ in (†), to see that if

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

then $\mathrm{E}\left[\{\psi(\boldsymbol{Y}) - \psi'(\boldsymbol{Y})\}^2\right] = 0$, i.e. $\psi(\boldsymbol{Y}) \overset{\text{ms}}{=} \psi'(\boldsymbol{Y})$. □

The definition and notation of conditional expectation all originates from Thm 1.13.

Definition 1.9 (Conditional expectation)**.**
Let $\mathcal{E}(X \mid \boldsymbol{Y})$ denote the set of functions $\psi : \mathcal{Y} \to \mathbb{R}$ for which

$$\mathrm{E}\left[\{X - \psi(\boldsymbol{Y})\}\, g(\boldsymbol{Y})\right] = 0 \quad \textit{for all } g. \tag{1.5}$$

Then $\mathbb{E}(X \mid \boldsymbol{Y})$ is defined to be any member of the set of random quantities

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

Each $\mathbb{E}(X \mid \boldsymbol{Y})$ is termed a version *of the conditional expectation of X given $\boldsymbol{Y}$.*

So, to summarise, a conditional expectation $\mathbb{E}(X \mid \boldsymbol{Y})$ is a random quantity, and it is not uniquely defined, although all such conditional expectations are mean-square equivalent. I am using a double-barred '$\mathbb{E}$' to indicate a conditional expectation.

Multiple versions of $\mathbb{E}(X \mid \boldsymbol{Y})$ arise whenever there are elements of $\mathcal{Y}$ for which $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) = 0$. The following result makes this clear.

Theorem 1.14. *Let ψ and ψ' be two elements of $\mathcal{E}(X \mid \boldsymbol{Y})$. If $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) > 0$, then $\psi(\boldsymbol{y}) = \psi'(\boldsymbol{y})$.*

Proof. From the FTP, we have

$$\mathrm{E}\left[\{\psi(\boldsymbol{Y}) - \psi'(\boldsymbol{Y})\}^2\right] = \sum\nolimits_{\boldsymbol{y}} \{\psi(\boldsymbol{y}) - \psi'(\boldsymbol{y})\}^2 \cdot \Pr(\boldsymbol{Y} \doteq \boldsymbol{y}).$$

But $\psi(\boldsymbol{Y}) \overset{\text{ms}}{=} \psi'(\boldsymbol{Y})$ according to Thm 1.13, and hence this expectation must be zero, which implies that $\psi(\boldsymbol{y}) - \psi'(\boldsymbol{y}) = 0$ whenever $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) > 0$. □

This result gives us an inkling of why probability theory gets so complicated when realms become infinite and non-countable. In this case there may be no $\boldsymbol{y} \in \mathcal{Y}$ for which $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) > 0$. That is not to say that conditional expectations are not well-defined—they are perfectly well-defined, but there are *lots* of them, and no conditional expectation (or conditional probability) is unambiguously defined. If you are a student, this may have been concealed from you up until now. We will stick with finite realms, but we must still deal with the possibility that $\mathcal{E}(X \mid \boldsymbol{Y})$ contains more than one element, and hence that there is more than one version of $\mathbb{E}(X \mid \boldsymbol{Y})$, even if they are all mean-square equivalent.

This notion that the conditional expectation is a random quantity is not consistent with the conventional interpretation of hypothetical expectations, where $\mathrm{E}(X \mid Q)$ is a value when Q is a random

proposition such as $Y \doteq 3$. But this is attributable to the difference between conditioning on a set of random quantities and conditioning on a random proposition. It is important to be clear that Q is a random proposition, but $\mathbb{1}_Q$ is a random quantity.

Definition 1.10 (Hypothetical expectation)**.**
If X is a random quantity and Q is a random proposition, then

$$\mathrm{E}(X \mid Q) := \phi(1)$$

for any $\phi \in \mathcal{E}(X \mid \mathbb{1}_Q)$.

This makes $\mathrm{E}(X \mid Q)$ a value with the meaning 'the expectation of X conditional on Q being true', which is why I term it a *hypothetical expectation*. The definition might seem ambiguous, given that $\mathcal{E}(X \mid \mathbb{1}_Q)$ may contain many elements, but for the following result.

Theorem 1.15. *If $\Pr(Q) > 0$ then $\mathrm{E}(X \mid Q)$ is uniquely defined.*

Proof. Follows directly from Def. 1.10 and Thm 1.14, because $\Pr(Q) = \Pr(\mathbb{1}_Q \doteq 1)$. □

Therefore the purpose of adding the rider 'provided that $\Pr(Q) > 0$' to hypothetical expectations such as $\mathrm{E}(X \mid Q)$ is not because such things do not exist otherwise—they do exist, but unless $\Pr(Q) > 0$ they are not unique.

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 Def. 1.9,
$\mathbb{E}(X \mid \boldsymbol{Y})$: Any member of a set of random quantities, defined in Def. 1.9,
$\mathrm{E}(X \mid Q)$: A value, unique if $\Pr(Q) > 0$, defined in Def. 1.10.

1.7.3 Explicit formulas

As we are taking all realms to be finite, we can find a precise expression for the nature of the functions in $\mathcal{E}(X \mid \boldsymbol{Y})$. The expression also works in the more general case of non-finite and possibly unbounded realms, although only with a stronger additivity axiom (see p. 11).

Theorem 1.16. *Let $\psi \in \mathcal{E}(X \mid \boldsymbol{Y})$ where $\boldsymbol{Y}$ has a finite realm. Then*

$$\mathrm{E}(X\, \mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}}) = \psi(\boldsymbol{y}) \Pr(\boldsymbol{Y} \doteq \boldsymbol{y})$$

for each $\boldsymbol{y} \in \mathcal{Y}$.

Proof. I will write Y for $\boldsymbol{Y}$, likewise y for $\boldsymbol{y}$, to avoid too much ink on the page. Let the realm of Y be $\mathcal{Y} := \{y^{(1)}, \ldots, y^{(r)}\}$. Any function of y can be written as

$$g(y) = \textstyle\sum_i \alpha_i \mathbb{1}_{y \doteq y^{(i)}} \quad \text{for some } (\alpha_1, \ldots, \alpha_r) \in \mathbb{R}^r. \qquad (\dagger)$$

We want to find the β's for which

$$\psi(y) = \textstyle\sum_j \beta_j \mathbb{1}_{y \doteq y^{(j)}},$$

and then $\psi(y^{(i)}) = \beta_i$. From the Projection Theorem (Thm 1.13), ψ must satisfy

$$\mathrm{E}\left\{\left[X - \psi(Y)\right] g(Y)\right\} = 0 \quad \text{for all } g.$$

From (†), this is true if and only if ψ satisfies

$$\mathrm{E}\left\{\left[X - \psi(Y)\right] \mathbb{1}_{Y \doteq y^{(i)}}\right\} = 0 \quad i = 1, \ldots, r.$$

Substituting for ψ and then multiplying out gives

$$\mathrm{E}(X\mathbb{1}_{Y \doteq y^{(i)}}) - \textstyle\sum_j \beta_j \, \mathrm{E}(\mathbb{1}_{Y \doteq y^{(j)}} \, \mathbb{1}_{Y \doteq y^{(i)}}) = 0 \quad i = 1, \ldots, r.$$

But

$$\mathbb{1}_{Y \doteq y^{(j)}} \, \mathbb{1}_{Y \doteq y^{(i)}} = \begin{cases} \mathbb{1}_{Y \doteq y^{(i)}} & i = j \\ 0 & \text{otherwise} \end{cases}$$

and so we get

$$\mathrm{E}(X\mathbb{1}_{Y \doteq y^{(i)}}) - \beta_i \, \mathrm{E}(\mathbb{1}_{Y \doteq y^{(i)}}) = 0 \quad i = 1, \ldots, r. \qquad (‡)$$

Substituting $\psi(y^{(i)})$ for β_i and $\Pr(Y \doteq y^{(i)})$ for $\mathrm{E}(\mathbb{1}_{Y \doteq y^{(i)}})$ gives the displayed equation in Thm 1.16, which holds for all $i = 1, \ldots, r$, i.e. all $y \in \mathcal{Y}$. □

Clearly if $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}^{(i)}) > 0$ then the equality in Thm 1.16 can be rearranged to provide a unique value for $\psi(\boldsymbol{y}_i)$. Otherwise, if $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}^{(i)}) = 0$, then by Schwarz's inequality (Thm 1.3)

$$\{\mathrm{E}(X\mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}^{(i)}})\}^2 \leq \mathrm{E}(X^2) \, \mathrm{E}(\mathbb{1}^2_{\boldsymbol{Y} \doteq \boldsymbol{y}^{(i)}}) = \mathrm{E}(X^2) \, \Pr(Y \doteq y^{(i)}) = 0.$$

Hence $\mathrm{E}(X\mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}^{(i)}}) = 0$ and (‡) has the form $0 - \beta^{(i)} \cdot 0 = 0$, and so the value of β_i, i.e. $\psi(\boldsymbol{y}^{(i)})$, is arbitrary.

The next two results follow directly from Thm 1.16. The first result gives an explicit expression for the hypothetical expectation $\mathrm{E}(X \mid Q)$, and is the basis for all of the conditional probability results of Sec. 1.8.2.

Theorem 1.17. *If Q is a random proposition, then*

$$\mathrm{E}(X\mathbb{1}_Q) = \mathrm{E}(X \mid Q) \Pr(Q) \quad \textit{where } \phi \in \mathcal{E}(X \mid \mathbb{1}_Q).$$

Proof. Follows from Thm 1.16 after setting $Y \leftarrow \mathbb{1}_Q$, taking $y \leftarrow 1$, and using Def. 1.10. □

Hence if $\Pr(Q) > 0$ then $\mathrm{E}(X \mid Q) = \mathrm{E}(X\mathbb{1}_Q) \big/ \Pr(Q)$.

The next result closes the gap between $\psi \in \mathcal{E}(X \mid \boldsymbol{Y})$ and $\mathrm{E}(X \mid \boldsymbol{Y} \doteq \boldsymbol{y})$: anything other than this result would be extremely alarming! It is used extensively in Sec. 5.2.

Theorem 1.18. *If $\psi \in \mathcal{E}(X \mid \boldsymbol{Y})$ and $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) > 0$ then*

$$\psi(\boldsymbol{y}) = \mathrm{E}(X \mid \boldsymbol{Y} \doteq \boldsymbol{y}).$$

Proof. We have both

$$\begin{aligned} \mathrm{E}(X\mathbb{1}_{\boldsymbol{Y}\doteq\boldsymbol{y}}) &= \psi(\boldsymbol{y})\ \Pr(\boldsymbol{Y}\doteq\boldsymbol{y}) \qquad & \psi &\in \mathcal{E}(X \mid \boldsymbol{Y}) \\ \text{and } \mathrm{E}(X\mathbb{1}_{\boldsymbol{Y}\doteq\boldsymbol{y}}) &= \phi(1)\ \Pr(\boldsymbol{Y}\doteq\boldsymbol{y}) \qquad & \phi &\in \mathcal{E}(X \mid \mathbb{1}_{\boldsymbol{Y}\doteq\boldsymbol{y}}) \end{aligned}$$

the first from Thm 1.16, and the second after setting $Q \leftarrow \boldsymbol{Y} \doteq \boldsymbol{y}$ in Thm 1.17. If $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) > 0$ then these two relations can be rearranged to show

$$\psi(\boldsymbol{y}) = \frac{\mathrm{E}(X\mathbb{1}_{\boldsymbol{Y}\doteq\boldsymbol{y}})}{\Pr(\boldsymbol{Y}\doteq\boldsymbol{y})} = \phi(1) = \mathrm{E}(X \mid \boldsymbol{Y}\doteq\boldsymbol{y})$$

as required. □

1.8 More on conditional expectation

The previous section explained the motivation for introducing conditional and hypothetical expectations: they provide us with a much richer palette with which to specify our beliefs and, one hopes, this results in less 'undecided' for expectations we do not feel we can specify directly. This section explores more properties of conditional expectation, and conditional probability as a special case. These properties are useful in exactly the same sense as given on p. 10, and they will be used extensively in the following chapters.

1.8.1 Some useful results

The following results follow directly from the definition of $\mathcal{E}$ in Def. 1.9.

Theorem 1.19.

1. *If a is a constant then $x \in \mathcal{E}(X \mid a)$ and $a \in \mathcal{E}(a \mid X)$.*
2. $x \in \mathcal{E}(X \mid X)$.
3. *If $\psi \in \mathcal{E}(X \mid \boldsymbol{Y})$ then $\psi(\boldsymbol{y})\, g(\boldsymbol{y}) \in \mathcal{E}\{X\, g(\boldsymbol{Y}) \mid \boldsymbol{Y}\}$.*
4. *If $\psi \in \mathcal{E}(X \mid \boldsymbol{Y}, \boldsymbol{Z})$ and $\psi(\boldsymbol{y}, \boldsymbol{z}) = \phi(\boldsymbol{y})$ then $\phi \in \mathcal{E}(X \mid \boldsymbol{Y})$.*

Proof. These can all be verified by substitution into (1.5). □

This next result is very powerful, because it extends all of the results about expectations to hypothetical expectations.

Theorem 1.20. *If Q is a random proposition and $\Pr(Q) > 0$ then $\mathrm{E}(\cdot \mid Q)$ is an expectation.*

Proof. This is just a matter of checking the three properties given in Def. 1.4.

0. (Existence) For any X, if $\Pr(Q) > 0$ then $\mathrm{E}(X \mid Q)$ exists and is unique, according to Thm 1.17.

1. (Lower boundedness) First, note that if $\Pr(Q) > 0$ then we have the normalisation property $\mathrm{E}(a \mid Q) = a$, from Thm 1.15 and Thm 1.19.

   Now, we need to show that $\mathrm{E}(X \mid Q) \geq \min \mathcal{X}$, where $\mathcal{X}$ is the realm of X. Define $Y := X - \min \mathcal{X}$, so that Y is non-negative. Then

   $$\begin{aligned} \mathrm{E}(Y \mid Q) &= \mathrm{E}(Y \mathbb{1}_Q) / \Pr(Q) && \text{by Thm 1.17} \\ &\geq 0 && \text{by lower-boundedness,} \end{aligned}$$

   as $Y\mathbb{1}_Q$ is non-negative. Then

   $$\begin{aligned} \mathrm{E}(X \mid Q) &= \mathrm{E}(Y + \min \mathcal{X} \mid Q) \\ &= \mathrm{E}(Y \mid Q) + \min \mathcal{X} && \text{by additivity and normalisation} \\ &\geq \min \mathcal{X} \end{aligned}$$

   where additivity is proved immediately below.

2. (Additivity)

   $$\begin{aligned} \mathrm{E}(X + Y \mid Q) &= \frac{\mathrm{E}\{(X+Y)\mathbb{1}_Q\}}{\Pr(Q)} && \text{by Thm 1.17} \\ &= \frac{\mathrm{E}(X\mathbb{1}_Q + Y\mathbb{1}_Q)}{\Pr(Q)} \\ &= \frac{\mathrm{E}(X\mathbb{1}_Q)}{\Pr(Q)} + \frac{\mathrm{E}(Y\mathbb{1}_Q)}{\Pr(Q)} && \text{by additivity of } \mathrm{E}(\cdot) \\ &= \mathrm{E}(X \mid Q) + \mathrm{E}(Y \mid Q) && \text{Thm 1.17 again.} \end{aligned}$$

□

This result entitles me to insert a '$|Q$' into any expectation, or any result involving expectations, provided that I do not believe Q to be impossible. For example, it implies that there is a conditional FTP, with

$$\mathrm{E}\{g(\boldsymbol{X}) \mid Q\} = \textstyle\sum_i g(\boldsymbol{x}^{(i)}) \cdot q_i \tag{1.6}$$

in place of the unconditional statement in Thm 1.5, where $q_i = \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(i)} \mid Q)$. Likewise, there is a conditional marginalisation theorem,

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \Pr(\boldsymbol{X} \doteq \boldsymbol{x}, \boldsymbol{Y} \doteq \boldsymbol{y} \mid Q),$$

and so on. Both of these results involve hypothetical probabilities of the form $\Pr(P \mid Q)$ where both P and Q are random propositions; these will be defined in Sec. 1.8.2 (but there are no surprises).

Next we have a celebrated result, which is a cornerstone of the very elegant and powerful theory of martingales.

Theorem 1.21 (Law of the Iterated Expectation, LIE)**.**

$$\mathrm{E}\{\mathbb{E}(X \mid \boldsymbol{Y})\} = \mathrm{E}(X).$$

Proof.

$$\begin{aligned}
\mathrm{E}\{\mathbb{E}(X \mid \boldsymbol{Y})\} &= \mathrm{E}\{\psi(\boldsymbol{Y})\} && \text{where } \psi \in \mathcal{E}(X \mid \boldsymbol{Y}) \\
&= \textstyle\sum_{\boldsymbol{y}} \psi(\boldsymbol{y}) \cdot \Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) && \text{by the FTP, Thm 1.5} \\
&= \textstyle\sum_{\boldsymbol{y}} \mathrm{E}(X \mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}}) && \text{by Thm 1.16} \\
&= \mathrm{E}\left(X \textstyle\sum_{\boldsymbol{y}} \mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}}\right) && \text{by linearity} \\
&= \mathrm{E}(X)
\end{aligned}$$

because $\sum_{\boldsymbol{y}} \mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}} = 1$. □

Working backwards through this proof, and remembering that $\psi(\boldsymbol{y}) = \mathrm{E}(X \mid \boldsymbol{Y} \doteq \boldsymbol{y})$ when $\Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) > 0$ (Thm 1.18), the LIE can also be expressed as

$$\mathrm{E}(X) = \sum_{\boldsymbol{y}} \mathrm{E}(X \mid \boldsymbol{Y} \doteq \boldsymbol{y}) \cdot \Pr(\boldsymbol{Y} \doteq \boldsymbol{y})$$

which may be familiar. Sometimes $\mathrm{E}(X \mid \boldsymbol{Y} \doteq \boldsymbol{y})$ will be quite easy (or uncontroversial) to assess for each $\boldsymbol{y}$, but $\boldsymbol{Y}$ itself is a collection of random quantities about which I have limited beliefs. 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 set

$$\big\{\mathrm{E}(X \mid \boldsymbol{Y} \doteq \boldsymbol{y}) : \Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) > 0\big\}.$$

Finally, the following simple result can be useful.

Theorem 1.22. *Let* $\boldsymbol{X} := (\boldsymbol{Y}, \boldsymbol{Z})$ *and suppose the truth of* Q *implies that* $\boldsymbol{Y} = \boldsymbol{y}$*. If* $\Pr(Q) > 0$ *then*

$$\mathrm{E}\{h(\boldsymbol{Y}, \boldsymbol{Z}) \mid Q\} = \mathrm{E}\{h(\boldsymbol{y}, \boldsymbol{Z}) \mid Q\}.$$

Proof. Follows because $\boldsymbol{Y} \neq \boldsymbol{y}$ implies that $\mathbb{1}_Q = 0$, and hence

$$\begin{aligned}
\mathrm{E}\{h(\boldsymbol{X}) \mid Q\} &= \mathrm{E}\{h(\boldsymbol{Y}, \boldsymbol{Z}) \mid Q\} \\
&= \frac{\mathrm{E}\{h(\boldsymbol{Y}, \boldsymbol{Z}) \mathbb{1}_Q\}}{\Pr(Q)} && \text{by Thm 1.17} \\
&= \frac{\mathrm{E}\{h(\boldsymbol{y}, \boldsymbol{Z}) \mathbb{1}_Q\}}{\Pr(Q)} && \text{see above} \\
&= \mathrm{E}\{h(\boldsymbol{y}, \boldsymbol{Z}) \mid Q\} && \text{Thm 1.17 again.}
\end{aligned}$$

A similar argument was used in Thm 1.12. □

1.8.2 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 Thm 1.16 and the following definition.

Definition 1.11 (Conditional probability)**.** *Let* P *and* Q *be random propositions. Then*

$$\Pr(P \mid Q) := \mathrm{E}(\mathbb{1}_P \mid Q).$$

Then by Thm 1.16 we have (after the substitution $X \leftarrow \mathbb{1}_P$)

$$\Pr(P, Q) = \Pr(P \mid Q) \Pr(Q). \tag{1.7}$$

Eq. (1.7) is often rearranged to provide the 'definition' for conditional probability, which requires $\Pr(Q) > 0$. It is important to understand this rearrangement is *not* the definition of conditional probability—it is a result that arises from the definitions for $\mathbb{E}(X \mid \boldsymbol{Y})$ and for $\mathrm{E}(X \mid Q)$ in Sec. 1.7.2, plus Def. 1.11. What is distinctive about (1.7) is that it is always true. The case where $\Pr(Q) = 0$ causes no particular difficulties, except for the possibly uncomfortable implication that $\Pr(P \mid Q)$ is an arbitrary value in the interval $[0, 1]$.

There are lots of very useful results which follow directly from (1.7); in fact, they are all the same result, more-or-less. The following two may be generalised in the obvious way to any finite number of random propositions.

Theorem 1.23 (Factorisation theorem)**.**
Let P, Q, and R be random propositions. Then

$$\Pr(P, Q, R) = \Pr(P \mid Q, R) \Pr(Q \mid R) \Pr(R).$$

Proof. Follows immediately from two applications of (1.7):

$$\begin{aligned} \Pr(P, Q, R) &= \Pr(P \mid Q, R) \Pr(Q, R) \\ &= \Pr(P \mid Q, R) \Pr(Q \mid R) \Pr(R), \end{aligned} \tag{†}$$

because $\mathbb{1}_{P,Q,R} = \mathbb{1}_P \mathbb{1}_{Q,R}$ and $\mathbb{1}_{Q,R} = \mathbb{1}_Q \mathbb{1}_R$. □

This result leads immediately to the following.

Theorem 1.24 (Sequential conditioning)**.**
Let P, Q, and R be random propositions. If $\Pr(R) > 0$ then

$$\Pr(P, Q \mid R) = \Pr(P \mid Q, R) \Pr(Q \mid R).$$

Proof. Because $\mathbb{1}_{P,Q,R} = \mathbb{1}_{P,Q} \mathbb{1}_R$, (1.7) also implies that

$$\Pr(P, Q, R) = \Pr(P, Q \mid R) \Pr(R). \tag{‡}$$

Equating (‡) and (†) gives

$$\Pr(P, Q \mid R) \Pr(R) = \Pr(P \mid Q, R) \Pr(Q \mid R) \Pr(R),$$

and if $\Pr(R) > 0$ the final term can be cancelled from both sides to give the result. □

Then there is the very useful *Law of Total Probability (LTP)*, also known as the *Partition Theorem*. A partition is a collection of random propositions, exactly one of which must be true.

Theorem 1.25 (Law of Total Probability)**.**
Let P be a random proposition and $\mathcal{Q} := \{Q_1, \ldots, Q_k\}$ be any finite partition. Then

$$\Pr(P) = \sum_{i=1}^{k} \Pr(P \mid Q_i) \Pr(Q_i).$$

Proof. As $\sum_i \mathbb{1}_{Q_i} = 1$, we have

$$\mathbb{1}_P = \mathbb{1}_P\Big(\sum_{i=1}^{m} \mathbb{1}_{Q_i}\Big) = \sum_i \mathbb{1}_P \cdot \mathbb{1}_{Q_i} = \sum_i \mathbb{1}_{P,Q_i}.$$

The result follows from taking expectations of both sides and writing $\Pr(P, Q_i) = \Pr(P \mid Q_i)\Pr(Q_i)$ from (1.7). □

The LTP plays the same role as the LIE (Thm 1.21). In particular, in situations where it is hard to assess $\Pr(P)$ directly, it is possible to bound $\Pr(P)$ using the lower and upper bounds of the set

$$\big\{\Pr(P \mid Q_i) : \Pr(Q_i) > 0\big\}.$$

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

Theorem 1.26 (Bayes's theorem)**.** *If* $\Pr(Q) > 0$ *then*

$$\Pr(P \mid Q) = \frac{\Pr(Q \mid P)\Pr(P)}{\Pr(Q)}.$$

Proof. Follows immediately from (1.7),

$$\Pr(P, Q) = \Pr(P \mid Q)\Pr(Q) = \Pr(Q \mid P)\Pr(P),$$

and then rearranging the second equality. □

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

$$\Pr(P \mid Q_2, Q_1) = \frac{\Pr(Q_2 \mid P, Q_1)}{\Pr(Q_2 \mid Q_1)}\ \Pr(P \mid Q_1)$$

if $\Pr(Q_2, Q_1) > 0$. And there is Bayes's theorem for a finite partition, $\mathcal{P} := \{P_1, \ldots, P_m\}$:

$$\Pr(P_i \mid Q) = \frac{\Pr(Q \mid P_i)\Pr(P_i)}{\sum_j \Pr(Q \mid P_j)\Pr(P_j)} \quad i = 1, \ldots, m$$

if $\Pr(Q) > 0$, which uses the LTP in the denominator. And there is a Bayes's theorem in odds form,

$$\frac{\Pr(P_i \mid Q)}{\Pr(P_j \mid Q)} = \frac{\Pr(Q \mid P_i)}{\Pr(Q \mid P_j)}\,\frac{\Pr(P_i)}{\Pr(P_j)} \quad i, j = 1, \ldots, m$$

if $\Pr(P_j, Q) > 0$.

1.8.3 Probability Mass Functions

There is a very useful notation which allows us to compress certain expressions involving random propositions, and also to express sets of equalities concisely. For the time being we can think of it simply as a notation, but from Sec. 2.3 onward it becomes the primitive object of our belief specifications.

Definition 1.12 (Probability Mass Function, PMF)**.**
$f_{\boldsymbol{X}}$ *is a Probability Mass Function exactly when*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) := \Pr(\boldsymbol{X} \doteq \boldsymbol{x})$$

where $\boldsymbol{X} \doteq \boldsymbol{x}$ *denotes* $\bigwedge_i (X_i \doteq x_i)$.

According to the comma notation introduced in Sec. 1.4, we can also write more complicated PMFs, such as

$$f_{X,Y}(x,y) := \Pr(X \doteq x, Y \doteq y)$$

and so on, with the random propositions being taken in conjunction in the natural way. It is conventional to specify f_X for all real values of x, but set to zero if $x \notin \mathcal{X}$, but I will restrict the domain of the PMF to the product of the realms of its arguments. According to the FTP, to specify a PMF f_X is to specify the expectation of every possible function of X.

Conditional PMFs can be defined in exactly the same way, except with the *proviso* that $f_{X|Y}(\cdot \mid y)$ is undefined if $f_Y(y) = 0$. However, I will tend to ignore this when the ambiguity of $f_{X|Y}(\cdot \mid y)$ has no practical effect. Consider, for example, the restatement of (1.7) in terms of PMFs,

$$f_{X,Y}(x,y) = f_{X|Y}(x \mid y)\, f_Y(y). \tag{†}$$

The ambiguity of the first term on the righthand side is of no consequence if $f_Y(y) = 0$, because this implies that $f_{X,Y}(x,y) = 0$, and the equality holds for all (x,y).

Eq. (†) is an example of a *functional equality*. My convention is that this type of functional equality represents a set of equalities, one for every point in the product domain

$$(x,y) \in \mathcal{X} \times \mathcal{Y}.$$

However, the domain of some functional equalities need to be qualified, precisely because they cannot tolerate ambiguity in the value of $f_{X|Y}(\cdot \mid y)$. Bayes's theorem (Thm 1.26), for example, can be written as

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)\, f_X(x)}{f_Y(y)},$$

but this only holds for those y for which $f_Y(y) > 0$.

To clarify this constraint, we introduce the notion of the *support* of a random quantity or a collection of random quantities,

$$\operatorname{supp}(X) := \left\{ x \in \mathcal{X} : f_X(x) > 0 \right\}.$$

In other words, those elements of the joint realm where the probability is positive. We must have

$$\operatorname{supp}(X) \subset \prod_i \mathcal{X}_i$$

because $f_{X_i}(x_i) = 0$ implies that $f_X(\cdots x_i \cdots) = 0$. Using this notation, the correct domain for Bayes's theorem is

$$(x,y) \in \mathcal{X} \times \operatorname{supp}(Y).$$

These issues become critical in Sec. 5.2.

1.A* *Concepts from first order logic*

Here is a fairly precise statement about commonly-used mathematical terms in first order logic; this account is a *précis* of several sources, including Keisler (2007, ch. 15). First order logic for real numbers is used to define a random proposition and a probability (Sec. 1.4), and an unfamiliar notation is used (e.g. '$\doteq$') to disambiguate a commonly-used notation in statistics.

The language of first order logic comprises functions and variables, predicates, connectives, quantifiers, and punctuation (parentheses and commas). Functions are n-ary, indicating that they take n arguments, where $n \geq 0$. Functions that are 0-ary are called *constants*. Variables range over the set of all constants. The meanings of functions (including constants) and predicates depends on the interpretation of the language, but variables, connectives, quantifiers and punctuation have a fixed (conventional) meanings. In these notes, functions, constants, and variables will be real-valued, and predicates will be binary relations.

A *term* is a finite sequence of symbols defined inductively according to:

1. Every constant and every variable is a term;
2. If $t_1, \ldots, t_n$ are terms and f is an n-ary function with $n \geq 1$, then $f(t_1, \ldots, t_n)$ is a term.

Binary relations have the form $s\, R\, t$, where s and t are terms. The binary relations comprise

$$\doteq,\ \dot{\neq},\ \dot{<},\ \dot{\leq},\ \dot{\geq},\ \text{and}\ \dot{>}.$$

The dot over each symbol indicates that these are predicates, and so mean something different from their usual 'undotted' usage. This is explained further after the description of a first order sentence on p. 36. *Connectives* comprise

$$\neg\ \text{(not)},\ \wedge\ \text{(and)},\ \vee\ \text{(or)},\ \implies\ \text{(implies)},\ \text{and}\ \iff\ \text{(if and only if)},$$

each of which is defined in terms of the usual truth tables. *Quantifiers* comprise

$$\forall\ \text{(for all)},\ \text{and}\ \exists\ \text{(there exists)}.$$

There is some redundancy here, since all of these connectives and quantifiers can be constructed from the smaller set $\{\neg, \vee, \exists\}$, but it is much clearer to keep them all.

A *formula* is a finite sequence of symbols defined inductively according to:

1. If R is a relation and s and t are terms then $s\, R\, t$ is a formula.
2. If ψ and ϕ are formulae, then

$$\neg\psi,\ \psi \wedge \phi,\ \psi \vee \phi,\ \psi \implies \phi,\ \text{and}\ \psi \iff \phi$$

are formulae.

3. If $\psi(v)$ is a formula and v is a variable, then

$$\forall v\psi(v) \quad \text{and} \quad \exists v\psi(v)$$

are formulae.

In a formula, a variable can be either a *free variable* or a *bound variable*. It is free if it is not quantified, otherwise it is bound. For example, in the formula $\forall v(v\, R\, w)$ the variable v is bound and the variable w is free. A formula with no free variables is a *first order sentence*: these are the formulae with well-defined truth values. Thus if a and b are constants then $a \mathrel{\dot{\leq}} b$ is a sentence. If f and g are 1-ary functions, then

$$\forall v(f(v) \doteq g(v))$$

is a sentence, which is true if f and g are the same function, and false if they are not. If $\psi(v)$ is a formula with a free variable v and c is a constant, then $\psi(c)$ is a sentence. For example, $(v \mathrel{\dot{\leq}} 3)$ is a formula with a free variable v, and $(2 \mathrel{\dot{\leq}} 3)$ is a sentence.

The truth of a sentence is defined inductively according to:

1. If R is a binary relation then the sentence $a\, R\, b$ is true exactly when the constants a and b are defined and $(a, b) \in R$.

2. If ψ and ϕ are sentences and C is a connective then the truth of $\psi\, C\, \phi$ is determined according to the usual truth tables.

3. The sentence $\forall v\psi(v)$ is true exactly when $\psi(c)$ is true for all constants c.

4. The sentence $\exists v\psi(v)$ is true exactly when $\psi(c)$ is true for some constant c.

It should be clear now why it is important to distinguish the predicate '$\doteq$' from the more usual '$=$'. The first-order sentence '$\psi \doteq \phi$' evaluates to false or true, depending on the values of ψ and ϕ, but the equation '$\psi = \phi$' is an assertion that the objects ψ and ϕ are equal to each other.[16]

[16] In first-order logic, predicates are written $P(x, y, z)$. But when the predicates are binary predicates it is much clearer to write $P(x, y)$ as $x\, P\, y$, known as 'infix' notation. Unfortunately for us, this clashes with the more usual uses of symbols such as '$=$' and '$\leq$', which is why the infix predicates are ornamented with dots.