Jonathan Rougier

Department of Mathematics
University of Bristol

Lecture notes on Statistical Inference

Our mission: To help people make better choices under uncertainty.

Version 3, compiled on December 9, 2014.

Contents

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.

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] See footnote 4 on p. 7.

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 $\boldsymbol{X} := (X_1, \dots, X_m)$, and their joint realm as $\boldsymbol{\mathcal{X}}$, where $\boldsymbol{x} := (x_1, \dots, x_m)$ is an element of $\boldsymbol{\mathcal{X}}$, and

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

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,

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

> [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.

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.

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.

[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.

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

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.

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

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

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 g, to represent the physical impossibility of an orange with negative weight. I might believe that the weight cannot be less than 50 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.

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 $0 = X + (-X)$.

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

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]

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.

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

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. □

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,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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:

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

The indicator function is used to define a probability.

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

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

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

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 $\{A_1, \dots, A_k\}$ is a collection of random propositions, then define

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

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. □

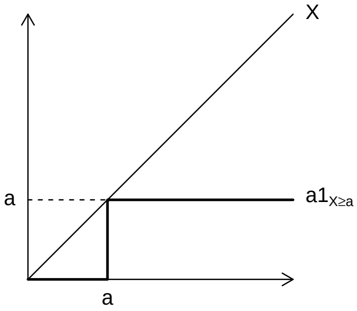

Figure 1.1: Markov's inequality.

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 $X \mathrel{\dot{\geq}} a \iff g(X) \mathrel{\dot{\geq}} g(a)$. A useful application of this generalisation is

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

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

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

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.

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 $\boldsymbol{X} := (X_1, \dots, X_m)$ be any finite collection of random quantities (with finite realms) and let

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

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

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

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

as required.

($\Rightarrow$). Note that

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

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*

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

where $\boldsymbol{\mathcal{Y}}$ 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

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

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

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

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

Theorem 1.7 (Jensen's inequality)**.**
Let $\boldsymbol{X} := (X_1, \dots, X_m)$. If g is a convex function of $\boldsymbol{x}$, then $\mathrm{E}\{g(\boldsymbol{X})\} \ge g(\mathrm{E}\{\boldsymbol{X}\})$.

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 $\mathcal{X} := \big\{\boldsymbol{x}^{(1)}, \dots \boldsymbol{x}^{(s)}\big\}$. 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 $\boldsymbol{p}^{(s)} := (p_1^{(s)}, \dots, p_s^{(s)}) \in \mathbb{S}^{s-1}$. I'll drop the superscript on $\boldsymbol{p}$ to avoid clutter.

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

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

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 $(p_1, \dots, p_{s+1})$ 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

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

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

Jensen's inequality is the basis for the very powerful *Gibbs's inequality*. This will appear in Sec. 4.6.

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

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

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

Proof.

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

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, $\boldsymbol{X} := (X_1, \dots, X_m)$. 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 $\boldsymbol{X} := (X_1, \dots, X_m)$ and let

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

be their joint realm. Let $\{g_1, \dots, g_k\}$ 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

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

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 $\boldsymbol{v} := (v_1, \dots, v_k)$.

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 $\{h_1, \dots, h_n\}$ 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 $h_1(\boldsymbol{X}), \dots, h_n(\boldsymbol{X})$ is

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

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

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

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 $g_1(\boldsymbol{X}), \dots, g_k(\boldsymbol{X})$. 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, $Q_1, Q_2, \dots$ might be different boundary conditions, representing different scenarios for future greenhouse gas emissions.

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

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

Together, my w and this formula provide a new row for my $[G, \boldsymbol{v}]$ 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 $[G, \boldsymbol{v}]$ 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.

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

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 $\boldsymbol{Y} := (Y_1, \dots, Y_n)$?" 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 $g : \boldsymbol{\mathcal{Y}} \to \mathbb{R}$, 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 $\psi : \boldsymbol{\mathcal{Y}} \to \mathbb{R}$ 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

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

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 $\boldsymbol{\mathcal{Y}}$ 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

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

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 $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ 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:

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

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

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 $\mathcal{Y} := \{y^{(1)}, \dots, y^{(r)}\}$. Any function of y can be written as

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

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

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

Substituting for ψ and then multiplying out gives

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

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

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

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 $i = 1, \dots, r$, 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 $\mathrm{E}(X \mid Q) = \mathrm{E}(X\mathbb{1}_Q) / \Pr(Q)$.

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

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

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)

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

□

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

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

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,

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

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

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

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 $\mathcal{Q} := \{Q_1, \dots, Q_k\}$ 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, $\mathcal{P} := \{P_1, \dots, P_m\}$:

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

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

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

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_X is a Probability Mass Function exactly when

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

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

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

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

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

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 $a \dot{\leq} b$ 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, $(v \dot{\leq} 3)$ is a formula with a free variable v, and $(2 \dot{\leq} 3)$ 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.

2
Modern Statistical Practice

The central part of this chapter describes modern statistical practice as a sequence of developments. This is *not* a history of statistics. Rather, it is a 'model' of statistics, where I understand a 'model' to be *an artefact used to organise our knowledge and beliefs*. I stand by my definition of statistical inference in Chapter 1, and its recognition of our limitations when quantifying uncertainty. And yet we find very little trace of these limitations in modern statistical practice. Sec. 2.2 to Sec. 2.6 is my model of this anomaly; its sequential structure is an organisational device.

Before these middle sections, the next section does some preliminary spade-work, dispelling a naive interpretation of statistical practice ('learning') and replacing it with a naturalistic one, which I hope will be recognisable to any applied statistician, despite my need to describe it in rather abstract terms. And then after the middle sections there are two additional starred sections that cover two abiding issues in statistical practice. Sec. 2.7 considers when two statisticians will agree with each other in their inferences. And Sec. 2.8 considers how strongly an inference is based on the available dataset. Finally, an appendix (Sec. 2.A) provides a very brief outline of some of the notions of Frequentist statistics, including estimators and confidence intervals.

2.1 *Some preliminary spadework*

This is a brief discussion about what statistics is not, and what it is (as actually practiced, not as theorised about).

2.1.1 *Statistical inference is not 'learning'*

Recall my definition of 'model' at the start of this chapter. There is a model of idealised learning, which runs as follows. There is a collection of random quantities, say $\boldsymbol{X} := (X_1, \dots, X_m)$, about which an agent has beliefs. These beliefs are represented as a conjunction of first-order sentences about $\boldsymbol{X}$ which he believes to be true, denoted by the proposition Ψ. The agent's complete set of beliefs is written as Bel_Ψ, where $\text{Bel}_\Psi(P)$ is his strength of belief in the proposition P. 'Learning' consists of adding new sentences to Ψ, and is represented formally as the arrow in

$$\mathrm{Bel}_{\Psi}(P) \longrightarrow \mathrm{Bel}_{Q \wedge \Psi}(P),$$

where Q is a sentence now believed by the agent to be true, and for which $Q \wedge \Psi$ is not a contradiction. $\mathrm{Bel}_{Q \wedge \Psi}$ might be termed the agent's *updated* beliefs.

Now we could choose to represent the belief function $\mathrm{Bel}_{\Psi}(\cdot)$ by the conditional probability $\Pr(\cdot \mid \Psi)$, and hence we could represent learning Q as

$$\Pr(P \mid \Psi) \longrightarrow \Pr(P \mid Q, \Psi).$$

This is known as *Bayesian conditionalisation*. Paris (1994) provides a very clear description of this model for learning from the point of view of computer science, and it has been popular with physicists such as R.T. Cox and Edwin Jaynes (see, e.g., Jaynes, 2003), and also philosophers (see, e.g., Jeffrey, 2004; Howson and Urbach, 2006). Note that all parties see conditionalisation as a model for learning, and not a description for how we learn. As Jaynes expresses it, this model describes how we might program an agent such as a robot to operate on our behalf.

The learning rule in Bayesian conditionalisation has an interesting and attractive form, summarised in the following result.[1]

Theorem 2.1 (Muddy table theorem)**.** *Let* $\boldsymbol{X} := (X_1, \dots, X_m)$*,* Ψ *be a random proposition, and* $q(\boldsymbol{x})$ *be a first-order sentence with* $Q := q(\boldsymbol{X})$ *and* $\Pr(Q, \Psi) > 0$*. Then*

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q, \Psi) \propto \mathbb{1}_{q(\boldsymbol{x})} \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \Psi)$$

where the constant of proportionality is $\Pr(Q \mid \Psi)^{-1}$*.*

Proof. The result is clearly true if $\Pr(\boldsymbol{X} \doteq \boldsymbol{x}, \Psi) = 0$, and so let $\Pr(\boldsymbol{X} \doteq \boldsymbol{x}, \Psi) > 0$. From the sequential Bayes's theorem (after Thm 1.26),

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q, \Psi) = \frac{\Pr(Q \mid \boldsymbol{X} \doteq \boldsymbol{x}, \Psi) \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \Psi)}{\Pr(Q \mid \Psi)}.$$

Then applying the conditional FTP (eq. 1.6) to the first term gives

$$\begin{aligned} \Pr(Q \mid \boldsymbol{X} \doteq \boldsymbol{x}, \Psi) &= \textstyle\sum_j \mathbb{1}_{q(\boldsymbol{x}^{(j)})} \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)} \mid \boldsymbol{X} \doteq \boldsymbol{x}, \Psi) \\ &= \textstyle\sum_j \mathbb{1}_{q(\boldsymbol{x}^{(j)})} \mathbb{1}_{\boldsymbol{x}^{(j)} \doteq \boldsymbol{x}} \\ &= \mathbb{1}_{q(\boldsymbol{x})} \end{aligned}$$

completing the proof. □

Bas van Fraassen (1989, ch. 7) describes this learning rule in the following terms. We start with $\mathcal{X}$ represented as tiles on a tabletop (he suggested a Venn diagram). Each tile contains a heap of mud whose proportion in the total represents the agent's $\Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)} \mid \Psi)$, for tile j. When the agent learns that Q is true, he sweeps the mud off all the tiles for which $q(\boldsymbol{x}^{(j)})$ is false; i.e. all the tiles that are ruled out by the truth of Q. So Thm 2.1 is the *Muddy table theorem*.

[1] Recall the comment on functional equalities in Sec. 1.8.3. The equality in Thm 2.1 holds for all $\boldsymbol{x}$ in the realm of $\boldsymbol{X}$.

Bayesian conditionalisation has three very attractive properties. First, it is consistent, so that if $\Psi \implies P$ then $\text{Bel}_\Psi(P) = 1$. Second, it is property-preserving. If Bel_Ψ satisfies the three axioms of probability, then so will $\text{Bel}_{Q \wedge \Psi}$; this follows from Thm 1.20. Third, it is order-invariant. If $q(x) = q_1(x) \wedge q_2(x)$, then $\text{Bel}_{Q \wedge \Psi}$ will be the same whether the update is Q_1 then Q_2, or Q_2 then Q_1, or both together. These properties are so attractive that we should not be surprised to find that conditionalisation is also the basis of statistical practice.

Attractive as the learning model is, inference as practised by statisticians is *not* learning as described here. As described in Sec. 2.1.2, statistical practice tends to involve working *backwards* from the dataset (represented as the truth of Q above): this reverse direction is the antithesis of learning. But anything else would be completely impractical, because there is no end to the list of relevant things that might be learnt between two time-points, even in a highly controlled experiment. To have to anticipate all of these would unworkable, and even accounting for just a fraction of them would require a really massive $\mathfrak{X}$: far larger than we could compute with, and most of which would then be ruled out by the dataset.

One might think that this was too obvious to mention, but for the fact that statistical inference is constantly being confused with learning, a confusion that is even enshrined in our vocabulary—'prior' and 'posterior' distributions, for example: see Sec. 2.6. This confusion is also represented in the inane advice 'not to use the dataset more than once'. And in the practice of setting the characteristics of an inference ahead of analysing the dataset (as in hypothesis tests with prescribed error rates, or fixed thresholds for significance levels). Interestingly, it has affected some of the deepest thinkers in our profession:

> We are sometimes asked "If all rests ultimately on personal opinion, why should a person confronted with data not simply decide intuitively what opinions he finds himself with, having seen the data, rather than trouble to calculate these opinions from initial opinions and opinions about the mechanism of the experiment by means of Bayes' theorem?" The question has the merit of pointing out a false asymmetry between initial and final opinions. For it is only chronologically, not logically, that one has priority over the other.[2] (from the English summary of de Finetti and Savage, 1962)

The use of 'initially', 'final', and 'chronologically' all point to a learning interpretation, although the continuation of the quote (in the footnote) suggests a more iterative process.

2.1.2 *A more naturalistic description*

What happens in practice? Typically the statistician consults with the client, and takes delivery of a dataset, some beliefs, and an objective. In the simplest possible case the dataset is just a collection of values. In the most common case it is a *flat database*, such as a spreadsheet, where each row is a case (e.g., a person, a donkey, a

[2] The quote continues: "The reason to make Bayesian and other probability calculations is to allow a person to confront his various opinions with one another to see if they are coherent. If they are not, he will generally be able so to modify his opinions as to be satisfied that he has improved them. Often, but not always, it is the conclusions intuitively suggested by the experiment that will be so modified."

country, an experimental plot), and each column is a variable. Every filled-in cell in this spreadsheet represents a measurement of the form $X_i \to x_i$, where the definition X_i is inferred from the row and the column, and x_i is the value.[3]

A brief but important aside on datasets: *do not assume that the dataset provided by the client is error free*. I have not yet taken delivery of an error-free dataset. Remember that statisticians are *data scientists*, but the client and her research team may not be. Statisticians have powerful tools to sanity-check a dataset, including the humble but effective *parallel coordinates plot*.[4] When you find an error, *do not* make a local modification to your download, or create a patch in your code. Instead, push it back to the client to correct in her source. Encourage her to put the source under version control, and ensure that the version information is always embedded in the download that you are given.[5]

The client's beliefs concern some wider collection of random quantities $\boldsymbol{X} := (X_1, \ldots, X_m)$, and we can assume that the dataset concerns a subset of these. Her beliefs can be represented as expectations of specified functions of $\boldsymbol{X}$, just like the g_i functions in Sec. 1.6.1, and, as explained in Chapter 1, these can include probabilities, hypothetical expectations and hypothetical probabilities. Her objective can be represented as an inference about $h(\boldsymbol{X})$, or a set of such functions. The statistician and the client work together to specify $\boldsymbol{X}$, the g's, and the h's.

This is a backward-and-forward process that cannot avoid involving the dataset as well. For a start, the dataset originated with the client, and she will undoubedly carry some of its values into her beliefs about $\boldsymbol{X}$.

Second, not all of the dataset will be 'modellable'. Some aspects will be hard to model, such as missing or unreliable values; see Gelman *et al.* (2014, chs 8, 18). It is often easier for the statistician to discard variables with lots of missing values, than to attempt to model the 'missingness'; ditto with reliability. Of course the motivation here is not to give the statistician an easy life. But if the resulting inference is to be accepted by the client's peers and her stakeholders, then sometimes it is better to skirt around contentious areas, than to confront them. Often 'unmodellability' will be discovered the hard way, as a failure in diagnostic assessment (to be discussed in Sec. 6.1).

To give one example, the under-recording of large volcanic eruptions in current databases increases as we go backwards in time, but at an unknown rate which depends on location and magnitude. Moreover, the reliability of the magnitude value of recorded eruptions decreases as we go back in time, but at an unknown rate that depends on location.[6] So the statistician and the client might decide, after some difficulties, to use only the observations from the recent past, in which there is effectively no under-reporting and the magnitude values are reliable, rather than to incorporate models for under-recording and unreliability for

[3] Actually, it is just as common for the dataset to be a collection of tables, and then the first task of the statistician is to join them into a single data object, remembering that some tables may be updates/patches on others.

[4] `parcoord` in the `MASS` package in `R`. Errors can also show up as highly influential observations: see Sec. 2.8.

[5] Also, do not trust a dataset that has been extracted by hand from a computer file. If necessary, take the file yourself, and write a script to do the processing. If you aspire to be a data scientist, then data processing is something you must learn to do well. Turning digital information into a usable dataset and more is known as 'data wrangling' or 'data munging'. This is rapidly becoming one of the high-value skills of the C21st. O'Neil and Schutt (2014) is a good place to start.

[6] In both of these cases location is a proxy for other variables such as population density, archiving, and closeness to trade routes.

which the client has only the vaguest beliefs.

So here are the steps in an actual statistical inference, from the point-of-view of the statistician:

1. Meet the client, take delivery of a dataset, a research objective, and some beliefs.

2. Sanity-check the dataset, and push errors back to the client. Keep doing this throughout the inference.

3. Meet the client again, refine the research objective and collect additional beliefs.

4. Specify a set of random quantities $\boldsymbol{X}$ which encompass some of the dataset, some of the client's beliefs, and her research objective, h.

5. Have a go at the inference about $h(\boldsymbol{X})$. Be prepared to return to any of the previous steps.

The objective at the end of this process is to report $\mathrm{E}^*\{h(\boldsymbol{X})\}$, where E^* is an expectation which reflects the client's beliefs, and also those elements of the dataset you (jointly) have decided to use. In general I represent the dataset as the truth of the random proposition $Q := q(\boldsymbol{X})$, where $q(\boldsymbol{x})$ is a first-order sentence. This is notationally efficient, and it also allows a high level of freedom about what constitutes 'data'.[7] The crucial feature of E^* is that $\Pr^*(Q) = 1$. This is what 'grounds' the inference $\mathrm{E}^*\{h(\boldsymbol{X})\}$. One strong reason to accept the value $\mathrm{E}^*\{h(\boldsymbol{X})\}$ as a valid inference about $h(\boldsymbol{X})$ is that this value is consistent with the dataset. Of course we do not simply hope that $\Pr^*(Q) = 1$, or accept $\Pr^*(Q) \approx 1$; instead, we build $\Pr^*(Q) = 1$ into our inference, so that it is automatically true. The next four sections (skipping starred section 2.5) all present ways of constructing an $\mathrm{E}^*(\cdot)$ with this property.

* * *

So, to wrap up this section, let us not confuse statistical inference with learning, but treat it as its own thing: a process in which we derive expectations for some specified random quantities $\boldsymbol{X}$, and for functions of $\boldsymbol{X}$, in which the random proposition Q, representing the dataset, is true.

The following sections outline different frameworks within which we practice statistical inference. I will conflate the roles of client and statistician, for simplicity, and focus on the latter. For convenience I will tend to refer to 'Bayesians' and 'Frequentists' as though they were two different tribes of statisticians, like the Houyhnhnms and Yahoos of *Gulliver's Travels*. But although many statisticians will self-identify as one or the other, generally a more pragmatic attitude prevails, and where I write, e.g., 'Bayesians' one ought to read this as 'statisticians operating in a Bayesian mode'.

[7] The more traditional data model is described in eq. (2.6).

2.2 Brief review of Chapter 1

The portrait of statistical inference given in Chapter 1 acknowledged from the outset our limitations as assessors of uncertainty. Hence the need for a calculus that imposes simple and intuitive rules. In the calculus of expectations, I specify my beliefs about $\boldsymbol{X}$ as expectations of a set of random quantities, including probabilities as a special case. Expectation is characterised, informally, in Def. 1.2, and defined by the properties given in Def. 1.4.

The approach of hypothetical expectations (and hypothetical probabilities) introduced in Sec. 1.7 provides a powerful approach to extending the set of expectations I can specify, for example by allowing me to reason causally within scenarios that I can construct, which may or may not happen.

Among my beliefs, represented as expectations, I will include the dataset, represented as the truth of the random proposition Q. Formally I specify the belief $\Pr(Q) \leftarrow 1$, but this is equivalent to thinning the joint realm $\mathcal{X}$, removing all of the $\boldsymbol{x}^{(j)}$ elements for which $q(\boldsymbol{x}^{(j)})$ is false. The mechanics of statistical inference were described in Sec. 1.6. My expectations can be checked for coherence, and they can be extended to expectations of those random quantities which are the objective of the inference. There is no need to distinguish between E and E^*.

Typically I will find that my expectations of many random quantities are not single values, but intervals. This 'undecided' aspect of my beliefs is a consequence of the limitations of my beliefs. The width of the interval can be reduced in a number of ways, all of which cost resources, but which might be justified by the need I have for tightly-constrained beliefs. For example, I might spend more time thinking about those beliefs which I have quantified. Or I might go out and extend my beliefs by polling experts. Or I might augment the dataset.

The computational tool is Linear Programming, see Sec. 1.6.3. But, as that section discussed, this tool does not scale well with large numbers of random quantities, because the size of the realm $\mathcal{X}$ is exponential in the number of random quantities: that is why this chapter does not stop right here. The calculation we ought to do is often not practical, and an approximation must be found. As will be seen in the next three sections (skipping starred section 2.5), the approximation is to adopt a level of personal omniscience which belies our limitations.

2.3 Bayesian statistical inference

Bayesians are bold. They laugh in the face of 'undecided', asserting that every expectation is specified. According to the FTP (Thm 1.5), this is equivalent to specifying a $\boldsymbol{p} \in \mathbb{S}^{s-1}$, where $p_j = \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)})$.[8] Were I being a Bayesian, I would have such a $\boldsymbol{p}$, and my expectation for any random quantity $h(\boldsymbol{X})$ would be

[8] As usual, $\mathbb{S}^{s-1}$ is the $(s-1)$-dimensional simplex, defined in (1.2), and s is the size of $\mathcal{X}$, the realm of $\boldsymbol{X}$.

computable using the FTP,

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

As explained in Chapter 1, this expression also covers probabilities, hypothetical expectations, and hypothetical probabilities.

Now $\mathcal{X}$ might be huge, comprising thousands if not millions of elements. Clearly our Bayesian statistician cannot think about each $\boldsymbol{x}^{(j)}$ and specify each p_j. Instead, he specifies a formula $f_{\boldsymbol{X}}$ for which

$$p_j \leftarrow f_{\boldsymbol{X}}(\boldsymbol{x}^{(j)}) \qquad j = 1, \dots, s.$$

Here $f_{\boldsymbol{X}}$ is termed the *probability mass function (PMF)* of $\boldsymbol{X}$. So were we to ask him what his probability was for the random proposition $\boldsymbol{X} \doteq \boldsymbol{x}$ he would say "Hang on while I plug $\boldsymbol{x}$ into my formula ... Ah ha! It's 0.002347843." Now is this really his probability? Well, it is now! This is the basic challenge of the Bayesian framework, to propose a defensible formula for specifying a probability for every element of $\mathcal{X}$. I come back to this in Sec. 2.6.

(This is the point in these notes alluded to in Sec. 1.8.3, at which the PMF stops being a convenient notation for expressing functional equalities, but becomes the essential object of an inference.)

Having surmounted this challenge, the Bayesian is then in good shape. The dataset, represented as the truth of the random proposition $Q := q(\boldsymbol{X})$, is incorporated into his beliefs by conditioning,

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) \propto \mathbb{1}_{q(\boldsymbol{x})} f_{\boldsymbol{X}}(\boldsymbol{x}) \tag{2.1a}$$

by the Muddy table theorem (Thm 2.1). Summing over the whole of $\mathcal{X}$ supplies the missing constant of proportionality, which is $\Pr(Q)^{-1}$, presuming that this is positive. Finally, expectations of interest are computed as

$$\begin{aligned} \mathrm{E}^*\{h(\boldsymbol{X})\} &:= \mathrm{E}\{h(\boldsymbol{X}) \mid Q\} \\ &= \textstyle\sum_j h(\boldsymbol{x}^{(j)}) \cdot \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)} \mid Q), \end{aligned} \tag{2.1b}$$

according to the conditional FTP (eq. 1.6).

Bayesians tend to take it for granted that the truth of Q is incorporated into beliefs by conditioning. The attractive features of this method were described in Sec. 2.1.1. But there is another justification from Decision Theory, which will be discussed in Sec. 3.7.

The last two decades have seen a statistical computing revolution in which this inferential calculation can be extended to joint realms which are non-finite and non-countable, using *Markov chain Monte Carlo (MCMC)* sampling techniques; see Besag (2004) for a summary, and Robert and Casella (2004) for details. With these techniques it is not necessary to enumerate $\mathcal{X}$, and so the size of $\mathcal{X}$ is not, in itself, an impediment to inference, as it would be for the Linear Programming calculations described in Sec. 1.6.3. Also it is not necessary to know $\Pr(Q)$. It is hard to overstate the way in which the simultaneous development of MCMC sampling algorithms and computing power has revolutionised Bayesian statistical inference. There are now tools in which one supplies an f_X and a q, and everything else is automated; see, e.g., Lunn *et al.* (2013) for a description of BUGS.[9]

2.4 *Frequentist statistical inference*

Frequentists are ostensibly more cautious than Bayesians. Unlike Bayesians, they are unwilling to commit up-front to a single $\boldsymbol{p}$, preserving some 'undecided' in all of their expectations. They do this by proposing a *statistical model*, which is a family of PMFs indexed by a *parameter* $\theta \in \Omega$, where Ω is the *parameter space*. For any particular choice of θ they have

$$p_j \leftarrow f_X(\boldsymbol{x}^{(j)};\theta) \quad j = 1,\dots,s.$$

With this statistical model, any expectation or probability is a function of θ:

$$\mathrm{E}\{h(\boldsymbol{X});\theta\} = \textstyle\sum_j h(\boldsymbol{x}^{(j)}) \cdot f_X(\boldsymbol{x}^{(j)};\theta),$$

by the FTP.[10] There is more discussion about families of distributions in Sec. 2.5, but the material in this section should be read first.

In fact, this notion of a statistical model is completely general. The vector $\boldsymbol{p}$ lives in $\mathbb{S}^{s-1}$, which has the cardinality of the continuum. Thus if Ω also has the cardinality of the continuum, e.g. the convex interval $[0,1]$, then we can arrange a bijective relationship between $\mathbb{S}^{s-1}$ and Ω, and every possible $\boldsymbol{p}$ can be represented by a $\theta \in \Omega$. But what actually happens, of course, is that the Frequentist chooses the statistical model and Ω to severely restrict the set of possible $\boldsymbol{p}$.

A simple definition of the effective dimension of the statistical model is the minimum number of expectations it takes to completely specify $\boldsymbol{p}$. In standard statistical models this equates to the dimension of Ω, which is usually treated as a product space (see Sec. 2.5). For example, $\theta = (\lambda)$ for a Poisson model for X, or $\theta = (\mu,\sigma^2)$ for a Normal model for X. In the Poisson model, $\mathrm{E}(X;\lambda) \leftarrow v$ is sufficient to specify (λ) and thus $\boldsymbol{p}$, while in the Normal model $\mathrm{E}(X;\mu,\sigma^2) \leftarrow v_1$ and $\mathrm{E}(X^2;\mu,\sigma^2) \leftarrow v_2$ are sufficient to specify (μ,σ^2) and thus $\boldsymbol{p}$. So we should not get carried away by the potential generality of the Frequentist approach: if the Bayesian approach has effective dimension zero ($\boldsymbol{p}$ specified directly), then typical Frequentist models might have effective dimension of 'few'. These are both a long way from the cardinality of the continuum!

Now to incorporate the belief that Q is true. The natural approach is for the Frequentist to consider each element of Ω. First,

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q;\theta) \propto \mathbb{1}_{q(\boldsymbol{x})}\, f(\boldsymbol{x};\theta) \tag{2.2a}$$

by the Muddy table theorem (Thm 2.1), where the missing constant of proportionality is $\Pr(Q;\theta)^{-1}$, presuming that this is positive.

[9] And I should mention JAGS, `http://mcmc-jags.sourceforge.net/` and the new kid on the block, STAN, `http://mc-stan.org/`.

[10] Another common notation is $\mathrm{E}_\theta\{h(\boldsymbol{X})\}$, which I use in Sec. 2.A.

Then

$$\begin{aligned}\mathrm{E}^{*}\{h(\boldsymbol{X});\theta\} &:= \mathrm{E}\{h(\boldsymbol{X}) \mid Q;\theta\} \\ &= \textstyle\sum_{j} h(\boldsymbol{x}^{(j)}) \cdot \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)} \mid Q;\theta) \qquad (2.2\mathrm{b})\end{aligned}$$

by the conditional FTP. So far, this is a direct analogue of the Bayesian approach, except with an added '$;\theta$'. What to do, though, about the 'width' of Ω, represented by the unspecified value θ? Our Frequentist could report the union of the set of values for $\mathrm{E}^{*}\{h(\boldsymbol{X});\theta\}$ that is generated over those $\theta \in \Omega$ for which $\Pr(Q;\theta) > 0$. Thus he remains 'undecided' because these expectations will not all coincide.[11]

Unfortunately, the resulting union is typically far too wide to be useful (or believable). The Frequentist solution is to constrain the domain of θ according to the truth of Q. By far the most popular approach is to use the *Maximum Likelihood (ML) estimate*

$$\hat{\theta} := \operatorname*{argmax}_{t \in \Omega} \Pr(Q;t).$$

If this estimate is plugged-in for θ, then we have the resulting inference

$$\hat{\mathrm{E}}^{*}\{h(\boldsymbol{X})\} := \mathrm{E}^{*}\{h(\boldsymbol{X});\hat{\theta}\}, \qquad (2.3)$$

which is a point value with no 'undecided'. Sec. 2.A has more details about the Frequentist approach, which involves several concepts not required in the Bayesian approach. The next section provides a justification for using the ML estimate as the 'plug-in' value for θ.

2.5 Models and parameters*

A family of models is simply a set $\mathcal{F}$, where each $f_{\boldsymbol{X}} \in \mathcal{F}$ is a PMF for $\boldsymbol{X}$. The Frequentist must specify $\mathcal{F}$ (likewise the Bayesian in Sec. 2.6). The crucial point, which can easily be obscured in textbook descriptions, is that the set $\mathcal{F}$ is the basis for any inference. In particular, inferences should be invariant to the way in which we label the elements of $\mathcal{F}$, because labels are ephemeral. So suppose we construct $\mathcal{F}$ using the model $f_{\boldsymbol{X}}^{\theta}$ for $\theta \in \Omega$, where I now show the parameterisation of $f_{\boldsymbol{X}}$ as a superscript (or a subscript, below). If

$$g : \theta \mapsto \phi$$

is a one-to-one[12] mapping between Ω and Φ, where $\Phi = g(\Omega)$, then the same $\mathcal{F}$ could be described using the model $f_{\boldsymbol{X}}^{\phi}$ for $\phi \in \Phi$ providing that

$$f_{\boldsymbol{X}}^{\theta}(\boldsymbol{x};t) = f_{\boldsymbol{X}}^{\phi}(\boldsymbol{x};g(t)). \qquad (\dagger)$$

So we must check that all inferences about $\boldsymbol{X}$ are invariant to the choice of the parameterisation θ or ϕ, in the case where (†) holds.[13]

First, we can confirm that if (†) holds and $\Pr(Q;t) > 0$, then

$$\Pr_{\theta}\{\boldsymbol{X} \doteq \boldsymbol{x} \mid Q;t\} = \Pr_{\phi}\{\boldsymbol{X} \doteq \boldsymbol{x} \mid Q;g(t)\},$$

[11] There is no guarantee that the set of expectations will be convex, though, so it would not be correct to call it an interval.

[12] Or 'injective'.

[13] Because g is one-to-one, this condition is equivalent to $f_{\boldsymbol{X}}^{\phi}(\boldsymbol{x};s) = f_{\boldsymbol{X}}^{\theta}(\boldsymbol{x};g^{-1}(s))$. The same equivalence holds for the similar results below.

directly from (2.2a). Then it follows from (2.2b) that

$$\mathrm{E}_{\theta}^{*}\{h(\boldsymbol{X}); t\} = \mathrm{E}_{\phi}^{*}\{h(\boldsymbol{X}); g(t)\}. \tag{‡}$$

So far so good. Finally, though, we want to deal with the width of the parameter space using a plug-in value for the parameter. We can make a strong argument in favour of the Maximum Likelihood Estimator (MLE) as the preferred plug-in, due to the following property (see Sec. 2.A for more details of the notation for the MLE).

Theorem 2.2. *Let θ and ϕ be two different parameterisations of the family of distributions $\mathcal{F}$, for which $g : \theta \mapsto \phi$ is one-to-one. If $\hat{\theta}$ and $\hat{\phi}$ are the MLEs of θ and ϕ respectively, then $\hat{\phi}(\boldsymbol{y}) = g(\hat{\theta}(\boldsymbol{y}))$.*

Proof. The MLE of θ satisfies

$$f_{\boldsymbol{Y}}^{\theta}(\boldsymbol{y}; \hat{\theta}(\boldsymbol{y})) \geq f_{\boldsymbol{Y}}^{\theta}(\boldsymbol{y}; t) \qquad \text{for all } t \in \Omega,$$

for any $\boldsymbol{y} \in \mathcal{Y}$. And then, substituting from (†),

$$f_{\boldsymbol{Y}}^{\phi}(\boldsymbol{y}; g(\hat{\theta}(\boldsymbol{y}))) \geq f_{\boldsymbol{Y}}^{\phi}(\boldsymbol{y}; g(t)) \qquad \text{for all } t \in \Omega.$$

But the two sets $\{g(t) : t \in \Omega\}$ and Φ are equivalent, and hence we have

$$f_{\boldsymbol{Y}}^{\phi}(\boldsymbol{y}; g(\hat{\theta}(\boldsymbol{y}))) \geq f_{\boldsymbol{Y}}^{\phi}(\boldsymbol{y}; s) \qquad \text{for all } s \in \Phi,$$

showing that $g(\hat{\theta}(\boldsymbol{y}))$ is the MLE of ϕ, as required. □

Now we can combine these results to prove the following.

Theorem 2.3. *The plug-in inference $\hat{\mathrm{E}}^{*}\{h(\boldsymbol{X})\}$ defined in* (2.3) *is invariant to the parameterisation of the family of models.*

Proof.

$$\begin{aligned}
\hat{\mathrm{E}}_{\theta}^{*}\{h(\boldsymbol{X})\} &= \mathrm{E}_{\theta}^{*}\{h(\boldsymbol{X}); \hat{\theta}(\boldsymbol{y})\} && \text{by definition, (2.3)} \\
&= \mathrm{E}_{\phi}^{*}\{h(\boldsymbol{X}); g(\hat{\theta}(\boldsymbol{y}))\} && \text{by (‡)} \\
&= \mathrm{E}_{\phi}^{*}\{h(\boldsymbol{X}); \hat{\phi}(\boldsymbol{y})\} && \text{by Thm 2.2} \\
&= \hat{\mathrm{E}}_{\phi}^{*}\{h(\boldsymbol{X})\} && \text{(2.3) again.}
\end{aligned}$$

□

It is important to appreciate that other estimators for the parameter may not have the invariance property (e.g. Method of Moments estimators), and thus using them as plug-ins makes the inference sensitive to the choice of parameterisation, which most people would regard as undesirable.

* * *

As long as his inferential method is parameterisation-invariant, the statistician can choose whichever parameterisation of $\mathcal{F}$ is most convenient for him. This is an important practical point, because some choices of parameterisation are far more convenient than others. For example, when maximising over the parameter space to find the value of the MLE, it is very convenient if the parameter space can be written in a product form, i.e. if $\theta = (\theta_1, \dots, \theta_p)$ then

$$\Omega = \Omega_1 \times \cdots \times \Omega_p,$$

because it is far easier to explore rectangular regions than non-rectangular ones. In this case the parameters are said to be *variation independent*. For example, in the Normal statistical model we have $\theta = (\mu, \sigma^2) \in \Omega = \mathbb{R} \times \mathbb{R}_{++}$ for a parameterisation in terms of the expectation and the variance. Almost all practical models have parameters that are variation independent, although there is no theoretical reason for this property to be favoured.

Also, the statistician can choose a parameterisation in which

$$h^*(\theta) := \mathrm{E}^*\{h(\boldsymbol{X}); \theta\}$$

has a simple form (if this does not conflict with the previous property of variation independence). Forms such as $h^*(\theta) = \theta_1$, or some linear combination of the elements of θ, are popular. So popular, in fact, that a label is available for those elements of θ which are not in $h^*(\theta)$: they are called *nuisance parameters*. 'Old fashioned' textbooks devote a lot of material to particular families and parameterisations in which the nuisance parameters can be circumvented.[14] This material has largely been superseded by empirical methods such as the *bootstrap* (see, e.g., Davison *et al.*, 2003), or by a Bayesian approach, as discussed in the next section.

2.6 Bayesian/Frequentist synthesis

In Sec. 2.3, where did the Bayesian's $f_{\boldsymbol{X}}$ come from? In the synthesis of the two approaches to statistical inference it comes from accepting the notion of a statistical model $f_{\boldsymbol{X}}$ and a parameter $\theta \in \Omega$, but choosing to treat θ itself as uncertain.[15] I hesitate to call θ a 'random quantity', because typically it does not have an operational definition. Instead I will call it a *random variable*. The Bayesian treats the statistical model as a conditional distribution, which is formally valid because $f_{\boldsymbol{X}}(\boldsymbol{x}; t)$ behaves exactly like the hypothetical probability $\Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \theta \doteq t)$.[16] And then he specifies a PMF for θ, denoted π_θ, from which he constructs the joint PMF

$$f_{\boldsymbol{X},\theta}(\boldsymbol{x}, t) \leftarrow f_{\boldsymbol{X}}(\boldsymbol{x}; t)\, \pi_\theta(t) \tag{†}$$

following the template in (1.7). The PMF π_θ is termed the *prior distribution*—a label I do not like for reasons given in Sec. 2.1.1, but we are stuck with it.[17]

The choice of statistical model and prior distribution induces a PMF for $\boldsymbol{X}$—found by marginalising out θ from the joint PMF for $(\boldsymbol{X}, \theta)$. But, as shown immediately below, it is more convenient to marginalise out θ *after* conditioning on dataset. Sec. 2.1.2 discussed

[14] See, e.g., Cox and Hinkley (1974). I say 'old fashioned', but this remains one of my favourite textbooks.

[15] This synthesis position is somewhat conciliatory. A more general treatment of statistical modelling is given in Chapter 5.

[16] Technically, the assertion is that $f_{\boldsymbol{X}}(\boldsymbol{x}; \cdot) \in \mathcal{E}(\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}} \mid \theta)$ for each $\boldsymbol{x} \in \mathcal{X}$; see Sec. 1.7.2.

[17] I am using π_θ in preference to f_θ to draw a distinction between random quantities and random variables, but this is somewhat pedantic. In Chapter 5 I will use f_θ, for reasons explained in footnote 4 on p. 100.

how it is effectively impossible in statistical practice to separate the choice of PMF for $\boldsymbol{X}$ from the dataset Q. But at least the Bayesian need not concern himself explicitly with Q when he is choosing his model and his prior distribution, because these data will be incorporated by conditioning. In published papers the choice of π_θ is often ostentatiously *not* influenced by the dataset, but experienced statisticians will take this with a pinch of salt, being aware that the published model, the dataset, and the prior distribution are likely to have been arrived at iteratively. Sec. 2.7 explains why it is sometimes acceptable to make ostentatious choices for π_θ.

With π_θ specified, we are back on-track for the Bayesian calculation, except now the uncertain quantities are $(\boldsymbol{X}, \theta)$ rather than just $\boldsymbol{X}$. Hence, provided that $\Pr(Q) > 0$,

$$\begin{aligned}
&\Pr(\boldsymbol{X} \doteq \boldsymbol{x}, \theta \doteq t \mid Q) \\
&\quad = \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid \theta \doteq t, Q) \Pr(\theta \doteq t \mid Q) \quad \text{by seq. cond. (Thm 1.24)} \\
&\quad = \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q; t)\, \pi^*_\theta(t) \qquad (2.4a)
\end{aligned}$$

where the first term was defined in (2.2a), and

$$\pi^*_\theta(t) := \Pr(\theta \doteq t \mid Q) = \frac{\Pr(Q; t)\, \pi_\theta(t)}{\Pr(Q)} \qquad (2.4b)$$

by Bayes's theorem; termed the *posterior distribution*. Treated as a function of $t \in \Omega$, $\Pr(Q; t)$ is termed the *likelihood function*. Hence the Bayesian mantra:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Finally,

$$\begin{aligned}
\mathrm{E}^*\{h(\boldsymbol{X})\} &:= \mathrm{E}\{h(\boldsymbol{X}) \mid Q\} \\
&= \textstyle\sum_t \sum_j h(\boldsymbol{x}^{(j)}) \cdot \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}, \theta \doteq t \mid Q) \quad \text{by the CFTP (eq. 1.6)} \\
&= \textstyle\sum_t \sum_j h(\boldsymbol{x}^{(j)}) \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)} \mid Q; t)\, \pi^*_\theta(t) \quad \text{from (2.4a)} \\
&= \textstyle\sum_t \mathrm{E}^*\left\{h(\boldsymbol{X}); t\right\} \pi^*_\theta(t), \qquad (2.4c)
\end{aligned}$$

where the first term in the summation previously occurred in (2.2b). In other words, the Bayesian knows precisely how to handle the 'width' of Ω: he averages $\mathrm{E}^*\{h(\boldsymbol{X}); \theta\}$ over the posterior distribution π^*_θ.

* * *

This, then, is the key difference between the Frequentist approach and the Bayesian approach. Let us suppose that all agree on the statistical model $f_{\boldsymbol{X}}(\cdot; \theta)$ with its parameter $\theta \in \Omega$ (but see the discussion at the end of Sec. 2.7). The Frequentist approach eschews the specification of a prior distribution π_θ but must then adduce an 'extra-probabilistic' principle for handling the width of Ω in the inference.[18] The difficulty for the Frequentist is that no compelling principle has been found. Although maximum likelihood is the most popular, collapsing Ω to a single point $\hat{\theta}$ is a drastic step, and

[18] 'Extra' as in 'outside' or 'beyond'.

seems hardly defensible if h is non-constant around $\hat{\theta}$, even less so if it is non-linear.

In contrast, the Bayesian approach specifies a prior distribution π_θ and is then able to handle the width of Ω within the standard rules of probability. The difficulty for the Bayesian approach is that θ is not operationally-defined, and so π_θ is not a very natural PMF to specify. And so, in this case also, 'extra-probabilistic' principles are often adduced to handle the choice of π_θ; see Kass and Wasserman (1996) or Robert (2007) for reviews. The next section tackles the tricky question of when it is relatively harmless for a Bayesian to replace a carefully considered prior distribution with a rule-based one.

2.7 *Stable estimation*

As already discussed, the course of an inference involves model development, in which the statistician and the client iterate through a sequence of models, possibly varying the subset of the dataset which is modelled directly. Each model might have a different parameter space, which is bad news for the Bayesian, who has to specify a prior distribution for the parameters of each model. Early on, he may well prefer to use a rather simple rule-based prior distribution, and focus his efforts, as the Frequentist would, on the development of the model. But he may find, as his choice for the model settles down, that the effect on his inference of changing the prior distribution is rather small. This insentitivity to the choice of prior distribution can be formally analysed, and the result is a set of qualitative guidelines under which the Bayesian can replace a carefully considered prior distribution with a rule-based one, at no serious detriment to his inference.

The analysis was presented in a classic paper, Edwards *et al.* (1963), but was almost certainly the work of the third author, L.J. Savage.[19] I will adapt Savage's analysis to my own notation. Let $\Omega := \{t^{(1)}, \dots, t^{(k)}\}$ be the parameter space, which I take to be finite but otherwise unstructured. Let $\boldsymbol{u} := (u_1, \dots, u_k)$ be proportional to the prior probabilities, and $\boldsymbol{v} := (v_1, \dots, v_k)$ be proportional to the likelihoods, i.e.

$$u_i \propto \pi_\theta(t^{(i)}) \quad \text{and} \quad v_i \propto \Pr(Q; t^{(i)}) \qquad i = 1, \dots, k.$$

The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are the primitive quantities; everything below is derived from these. Define

$$p_i := \frac{x_i}{\sum_j x_j} \quad \text{and} \quad q_i := \pi^*_\theta(t^{(i)}) = \frac{x_i y_i}{\sum_j x_j y_j} \qquad i = 1, \dots, k.$$

Thus $\boldsymbol{q} := (q_1, \dots, q_k)$ are the posterior probabilities, and $\boldsymbol{p} := (p_1, \dots, p_k)$ are the normalised likelihoods. In these terms our interest is in when we can approximate an inference based on $\boldsymbol{q}$ with one based on $\boldsymbol{p}$, for which we do not have to specify a prior distribution.

[19] This conclusion is based on my familiarity with Savage's style, but it was shared by Dennis Lindley (see Lindley, 1980), who was an authority on Savage's work.

Savage provided three *stability conditions* that are sufficient to provide a small upper bound on the error of such an approximation, based around a subset $B \subset \Omega$. In each case I give the formal condition and its interpretation.

1. There is an $\alpha \ll 1$ for which
$$\sum_{i \notin B} p_i \leq \alpha \sum_{i \in B} p_i.$$
▶ The subset B contains almost all of the relative likelihood.

2. There is a $\beta \ll 1$ for which
$$\psi \leq u_i \leq \psi(1+\beta) \quad \text{for all } i \in B$$
(the value of ψ is unimportant).

   ▶ The prior probabilities change very little in B.

3. There is some θ for which
$$u_i \leq \theta\psi \quad \text{for all } i,$$
where $\gamma := \alpha\theta \ll 1$.

   ▶ The prior probabilities are nowhere very large compared to their (nearly constant) values in B.

Savage proved the following result (Implication 5, p. 203).[20]

Theorem 2.4 (Stable estimation theorem)**.**
Let $g : \Omega \to \mathbb{R}$ have upper bound G. Then
$$\left| \textstyle\sum_t g(t^{(i)})\, p_i - \sum_t g(t^{(i)})\, q_i \right| \leq G \cdot \big(\alpha + \beta + \gamma + \max\{\alpha, \gamma\}\big).$$

As an immediate corollary, set $g(t) \leftarrow \mathbb{1}_{t \in C}$ for any $C \subset \Omega$, and then Thm 2.4 implies that the *total variation distance* between the normalised likelihood and the posterior distribution is bounded above by $\alpha + \beta + \gamma + \max\{\alpha, \gamma\}$.

Referring back to Sec. 2.6, we are interested in the inference
$$\mathrm{E}^*\{h(\boldsymbol{X})\} = \textstyle\sum_t \mathrm{E}\{h(\boldsymbol{X}) \mid Q; t\}\, \pi^*_\theta(t) = \sum_i g(t^{(i)})\, q_i$$
taking $g(t) \leftarrow \mathrm{E}\{h(\boldsymbol{X}) \mid Q; t\}$. We might consider instead the approximation
$$\tilde{\mathrm{E}}^*\{h(\boldsymbol{X})\} := \textstyle\sum_i g(t^{(i)})\, p_i,$$
replacing the posterior distribution with the normalised likelihood. Thm 2.4 asserts that the relative absolute error in replacing E^* with $\tilde{\mathrm{E}}^*$ is bounded above by $\alpha + \beta + \gamma + \max\{\alpha, \gamma\}$. If the three stability conditions hold, then this value is close to zero, and

(i) the normalised likelihood is close to the posterior distribution in total variation distance, and

[20] For those referring to the original paper, I have simplified the expression by approximating δ and ε in terms of α, β, and γ, using $\delta \approx \beta + \gamma$ and $\varepsilon \approx \alpha + \beta$. So technically the result as stated is not quite true.

(ii) the approximate inference $\tilde{\mathrm{E}}^*\{h(\boldsymbol{X})\}$ is close to the actual inference $\mathrm{E}^*\{h(\boldsymbol{X})\}$ in relative absolute error.

In principle the crucial set B can be any subset of Ω. Usually, however, Ω has a topology. In this case there is a compelling reason to restrict B to a contiguous subset of Ω, because smoothness in the prior distribution will then imply a smaller β in condition 2 than would otherwise be the case. Suppose that $\theta := (\theta_1, \dots, \theta_p)$ with $\Omega \subset \mathbb{R}^p$. In this case a simple and effective strategy for identifying a B which respects the topology of Ω is to define it as a level set for the likelihood, i.e.

$$B := \{i : v_i \geq c\},$$

and adjust c from the maximum likelihood value downwards until a sufficiently small α is reached. This tends to generate connected B's because small perturbations in the parameter value tend to cause only small perturbations in the likelihood. Helpfully, the level sets of the likelihood are transformation-invariant, so it would not matter whether the parameter was θ or some one-to-one transformation (see Sec. 2.5).

Once a B with a small α has been found, the statistician must decide whether β and γ are sufficiently small, to satisfy stability conditions 2 and 3. Of course he could do this explicitly if he has specified a prior distribution. But the attraction of Savage's stability conditions is that he may be able to do this qualitatively, without specifying a prior distribution. Edwards *et al.* (1963) provide a detailed illustration. This is not a trivial exercise, even in the simple models that were ubiquitous thirty years ago. Now, however, we can compute with a very diverse set of models, and there has also been a widening of the kinds of applications that statisticians consider. So it is really quite hard to know whether, for one particular model, stability conditions 2 and 3 apply.[21]

There are two guidelines that can help. First, to construct models in which the parameters are meaningful. For example, to map individual parameters to distinct beliefs about $\boldsymbol{X}$. It is tempting to treat the parameters in a purist manner, devoid of meaning except as an index into a family of distributions (Sec. 2.5). But in this case it is impossible to have well-formed beliefs about the prior probabilities, and in particular beliefs about whether these prior probabilities might be much larger in the complement of B than in B, violating stability condition 3. Hierarchical models (Sec. 5.4) are a powerful framework for constructing models with meaningful parameters.

The second guideline is to favour models with fewer parameters. This limits the opportunity for one parameter to offset another in the likelihood, and should result in a more compact B with a smaller β in stability condition 2. Also, of course, the more parameters there are, the harder it is to ascribe a distinct meaning to each parameter.

Both of these guidelines are basic tenets of statistical modelling,

[21] Lindley (1980) alluded to this issue in his review of Savage's work.

followed by experienced statisticians of all tribes; see, e.g., Lehmann (1990) and Cox (1990), in the same volume of *Statistical Science* (vol. 5, num. 2). For the Bayesian, though, they have the beneficial side-effect of enabling an assessment of whether the stability conditions hold. If so, the Bayesian can simplify his inference by replacing a carefully-specified prior distribution with a simple and tractable one, confident in the knowledge that the relative absolute error in his inference will be small.

When we inspect a published Bayesian inference, we often find such simple and tractable prior distributions, such as $\pi_{\theta_1}(t_1) \propto 1$ or $\pi_{\theta_2}(t_2) \propto 1/t_2$. It is vey important to appreciate that in this case the $\boldsymbol{X}$-margin of the joint distribution $(\boldsymbol{X}, \theta)$ is *not* a representation of the Bayesian's beliefs about $\boldsymbol{X}$. Instead, the Bayesian has replaced his actual joint distribution for $(\boldsymbol{X}, \theta)$ with another distribution which does not have the right $\boldsymbol{X}$-margin, but which still gives approximately the right posterior distribution and inference about $h(\boldsymbol{X})$ after conditioning on the dataset. This issue will resurface in Chapter 6.

* * *

I deliberately refrained from framing the previous material in terms of the question "When might two statisticians agree in their inference?" This has been a traditional concern, and is a major preoccupation of 'learning' theories. The challenge for such theories is to explain why, if beliefs are subjective, we often come to hold many beliefs in common. A superficially attractive answer is to point to the Stable estimation theorem, or something like it, which indicates that agreement on the model and a sufficiently large dataset might do the trick, for Frequentists and Bayesians alike.

However, this answer fails, from a statistical point of view, because there is no reason for two statisticians to have the same model, or to make the same choice about which portion of the dataset to condition on. Hence the likelihood is just as subjective as the prior distribution.[22] This is not to say that the choices are not similar enough to lead to the same inferences. But we cannot prove similarity of beliefs in a formal sense unless there are strict conditions on the model.

If you find yourself working in an area where everyone has roughly the same likelihood function then you are experiencing a convergence of epistemic norms: everyone makes the same subjective choice. This does not make it an objective choice, and nor does it make it the right choice, except according to the epistemic standards that currently prevail. You should remember the phrase 'pessimistic meta-induction' (Ladyman, 2002, ch. 8). This is a philosphical position which argues that our belief in the 'rightness' of scientific models ('scientific realism') has been repeatedly overturned, and therefore we have no defensible reason for thinking that our current model is immune from this pattern.

[22] In fact I would go further, based on my own experience, which I doubt is unusual. I would be unlikely to choose the same model and dataset from one year to the next. New modelling frameworks are appearing all the time; often these were previously ruled out for computational reasons that are being eroded by faster CPUs, more memory, and parallel computation. Also, I am getting more experienced.

2.8* Strength of belief

The section introduces several interrelated concepts that are often muddled together. I think the perspective provided by Chapter 1 helps to disentangle them. 'Strength of belief' is a nebulous concept, but it seems reasonable to quantify it as the width of 'undecided' in my inference about $h(\boldsymbol{X})$. So one of the things that we lose in the Bayesian approach and the plug-in Frequentist approach is any assessment of the strength of my beliefs, because we have eliminated the 'undecided'. But that does not stop us from asking related questions that ought to be highly informative in the wider context in which an inference about $h(\boldsymbol{X})$ is required.

Two questions spring to mind:

1. *Tenability*: Accepting my statistical choices as appropriate, what would be the effect on my beliefs of using a slightly larger number of obervations?

2. *Robustness*: Keeping the dataset the same, what would be the effect on my beliefs of a small perturbation in my statistical choices?

Both of these questions can be addressed with and without 'undecided', and therefore they do not directly address the issue of how the elimination of 'undecided' has affected my beliefs. However, they address this issue indirectly because they show the degree to which the dataset dominates in my beliefs.

I think 'tenability' is a fascinating question, and a crucial one in evidence-based policy. In policy it is often the case that we could wait a bit longer, or spend a bit more money, and acquire a few more observations. The question of whether we should do this can be analysed in detail in Decision Theory (Sec. 3.5), and this would be the right thing to do if the choice of action is a critical one. But it would be useful to have a more lightweight assessment, which could be routinely performed alongside the inference. I will focus on 'tenability' in this section.

So consider the question of how much my beliefs about $h(\boldsymbol{X})$ might change if I acquired one more observation. In this section it is helpful to adopt the simple observation model given in (2.6), and, in addition, to require that the first n components of $\boldsymbol{X}$ (the observations) represent the same type of random quantity. This is the context in which it makes the most sense to consider acquiring more observations, but it is also the context in which the approximations below are most accurate.[23] In this special case when n large we can approximate the effect of acquiring an additional observation by examining the effect of deleting one of the current observations.

I will focus on a plug-in Frequentist inference, although the same approach may be used in a Bayesian inference.[24] Define $\boldsymbol{y}_{-j} := (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$, the dataset missing the jth obser-

[23] My use of the notation $(X_1, \dots, X_m)$ belies the generality of the random quantities in our inference: up until now there has been no need for X_i and X_j to be of the same type.

[24] Now would be a good time to look at Sec. 2.A.

vation, similarly for $\boldsymbol{Y}_{-j}$, and then define

$$\hat{\theta}_{-j}(\boldsymbol{y}) := \operatorname*{argmax}_{t \in \Omega} f_{\boldsymbol{Y}_{-j}}(\boldsymbol{y}_{-j}; t),$$

the MLE for θ based on all but the jth observation. Set $H := h(\boldsymbol{X})$, and now define

$$\begin{array}{ll} h^*(\theta) := \mathrm{E}\{H \mid \boldsymbol{Y} \doteq \boldsymbol{y}; \theta\} & \text{our objective} \\ \hat{h}^*(\boldsymbol{y}) := h^*(\hat{\theta}(\boldsymbol{y})) & \text{the MLE of } h^*(\theta)\text{, see Sec. 2.A.1} \\ h^*_{-j}(\theta) := \mathrm{E}\{H \mid \boldsymbol{Y}_{-j} \doteq \boldsymbol{y}_{-j}; \theta\} & h^*(\theta)\text{, but dropping the } j\text{th observation} \\ \hat{h}^*_{-j}(\boldsymbol{y}) := h^*_{-j}(\hat{\theta}_{-j}(\boldsymbol{y})), & \text{then plugging in } \hat{\theta}_{-j}(\boldsymbol{y}) \end{array}$$

so that $\hat{h}^*_{-j}$ is the MLE for $h^*(\theta)$ based on all but the jth observation. Now quantify the effect of one additional observation on my inference about H as the 'variance'

$$\sigma^2_{\text{LOO}}(h^*, \boldsymbol{y}) := \frac{1}{n} \sum_{j=1}^{n} \left\{ \hat{h}^*_{-j}(\boldsymbol{y}) - \hat{h}^*(\boldsymbol{y}) \right\}^2, \tag{†}$$

where 'LOO' denotes 'leave one out'. The square-root of this value, denoted $\sigma_{\text{LOO}}(h^*, \boldsymbol{y})$, is a simple inverse measure of the tenability of my beliefs about H, in the same units as H. A small value indicates that an additional observation like the n I already have will likely make only a small difference to my inference about H.

So although a point-value for my expectation of H conveys no 'undecided', we are able to assess the tenability of this value to convey how much we would expect it to change were we to acquire an additional observation. A useful side-effect of this calculation is that we can assess the *influence* of each observation on the result. Highly influential observations are always worth a second look. In my experience they will often be either mis-recorded or atypical. In the latter case, the statistician must choose whether to exclude them, to include them within the current statistical model, or to include them within an extended model. This issue of influential observations is one of the main drivers for the iterations in the modelling process mentioned in Sec. 2.1.2.

The only catch with this approach to computing the tenability of my inference about H is that it is expensive: n times as expensive as doing the inference itself. So now we turn to how the value of $\sigma_{\text{LOO}}(h^*, \boldsymbol{y})$ can be approximated from within the single inference based on all n observations. The crucial result is that when the asymptotic conditions described at the end of Sec. 2.A.1 hold,[25]

$$\widehat{\mathrm{SE}}(h^*, \boldsymbol{y})^2 \approx \frac{n-1}{n} \sum_{j=1}^{n} \left\{ \hat{h}^*_{-j}(\boldsymbol{y}) - \hat{h}^*(\boldsymbol{y}) \right\}^2; \tag{‡}$$

an heuristic argument is given at the end of this section. In this case,

$$\sigma_{\text{LOO}}(h^*, \boldsymbol{y}) \approx \frac{\widehat{\mathrm{SE}}(h^*, \boldsymbol{y})}{\sqrt{n-1}},$$

combining (†) and (‡).

[25] We also need $h^*(\theta)$ to be a smooth function of θ around $\hat{\theta}(\boldsymbol{y})$.

Often, uncertainties in Frequentist inference are given as 95% confidence intervals (see Sec. 2.A.2). Since a 95% confidence interval has a width of about 4 standard errors, we have the approximation

$$\sigma_{\text{LOO}}(h^*, \boldsymbol{y}) \approx \frac{u(\boldsymbol{y}) - \ell(\boldsymbol{y})}{4\sqrt{n-1}} \tag{2.5}$$

where $[\ell(\boldsymbol{y}), u(\boldsymbol{y})]$ is the 95% confidence interval for $h^*(\theta)$. This, it seems to me, is a simple way to interpret confidence intervals: in terms of the leave-one-out measure of tenability—see the discussion at the end of Sec. 2.A.2.

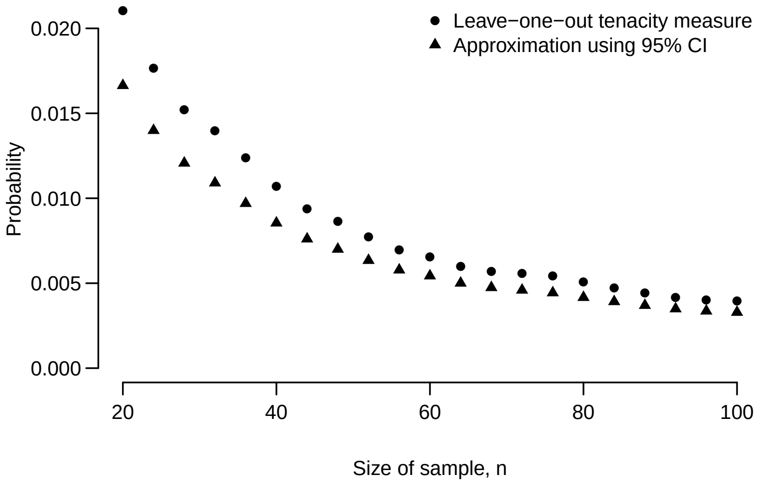

Figure 2.1: Leave-one-out tenability measure for the expected number of successes of an IID Bernoulli model, with an approximation based on a 95% confidence interval, given in (2.5).

Figure 2.1 shows an example of the convergence of the leave-one-out tenability measure with its approximation based on a 95% confidence interval, for the probability of success in an IID Bernoulli model. The observations were randomly generated with $\mathrm{E}(X_i) \leftarrow 0.2$. In this case the 95% confidence interval was computed as the 95% equi-tailed credible interval with a Jeffreys prior, as recommended by Brown *et al.* (2001).[26] At $n = 20$, one extra observation can affect my inference about the probability of success by about 2 percentage points, but by $n = 100$ this has fallen to 0.5 percentage points.

* * *

How can we understand the approximation in (‡)? The crucial result is stated at the end of Sec. 2.A.1, which is that, for sufficiently large n and under some conditions on the model, $\hat{\theta}(\boldsymbol{Y})$ behaves like the sample mean of an IID sample of size n. The same is true of $\hat{h}^*(\boldsymbol{Y}) := h^*(\hat{\theta}(\boldsymbol{Y}))$ if h^* is a smooth function of θ. So if we can establish that (‡) holds for the sample mean, then we can infer that it holds, approximately and under some conditions, for $\hat{h}^*$ as well.

So consider the squared estimated standard error of the sample mean, $\overline{\mu}(\boldsymbol{y}) := n^{-1} \sum_{i=1}^{n} y_i$,

$$\widehat{\mathrm{SE}}(\overline{\mu}, \boldsymbol{y})^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left[y_i - \overline{\mu}(\boldsymbol{y})\right]^2$$

[26] I.e., the confidence interval and credible interval were the same interval.

see (2.9). Now define the sample mean after leaving out one observation,

$$\overline{\mu}_{-j}(\boldsymbol{y}) := \frac{1}{n-1} \sum_{i \neq j}^{n} y_i = \frac{n\overline{\mu}(\boldsymbol{y}) - y_j}{n-1} \qquad j = 1, \dots, n.$$

A simple rearrangement shows that

$$(n-1)\{\overline{\mu}_{-j}(\boldsymbol{y}) - \overline{\mu}(\boldsymbol{y})\} = \overline{\mu}(\boldsymbol{y}) - y_j.$$

Hence we can rewrite the squared estimated standard error as

$$\begin{aligned} \widehat{\mathrm{SE}}(\overline{\mu}, \boldsymbol{y})^2 &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \big[(n-1)\{\overline{\mu}_{-i}(\boldsymbol{y}) - \overline{\mu}(\boldsymbol{y})\}\big]^2 \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \{\overline{\mu}_{-i}(\boldsymbol{y}) - \overline{\mu}(\boldsymbol{y})\}^2. \end{aligned}$$

The righthand side is precisely (‡), but with $\overline{\mu}$ in place of $\hat{h}^*$, as required.

The general theory for this result concerns the *jackknife* estimator of the variance, see Efron (1982) or Efron and Gong (1983) for more details. Efron and Stein (1981) show that the confidence interval approximation should be biased downwards, as shown in Figure 2.1.

2.A Core Frequentist concepts*

This Appendix summarises some of the concepts used in Frequentist inference. These tools are adapted to a more constrained dataset representation than the truth of the random proposition Q, which I will refer to as the *simple observation model*. In this model the dataset is assumed to correspond to the n components of $\boldsymbol{X}$, the first n for simplicity, for which

$$Q := q(\boldsymbol{X}) := \bigwedge_{i=1}^{n} (X_i \doteq y_i), \tag{2.6}$$

where y_i is the measured value of X_i. I write $\boldsymbol{Y}$ for $(X_1, \dots, X_n)$ to enforce the distinction between those random quantities which are measured, and those which are not; $\boldsymbol{Y}$ are the *observations*. Starting with the statistical model $f_{\boldsymbol{X}}(\cdot; \theta)$, the PMF of the observations is computed by marginalising out $(X_{n+1}, \dots, X_m)$:

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) &= \Pr(X_1 \doteq y_1, \dots, X_n \doteq y_n; \theta) \\ &= \sum_{x_{n+1}, \dots, x_m} f_{\boldsymbol{X}}(y_1, \dots, y_n, x_{n+1}, \dots, x_m; \theta) \end{aligned} \tag{2.7}$$

see Thm 1.6. In the case where the X's are IID this multiple sum has a very simple form (see Sec. 2.A.3).

In this section I will write $\mathrm{E}(\cdot; \theta)$ and similar operators as $\mathrm{E}_\theta(\cdot)$, to reduce clutter—this is a common notation.

2.A.1* Estimators

Any function of $\boldsymbol{y}$ is termed a *statistic*. Each statistic has a PMF induced by the choice of $f_{\boldsymbol{X}}$, which depends only on $f_{\boldsymbol{Y}}$. A statistic which is designed to be like the specified function $h(\theta)$ is termed an *estimator* of $h(\theta)$, which I will denote as $\widetilde{h(\theta)}$. But because such an estimator is usually constructed from estimator for θ, using

$$\widetilde{h(\theta)}(\boldsymbol{y}) \leftarrow h(\tilde{\theta}(\boldsymbol{y})),$$

I will focus on estimators for θ. I will take the range of $\tilde{\theta}$ to be a subset of the parameter space Ω. Although we should not presuppose that $\tilde{\theta}(\boldsymbol{y}) \in \Omega$ for all $\boldsymbol{y}$, in practical situations a failure of this condition would often have undesirable consequences.

For simplicity I now assume that θ is a scalar; the material in this section generalises if θ is a vector. The PMF of the random quantity $\tilde{\theta}(\boldsymbol{Y})$ is termed the *sampling distribution* of $\tilde{\theta}$. A good estimator has a sampling distribution which is concentrated around the value θ, no matter where in Ω that value happens to be. By the same argument as in Sec. 1.7.2, ideally $\tilde{\theta}(\boldsymbol{Y})$ would be effectively equivalent to θ, for all $\theta \in \Omega$. This suggests that the quality of an estimator be measured by its *mean squared error (MSE)*,

$$\mathrm{MSE}_\theta(\tilde{\theta}) := \mathrm{E}_\theta\left[\{\tilde{\theta}(\boldsymbol{Y}) - \theta\}^2\right],$$

which depends on θ, as shown. Inserting $-\mathrm{E}_\theta\{\tilde{\theta}(\boldsymbol{Y})\} + \mathrm{E}_\theta\{\tilde{\theta}(\boldsymbol{Y})\}$ and multiplying out shows that

$$\mathrm{MSE}_\theta(\tilde{\theta}) = \mathrm{Var}_\theta\left[\tilde{\theta}(\boldsymbol{Y})\right] + \left[\mathrm{E}_\theta\{\tilde{\theta}(\boldsymbol{Y})\} - \theta\right]^2$$

where the first term is the *variance* of the estimator, and the second term is the square of the *bias* of the estimator. Typically, choosing between estimators involves managing a trade-off between the variance and the squared bias.

The square root of the variance is termed the estimator's *standard error*, denoted $\mathrm{SE}_\theta(\tilde{\theta})$. This is a function of θ, but we can replace θ or functions of θ that occur in $\mathrm{SE}_\theta(\tilde{\theta})$ by estimators to give the *estimated standard error*; for example

$$\widehat{\mathrm{SE}}(\tilde{\theta}, \boldsymbol{y}) \leftarrow \mathrm{SE}_{\tilde{\theta}(\boldsymbol{y})}(\tilde{\theta})$$

(hang in there!). This is itself a statistic, and so has a sampling distribution, and so on.

An estimator for which the sampling distribution has the property

$$\mathrm{E}_\theta\{\tilde{\theta}(\boldsymbol{Y})\} = \theta \quad \text{for all } \theta \in \Omega$$

is said to be *unbiased*. There is no particular reason to favour unbiased estimators, given that often they do not exist, and where they do, often there are biased estimators with smaller MSEs.

* * *

Of the infinity of possible estimators for θ, the most popular estimator is undoubtedly the *maximum likelihood estimator (MLE)*,

$$\hat{\theta}(\boldsymbol{y}) := \operatorname*{argmax}_{t \in \Omega} f_{\boldsymbol{Y}}(\boldsymbol{y}; t).$$

There are several reasons for this. First, in simple models the MLE often has an intuitive form as a function of $\boldsymbol{y}$. Second, in more complicated models it is often fairly easy to compute the value of the MLE for a given $\boldsymbol{y}$ using numerical maximisation. It is sometimes possible to prove that $\log f_{\boldsymbol{Y}}(\boldsymbol{y}; t)$ is a concave function of t, so that if a numerical maximisation converges, then the result must be the global maximum. Otherwise, one cannot be sure, but see the discussion on asymptotic behaviour and multiple starting-points at the end of this subsection.

Third, as shown in Thm 2.2, the MLE of θ is invariant to one-to-one transformations. This implies that if h is a one-to-one function, or can be embedded in a one-to-one function, then $\widehat{h(\theta)}(\boldsymbol{y}) = h(\hat{\theta}(\boldsymbol{y}))$. Fourth, and following on from that, the MLE is a good choice for a plug-in value for θ, as shown in Thm 2.3.

Finally, there are arguments based on the performance of the MLE in the case where n is large; see Cox (2006, ch. 6) for an outline, and van der Vaart (1998, ch. 5) for details. The strongest results come in the case where $Y_1, \dots, Y_n$ is IID, i.e.

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} f_Y(y_i; \theta)$$

for some family of PMFs f_Y (see Sec. 2.A.3). Under certain conditions on this family (which are typically satisfied for the usual choices), the estimated standard error $\widehat{\mathrm{SE}}(\hat{\theta}, \boldsymbol{Y})$ is $O_P(n^{-\frac{1}{2}})$, and

$$\frac{\hat{\theta}(\boldsymbol{Y}) - \theta}{\widehat{\mathrm{SE}}(\hat{\theta}, \boldsymbol{Y})} \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1) \quad \text{for all } \theta \in \Omega, \tag{2.8}$$

where $\xrightarrow{\mathrm{D}}$ denotes convergence in distribution as $n \to \infty$.[27] In other words, when n is large enough, $\hat{\theta}(\boldsymbol{Y})$ becomes more and more bell-shaped, and more and more concentrated around θ, as n increases further. Or, to make a useful analogy, $\hat{\theta}(\boldsymbol{Y})$ behaves more and more like the sample mean of an IID sample of size n; see Sec. 2.8. While other estimators also have this property, the MLE can be shown to have the smallest mean-squared error, as $n \to \infty$.[28] These asymptotic results still hold for some relaxations of the IID property.

It is complicated to determine whether n is big enough in any particular application for (2.8) to be a good approximation. Put briefly, the asymptotic results follow from the random function $\log f_{\boldsymbol{Y}}(\boldsymbol{Y}; t)$ being dominated by an approximately quadratic term which is maximised at a value of t near to θ (van der Vaart, 1998, ch. 7). But to find the maximum of this quadratic term for given $\boldsymbol{y}$, a hill-climbing numerical maximiser must start in its basin of attraction. This basin can be hard to find if the quadratic term is

[27] There is some new notation and concepts here, but they are not important: see the references above for details.

[28] Technically, the MLE achieves the Cramér-Rao lower bound, and so cannot be beaten by any other estimator. Conditions apply, see Casella and Berger (2002, sec. 7.3).

concentrated in a small volume of the parameter space, which it may well be if n is very large and the parameter space is quite large. Thousands of different starting points may be required. If the largest maximum thus found is much larger than the next largest maximum, and if $\log f_Y(y; t)$ is approximately quadratic around this maximum, then you are entitled to hope that the sampling distribution of $\hat{\theta}(Y)$ can be well-approximated by its asymptotic properties. Simply to presume that your n is large enough without going through this process would be reckless.

2.A.2 Confidence intervals*

This is a brief outline of the theory of confidence intervals and some of their difficulties. I continue to treat θ as a scalar, for simplicity. Confidence intervals are discussed in detail in Sec. 6.5.

As discussed at the end of Sec. 2.A.1, there are situations in which the sampling distribution of the Maximum Likelihood Estimator (MLE) is approximately unbiased and Normal, see (2.8). In this case

$$\Pr_\theta \left\{ \Phi^{-1}(\alpha/2) \leq \tfrac{\hat{\theta}(Y)-\theta}{\widehat{\mathrm{SE}}(\hat{\theta};Y)} \leq \Phi^{-1}(1-\alpha/2) \right\} \approx 1-\alpha \quad \text{for all } \theta \in \Omega,$$

where Φ^{-1} is the quantile function of the standard Normal distribution, so that $-\Phi^{-1}(\alpha/2) = \Phi^{-1}(1-\alpha/2) = z_{\alpha/2}$, say. Inverting this probability statement gives

$$\Pr_\theta \left\{ \theta \in \hat{\theta}(Y) \mp z_{\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta};Y) \right\} \approx 1-\alpha \quad \text{for all } \theta \in \Omega,$$

where I write '$a \mp b$' to denote the set $[a-b, a+b]$. Thus for a given estimator the interval statistic

$$\mathcal{C}_{1-\alpha}(y) := \left\{ t \in \Omega : t \in \hat{\theta}(y) \mp z_{\alpha/2}\,\widehat{\mathrm{SE}}(\hat{\theta};y) \right\}$$

is approximately a level-$(1-\alpha)$ *confidence interval* for θ, because it satisfies the defining property that

$$\Pr_\theta \left\{ \theta \in \mathcal{C}_{1-\alpha}(Y) \right\} \approx 1-\alpha \quad \text{for all } \theta \in \Omega.$$

Note that this is a probability statement about a random interval $\mathcal{C}_{1-\alpha}(Y)$ which holds for all $\theta \in \Omega$. The most commonly computed confidence interval for θ is an approximate 95% interval,

$$\hat{\theta}(y) \mp 1.96\,\widehat{\mathrm{SE}}(\hat{\theta}, y), \tag{†}$$

where $z_{0.05/2} = 1.959964\cdots \approx 1.96$. The origins of 5% as the conventional choice for α are explored in Cowles and Davis (1982).

One of the difficulties with confidence intervals is that it is hard to construct an accurate one; i.e. to specify a $\mathcal{C}_{1-\alpha}$ that can be shown to have the required property of containing θ with probability close to $1-\alpha$ for all $\theta \in \Omega$. The asymptotic approach illustrates the problem: it is only as $n \to \infty$ that an interval such as (†) actually has a 95% probability of containing θ for all $\theta \in \Omega$, and only then for that subset of all families of distributions which is centred around the IID case.

The probability of an interval statistic $\mathcal{C}$ containing θ is termed the *coverage* of $\mathcal{C}$ at θ, and we distinguish between the nominal coverage, $1 - \alpha$, and the actual coverage. The difference between these two is termed *level error*, which will depend on θ. Modern resampling techniques, notably the *bootstrap* (see, e.g., DiCiccio and Efron, 1996; Davison *et al.*, 2003), are designed to reduce level error relative to conventional choices such as (†), but cannot eliminate it. Another approach is based on *small-sample asymptotics* (Brazzale *et al.*, 2007). There is more discussion of level error at the end of Sec. 6.5.

* * *

Confidence intervals are often misinterpreted. Whenever an interval such as "$\mathcal{C}_{0.95}(\boldsymbol{y}) = [\ell, u]$" is given, ℓ and u being specified values, remember this mantra:

> $[\ell, u]$ *is one realisation of a random interval which has the property of containing the actual value of* θ *in about 95% of cases, no matter what that value happens to be.*

It is a mistake to state that $\Pr\{\theta \in [\ell, u]\} \approx 0.95$; this would be to confuse a random interval with a random θ. The only random thing in the Frequentist approach is $\boldsymbol{Y}$.

I write "95% of cases" but this conceals a subtle philosophical issue. We have precisely one dataset, so the idea that we might understand a confidence interval through its behaviour over multiple realisations of $\boldsymbol{Y}$ is purely hypothetical, and hardly practical. So what population do we refer to when we say "95% of cases"? The most defensible interpretation is that the population is all of the 95% confidence intervals that I might compute over the course of my lifetime; or, even wider, all 95% confidence intervals that anyone might compute. Hacking (2001, ch. 22) explores this rather startling escalation, which he traces back to the American scientist and philosopher C.S. Peirce. A similar point has been made by Wasserman (2004, sec. 6.3).

But to the client, this might seem rather specious: after all, she is not interested in my performance over the course of my lifetime, but my performance right now, in her application. Moreover, clients will usually misinterpret a 95% confidence interval as a probabilistic statement about θ (making the mistake described above); what is termed a *credible interval*. It seems obtuse to provide a client with an interval which she will misinterpret (Rubin, 1984). And, unfortunately, there are many simple cases where confidence and credible intervals are quite different from one another (see, e.g., Berger and Wolpert, 1984, ch. 2). So there are fundamental difficulties with confidence intervals as an assessment of uncertainty, as well as the practical difficulty of achieving a low level error.

My preference, following the argument outlined in Sec. 2.8, would be to interpret a confidence interval in terms of the leave-one-out measure of tenability for my inference about $h(\boldsymbol{X})$, denoted $\sigma_{\text{LOO}}(h)$. In this case a 95% confidence interval for $h(\theta)$ provides a quick approximation to $\sigma_{\text{LOO}}(h)$, using (2.5). This is a purely computational result, requiring no philosophy, although, as an approximation, it too can be compromised by level error.

2.A.3 The IID model*

If $\boldsymbol{X} := (X_1, \dots, X_m)$ and the statistical model for $\boldsymbol{X}$ has the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^{m} f_X(x_i; \theta)$$

for some f_X, then we also write

$$\boldsymbol{X} \overset{\text{iid}}{\sim} f_X(\cdot; \theta)$$

or that $\boldsymbol{X}$ is *independent and identically distributed (IID)*. Under the simple observation model (eq. 2.6), the PMF of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} f_X(y_i; \theta).$$

The IID model can be used to represent the belief that $\boldsymbol{X}$ is exchangeable (Chapter 4), and is also suitable for some types of random sample (Sec. 4.4). It is extremely tractable, and, in the pre-computer era, it was ubiquitous. It is less common now for the X's themselves, but it is often to be found somewhere inside the model for the X's, to represent the judgement of partial exchangeability (Sec. 4.5).

If $\boldsymbol{X}$ is IID, then the FTP shows that

$$\mathrm{E}_\theta\{g(\boldsymbol{X}_A)\, h(\boldsymbol{X}_B)\} = \mathrm{E}_\theta\{g(\boldsymbol{X}_A)\}\, \mathrm{E}_\theta\{h(\boldsymbol{X}_B)\} \qquad (\dagger)$$

whenever A and B are non-intersecting subsets of $(1, \dots, m)$. In particular, if $g(\boldsymbol{x}_A) \leftarrow x_i$ and $h(\boldsymbol{x}_B) \leftarrow x_j$ then

$$\mathrm{E}_\theta(X_i\, X_j) = \mathrm{E}_\theta(X_i)\, \mathrm{E}_\theta(X_j) = \mu(\theta)^2$$

where $\mu(\theta) := \mathrm{E}_\theta(X_i)$. This implies that $\mathrm{Cov}_\theta(X_i, X_j) = 0$, and that

$$\mathrm{Var}_\theta\Big(\sum_{i \in A} X_i\Big) = |A|\, \sigma^2(\theta)$$

where $\sigma^2(\theta) := \mathrm{Var}_\theta(X_i)$; see Sec. 1.3.1.

For the IID model, the *sample mean*

$$\overline{\mu}(\boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} y_i$$

is an unbiased estimator for $\mu(\theta)$:

$$\mathrm{E}_\theta\left\{\overline{\mu}(\boldsymbol{Y})\right\} = \mathrm{E}_\theta\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i\right\} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_\theta(Y_i) = \mu(\theta) \quad \text{for all } \theta \in \Omega.$$

Furthermore,

$$s^2(\boldsymbol{y}) := \frac{1}{n-1}\sum_{i=1}^{n}\{y_i - \overline{\mu}(\boldsymbol{y})\}^2$$

is an unbiased estimator for $\sigma^2(\theta)$, typically termed the *sample variance*. The proof is straightforward and can be found, e.g., DeGroot and Schervish (2002, sec. 7.7).

The variance of $\overline{\mu}$ is

$$\operatorname{Var}_\theta\left\{\overline{\mu}(\boldsymbol{Y})\right\} = \frac{1}{n^2}n\sigma^2(\theta) = \frac{1}{n}\sigma^2(\theta)$$

and so the estimated standard error of the estimator $\overline{\mu}$ is

$$\widehat{\mathrm{SE}}(\overline{\mu}, \boldsymbol{y}) = \sqrt{\frac{1}{n}\, s^2(\boldsymbol{y})} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left\{y_i - \overline{\mu}(\boldsymbol{y})\right\}^2} \tag{2.9}$$

after plugging in $s^2(\boldsymbol{y})$ for $\sigma^2(\theta)$. The *Central Limit Theorem* implies that $\overline{\mu}(\boldsymbol{Y})$ has an approximately Normal distribution for large n, and hence we have the canonical approximate 95% confidence interval for $\mu(\theta)$, the expectation of X:

$$\overline{\mu}(\boldsymbol{y}) \mp 1.96\sqrt{n^{-1}s^2(\boldsymbol{y})}\,.$$

3
Utility and Decision-Making

This chapter outlines the formal theory of expected utility maximisation as an approach to choosing between actions, and the way that decision problems both without and with observations are structured (Sec. 3.2 and Sec. 3.5). Other sections are more discursive, including a summary and reflection at the end of the chapter (Sec. 3.7). The important topic of prediction is covered in Sec. 3.6. For reading, DeGroot (1970) is about statistical decisions, and Smith (2010) is about more general decisions.

3.1 Utility functions

Suppose a person, e.g. myself, is contemplating a finite set of possible *outcomes*,

$$\mathcal{R} := \{r_1, \ldots, r_k\}$$

considered by me to be mutually exclusive and exhaustive. These outcomes can be fairly arbitrary, although each one should be operationally defined, to avoid any ambiguity. Typically each outcome will comprise a collection of circumstances. Thus, "I walk to work in my sandals and it rains" would qualify as an outcome.

Now suppose that I must choose between *gambles* over the outcomes in $\mathcal{R}$. A gamble is represented by a $\boldsymbol{p} \in \mathcal{P} \subset \mathbb{S}^{k-1}$, where

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}$$

indicates a situation where I believe I will obtain outcome r_1 with probability p_1, and so on.[1] Gambles include certainty as a special case; thus the gamble with $p_i \leftarrow 1$ and all other p's zero is the gamble where I believe I am certain to obtain outcome r_i.

Clearly I have preferences about gambles; not least, because I have preferences over $\mathcal{R}$. Denote these preferences using $\preceq$, where $\boldsymbol{p} \preceq \boldsymbol{p}'$ exactly when $\boldsymbol{p}$ is not preferred to $\boldsymbol{p}'$. We will assert that my preferences are a transitive and complete order on $\mathcal{P}$:

1. If $\boldsymbol{p} \preceq \boldsymbol{p}'$ and $\boldsymbol{p}' \preceq \boldsymbol{p}''$, then $\boldsymbol{p} \preceq \boldsymbol{p}''$, (Transitive)
2. For all $\boldsymbol{p}$ and $\boldsymbol{p}'$, either $\boldsymbol{p} \preceq \boldsymbol{p}'$ or $\boldsymbol{p}' \preceq \boldsymbol{p}$.[2] (Complete)

[1] Recollect that $\mathbb{S}^{k-1}$ is the $(k-1)$-dimensional unit simplex, see (1.2). The set of available gambles may be limited, which is why I have not written $\mathcal{P} = \mathbb{S}^{k-1}$.

[2] As is usual in mathematics, 'either/or' here means 'either/or/or both'.

Technically, any $\preceq$ satisfying these two properties is termed a *weak order* on $\mathcal{P}$.[3] Transitivity is intuitive: if my preferences were not transitive then I could willingly be reduced to a state of penury in which I swapped $\boldsymbol{p}$ for $\boldsymbol{p}'$, $\boldsymbol{p}'$ for $\boldsymbol{p}''$, and then paid to swap $\boldsymbol{p}''$ for $\boldsymbol{p}$.

But completeness is a strong assertion, given that outcomes can be quite complicated, and the set of possible outcomes can be quite large. Sometimes I would have to work hard to decide whether I would rather have $\boldsymbol{p}$ or $\boldsymbol{p}'$. This effort will be justified if the decision is sufficiently important. Otherwise, I might proceed in an approximate fashion, by grouping together outcomes that are similar, and then grouping together gambles which appear to be about equally attractive, on a cursory examination. For each pair of gambles in a given group, both $\boldsymbol{p} \preceq \boldsymbol{p}'$ and $\boldsymbol{p}' \preceq \boldsymbol{p}$, indicating that I am (superficially) indifferent between $\boldsymbol{p}$ and $\boldsymbol{p}'$.

The preference relation $\preceq$ would be complicated to work with, from a mathematical point of view. It would be much simpler if we could encapsulate any preference relation in terms of a mathematical function of $\boldsymbol{p}$, termed a *utility function*.

Definition 3.1 (Utility function)**.** *A function $u : \mathcal{P} \to \mathbb{R}$ is a utility function for the preference relation $\preceq$ exactly when*

$$\boldsymbol{p} \preceq \boldsymbol{p}' \iff u(\boldsymbol{p}) \leq u(\boldsymbol{p}').$$

Luckily, if $\preceq$ is transitive and complete there are plenty of utility functions available, under the simplifying condition that $\mathcal{P}$ is finite. This is because these two properties ensure that all of the elements of a finite $\mathcal{P}$ can be ordered such that

$$\boldsymbol{p}_{(1)} \preceq \boldsymbol{p}_{(2)} \preceq \cdots \preceq \boldsymbol{p}_{(m)}.$$

Then, for example,

$$v(\boldsymbol{p}) := \sharp\{\boldsymbol{p}' \in \mathcal{P} : \boldsymbol{p}' \preceq \boldsymbol{p} \text{ and } \boldsymbol{p} \npreceq \boldsymbol{p}'\}$$

is a utility function, where $\sharp\{\cdot\}$ is the number of elements in the set.[4] This result can be extended to a non-finite $\mathcal{P}$; see, e.g., Roberts (1984, ch. 3). If g is any increasing function, then $g(v(\boldsymbol{p}))$ is also a utility function. Thus we proceed on the basis that there is indeed a utility function for my preference relation.

Now let $\boldsymbol{p}', \boldsymbol{p}'' \in \mathcal{P}$, and consider the new gamble

$$\boldsymbol{p} = \alpha\boldsymbol{p}' + (1-\alpha)\boldsymbol{p}'' \quad \text{for some } \alpha \in [0,1].$$

In this case $\boldsymbol{p}$ might be constructed by first tossing a coin to choose between $\boldsymbol{p}'$ (heads) and $\boldsymbol{p}''$ (tails). If the probability of heads is α, then this compound gamble is equivalent to $\boldsymbol{p}$. So gambles such as $\boldsymbol{p}$ are also available to me (as long as I have an appropriate coin). The crucial question is how my preferences for $\boldsymbol{p}$ relate to my preferences for $\boldsymbol{p}'$ and $\boldsymbol{p}''$. We introduce an additional property of my preferences, that of *linearity*.

[3] See Roberts (1984, ch. 1) for more details about orders on sets.

[4] I.e., $v(\boldsymbol{p})$ denotes the number of elements of $\mathcal{P}$ that are worse than $\boldsymbol{p}$.

Definition 3.2 (Linear utility functions). *A utility function is linear exactly when*

$$u\big(\alpha \boldsymbol{p}' + (1-\alpha)\boldsymbol{p}''\big) = \alpha u(\boldsymbol{p}') + (1-\alpha)u(\boldsymbol{p}'').$$

This new property may or may not hold, for a given preference relation. Whether it does hold, or ought to hold, is something I must decide for myself. There are some simple implications which are hard to deny. For example, if I am indifferent between an apple or an orange, then linearity asserts that I would also be indifferent between an apple (or an orange) or a gamble between an apple and an orange based on the toss of an arbitrary coin. There are other implications, though, that are more subtle, although I do not propose to analyse them here. Rather, I turn to the following result.

Theorem 3.1. *A utility function is linear if and only if*

$$u(\boldsymbol{p}) = \sum_{i=1}^{k} u_i\, p_i \tag{3.1}$$

where u_i is the utility of receiving outcome r_i with certainty.

Proof.
($\Leftarrow$). Suppose that $u(\boldsymbol{p}) = \sum_i u_i\, p_i$. Then

$$\begin{aligned} u\big(\alpha\boldsymbol{p} + (1-\alpha)\boldsymbol{p}'\big) &= \textstyle\sum_i u_i\, \big(\alpha p_i + (1-\alpha)p_i'\big) \\ &= \alpha \textstyle\sum_i u_i\, p_i + (1-\alpha) \textstyle\sum_i u_i\, p_i' \\ &= \alpha u(\boldsymbol{p}) + (1-\alpha)u(\boldsymbol{p}'). \end{aligned}$$

($\Rightarrow$). Suppose that u is linear in $\boldsymbol{p}$. Define $\boldsymbol{e}_i$ to be the unit vector with a 1 in the ith position; note that $\boldsymbol{e}_i \in \mathcal{P}$. Then

$$\begin{aligned} u(\boldsymbol{p}) &= u(p_1\, \boldsymbol{e}_1 + \cdots + p_k\, \boldsymbol{e}_k) \\ &= p_1\, u(\boldsymbol{e}_1) + \cdots + p_k\, u(\boldsymbol{e}_k) \qquad \text{by linearity and iteration} \\ &= \textstyle\sum_i u_i\, p_i \end{aligned}$$

where $u_i := u(\boldsymbol{e}_i)$, the utility of receiving r_i with certainty. □

How does this help? Primarily in terms of the huge dividend of simplicity. Accepting linearity, all of my preferences over $\mathcal{P}$ can be expressed in terms of k numerical values for the outcomes in $\mathcal{R}$. Furthermore, these k values can be elicited as follows.

First, note that if u is linear, then so is

$$v(\boldsymbol{p}) := a + bu(\boldsymbol{p}) = \textstyle\sum_i (a + bu_i)p_i.$$

Furthermore,

$$\begin{aligned} v(\boldsymbol{p}) \le v(\boldsymbol{p}') &\iff v(\boldsymbol{p}) - v(\boldsymbol{p}') \le 0 \\ &\iff \textstyle\sum_i (a + bu_i)(p_i - p_i') \le 0 \\ &\iff b\big(u(\boldsymbol{p}) - u(\boldsymbol{p}')\big) \le 0 \\ &\iff u(\boldsymbol{p}) \le u(\boldsymbol{p}') \qquad \text{if } b > 0. \end{aligned}$$

Hence linear utility functions are invariant under positive linear transformations. As $\mathcal{R}$ is finite, there is a least-preferred outcome and a most-preferred outcome, with indices denoted $\underline{r}$ and $\overline{r}$ respectively. Assign these the utility values $u_{\underline{r}} \leftarrow 0$ and $u_{\overline{r}} \leftarrow 1$, which is permissible because utility is invariant under positive linear transformations. Then note that

$$\begin{aligned} u_i &= (1 - u_i) \cdot 0 + u_i \cdot 1 \\ &= (1 - u_i) \cdot u_{\underline{r}} + u_i \cdot u_{\overline{r}} \\ &= u\big((1 - u_i)\, \boldsymbol{e}_{\underline{r}} + u_i\, \boldsymbol{e}_{\overline{r}}\big) && \text{by linearity.} \end{aligned}$$

Hence, if $u_{\underline{r}} = 0$ and $u_{\overline{r}} = 1$, then u_i is precisely the probability at which I am indifferent between receiving r_i with certainty, and a gamble of $\underline{r}$ versus $\overline{r}$ with probabilities $(1 - u_i)$ and u_i.

So linearity goes further than providing a simple form for the utility function: it also provides us with a thought experiment through which the utility of each outcome in $\mathcal{R}$ can be elicited. For example, suppose that, for me, $\underline{r}$ corresponds to a day in a small boat, and $\overline{r}$ corresponds to a day in a wood. What about a day in the library? I ask myself: at what p would I be indifferent between a gamble of $1 - p$ probability of the boat versus a p probability of the wood, and the certainty of the library? Since a day in the library is quite attractive to me, almost as attractive as a day in the wood, I would set p quite high, since the gamble offers outcomes that are much worse and only a little better. So on a scale from 0 to 1, my utility for the library is, say, 0.9. In Sec. 3.3 I discuss how I might incorporate other suppressed factors into my utility assessment, such as—in this case—the weather.

With these powerful simplifications in mind, I should have a predisposition to linearity, unless I have clear indication in my preferences that linearity does not hold. Where linearity does not hold, I still have a utility function, but I do not know its form without much further reflection, which would undoubtedly be more arduous than the reflection necessary to determine the utility scores of the elements of $\mathcal{R}$. For agents making choices on behalf of other people, I would say that linearity is more-or-less mandatory, because its simplicity implies transparency.

When my utility function is linear, the utility value $u(\boldsymbol{p})$ given in (3.1) has the form of my expectation of u_R supposing that my probability distribution for R was $\Pr(R \doteq r_i) \leftarrow p_i$, according to the FTP (Thm 1.5). In this case, choosing the gamble with the largest utility is termed *expected utility maximisation (EUM)*. I will assume EUM from now on. Sec. 3.7 contains reflections on EUM in the context of the material in the previous chapters.

* * *

Other approaches for choosing between elements of $\mathcal{P}$ according to a weak order $\preceq$ derive the linear form of the utility function given in (3.1) from a set of simpler properties about preferences; see, for example, DeGroot (1970). I do not find the union of these simpler properties any more compelling than the linearity property (which is not surprising because they are effectively equivalent). For me, it is the simplicity and the transparency of the linearity property which appeals; plus my attitude that linearity is not obviously wrong and may indeed often be appropriate.

3.2 *Simple decision problems*

This section simply repackages EUM of the previous section in the form of a decision problem. This material, plus that in Sec. 3.5, is the basis of *Statistical Decision Theory*.

The two primitives of a decision problem are a set of possible actions,

$$a \in \mathcal{A} := \{a^{(1)}, \dots, a^{(m)}\},$$

the *action set*, and a set of random quantities,

$$X \in \mathcal{X} := \{x^{(1)}, \dots, x^{(s)}\},$$

the *predictands* (although calling X the 'state of nature' is common). X is typically a vector of random quantities, but in this chapter I will treat it as a scalar, simply to avoid too much ink on the page. An outcome combines both an action and a value for the predictands,

$$r = (a, x) \in \mathcal{A} \times \mathcal{X}.$$

The problem to be solved is simply which action I should choose, given my preferences on the outcomes, and beliefs about the predictands.

Accepting the properties of my preferences given in the previous sections, I can score each outcome according to its utility, $u_r = u_{a,x}$. Rather than utility, though, it is conventional to work in terms of its converse, termed *loss*, and defined as

$$L(a, x) := -u_{a,x}.$$

Thus EUM asserts that I should choose from among actions by selecting the one which minimises my expected loss.

My choice of action affects my probability distribution over outcomes. Thus I write $p_{a',x;a}$ to denote my probability for outcome (a', x) when choosing action a. Assuming that I am effective in implementing my choice of action, the probability of actions that are not a is zero, and hence

$$p_{a',x;a} = \begin{cases} f_X(x; a) & a' = a \\ 0 & \text{otherwise,} \end{cases}$$

where $f_X(x; a)$ is the probability I specify for $X \doteq x$ under action a. Thus my choice of action *always* affects my probabilities when the outcome is defined to include my action, but it may also affect my probabilities for the predictands.

For example, suppose that

$$\mathcal{A} = \{\textsf{wear sandals}, \textsf{wear shoes}\} \quad \text{and} \quad \mathcal{X} = \{\textsf{dry}, \textsf{rainy}\}.$$

I do not believe that my choice of footwear influences the weather, although it might sometimes seem like that, and hence, for me, f_X is invariant to a. On the other hand, suppose that

$$\mathcal{A} = \{\textsf{don't cloud seed}, \textsf{cloud seed}\} \quad \text{and} \quad \mathcal{X} = \{\textsf{dry}, \textsf{rainy}\}.$$

In this case, the purpose of the action is to influence the weather, and so there is a *prima facie* case that f_X varies with a.[5]

The expected loss of an action a is termed the *risk* of action a,

$$R(a) := \sum_{a',x} L(a', x)\, p_{a',x;a} = \sum_{x} L(a, x)\, f_X(x; a). \tag{3.2}$$

This is a smooth function of f_X, and so small perturbations in my $f_X(\cdot; a)$ will cause only small perturbations in my $R(a)$. According to EUM, a best action is any action which minimises the risk, termed the *Bayes action*,

$$a^* := \operatorname*{argmin}_{a \in \mathcal{A}} R(a)$$

and the resulting risk is the *Bayes risk*, $R(a^*)$.

Ordering all of the actions according to their risks, typically we would have strict inequalities of the form

$$R(a^*) < R(a') < \cdots,$$

where action a' has the next-smallest risk after a^*. Small perturbations in my $f_X(\cdot; a^*)$ and $f_X(\cdot; a')$ will cause only small perturbations in $R(a^*)$ and $R(a')$, and are unlikely to reverse my preferences between the two actions. So in fact I do not have to agonise about my probability distributions unless I discover that $R(a^*)$ and $R(a')$ are very similar. In this situation, though, I am more likely to take into account the non-quantifiable aspects of a^* and a', rather than try to resolve my difficulties by tweaking my probability distributions.

3.3 *Structuring decision problems*

There is no unique way to structure a decision problem, even in the case where the action set and loss function are unambiguous.

For example, suppose that you are the mayor of a coastal town, worried about the effect of sea-level rise on flooding, and contemplating building a sea-wall. Let $(X, Y_1, \ldots, Y_k)$ be the predictands, where X is the maximum sea-level in the harbour in 2050, and Y_j is 0 if property j is not flooded in 2050, and 1 if it is. Let a be the height of the proposed sea-wall. With this much detail in the predictands, the loss function might be as simple as

$$L(a, x, \boldsymbol{y}) = c(a) + c_y \textstyle\sum_j y_j,$$

[5] See Levin *et al.* (2010) for an assessment of the long-running Israeli cloud seeding experiment.

where $c(a)$ is the cost in £ of a metres of sea-wall, and c_y is a conversion factor with units of £ per property. More complicated functions of $\boldsymbol{y}$ are also possible, for example to account for the different values of different properties, or to account for your preferences, which may not be linear in the number of flooded properties.

Realistically, though, the level of detail represented by the y's is way beyond your budget. So instead, you suppress the y's in your assessment. It is quite clear from the structure of the risk function how this suppression must operate, because

$$\begin{aligned} R(a) &= \textstyle\sum_x \sum_{\boldsymbol{y}} L(a, x, \boldsymbol{y})\, f_{X,\boldsymbol{Y}}(x, \boldsymbol{y}; a) \\ &= \textstyle\sum_x \sum_{\boldsymbol{y}} L(a, x, \boldsymbol{y})\, f_{\boldsymbol{Y}|X}(\boldsymbol{y} \mid x; a)\, f_X(x; a) \quad \text{by (1.7)} \\ &= \textstyle\sum_x L(a, x)\, f_X(x; a) \end{aligned}$$

where

$$\begin{aligned} L(a, x) &:= \textstyle\sum_{\boldsymbol{y}} L(a, x, \boldsymbol{y})\, f_{\boldsymbol{Y}|X}(\boldsymbol{y} \mid x; a) \\ &= \mathrm{E}\{L(a, x, \boldsymbol{Y}) \mid x; a\} \qquad \text{by the CFTP, (1.6).} \end{aligned}$$

Hence your loss function over (a, x) should be your conditional *expected* loss function over $(a, x, \boldsymbol{Y})$. This illustrates a very important point:

> *Every loss function should be a conditional expected loss function, where the expectation is taken over all of the random quantities not present in the predictands.*

With the limited budget at your disposal, you and your experts must make the best loss assessment that you can—this assessment is nothing other than your expectation.

If your budget goes up, you can make previously suppressed random quantities explicit. For example, you might commission an elevation map of the town showing every property. This will allow you to group properties by elevation, and provide a much tighter coupling between the value of $X - a$ and the number of flooded properties. The question of how you spend your budget most effectively is also a decision problem; see Smith (2010, ch. 2).

In structuring your decision problem, you can do more than just suppress predictands that occur in your loss function. Another good modelling strategy to expand the set of predictands to include 'exogenous' predictands, which I will denote here as W. By 'exogenous' I mean that your distribution of W is unaffected by your choice of a. For example, W might be global sea-level rise by 2050. The specification of f_W is not really your concern, but ought to be provided by some central group of experts: in the case of sea-level rise, perhaps the scientists on the Intergovernmental Panel on Climate Change (IPCC). You and your experts can then concentrate on $f_{X|W}$ for your town. You will undoubtedly find it much easier to specify $f_{X|W}$ and then to deduce f_X from the Law of Total Probability (Thm 1.25)

$$f_X(x; a) = \textstyle\sum_w f_{X|W}(x \mid w; a)\, f_W(w),$$

using the IPCC f_W, than to specify f_X directly. O'Hagan (2012) refers to this type of elicitation strategy as *elaboration*.

3.4* A couple of digressions

Here are two issues that often come up in Decision Theory, which are red-herrings, but interesting nonetheless.

3.4.1* Mixed actions

The actions in $\mathcal{A}$ do not exhaust my possibilities. I can also choose an action at random according to any probability vector $\boldsymbol{w} \in \mathbb{S}^{m-1}$, where w_a is the probability of my choosing action a. Actions from $\mathcal{A}$ are termed *pure actions*, while actions chosen according to some probability vector $\boldsymbol{w}$ are *mixed actions*.

In fact, mixed actions are not very interesting, as they do not allow for any improvement in the Bayes risk over pure actions, but this is something that must be proved.

Theorem 3.2. *Let $A \in \mathcal{A}$ be a random quantity for which*

$$f_{X,A}(x,a) = f_X(x;a) \cdot w_a$$

where $\boldsymbol{w} \in \mathbb{S}^{m-1}$ is specified. Let $R(\boldsymbol{w}) := \mathrm{E}\{L(A,X);w\}$. Then $R(\boldsymbol{w}) \geq R(a^*)$ for all $\boldsymbol{w}$, where a^* is a Bayes action.

Proof. Starting with the LIE (Thm 1.21),

$$\begin{aligned} R(\boldsymbol{w}) &= \textstyle\sum_a \mathrm{E}\{L(a,X);a\} \cdot w_a \\ &= \textstyle\sum_a R(a) \cdot w_a \\ &\geq \min_{\boldsymbol{w}' \in \mathbb{S}^{m-1}} \textstyle\sum_a R(a) \cdot w'_a \\ &= \min_a R(a) = R(a^*), \end{aligned}$$

where $a^* \in \mathcal{A}$ is a pure Bayes action. □

This result confirms what many people would feel was entirely intuitive: that I can gain no benefit by bringing a random component into my decision-making.

3.4.2* Minimax actions

There will be times when specifying a probability distribution f_X seems hard, perhaps somewhat arbitrary. And this task only gets harder in the situation where f_X varies with the choice of action. In this case it is natural to ask whether there is another approach to choosing between actions which uses only my loss function L.

One such approach is to choose the *minimax action*. This is the action for which the maximum loss (across x) is minimised:

$$a_{\mathrm{M}} := \min_{a \in \mathcal{A}} \max_{x \in \mathcal{X}} L(a,x),$$

if we restrict ourselves to pure actions.[6] Minimax represents a

[6] It would be better to define this as the 'minimax loss' action, in contrast to 'minimax regret', defined below.

scenario in which we play first, choosing an a, and then nature does its worst. In the case where $\mathcal{X} = \{x^{(1)}, x^{(2)}\}$ we can visualise the minimax rule, as in the lefthand panel of Figure 3.1.

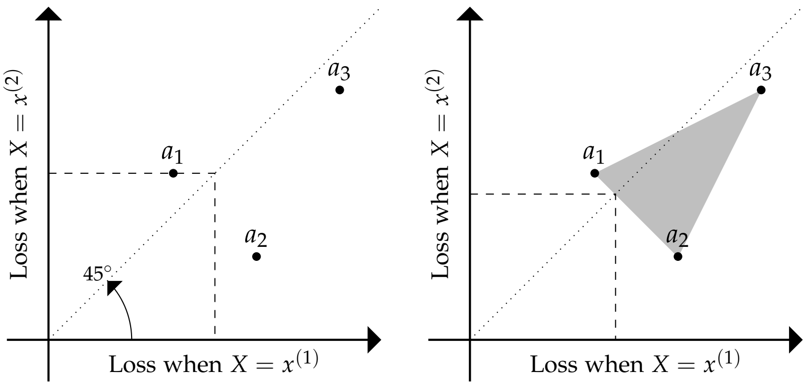

Figure 3.1: Minimax actions in the case where $\mathcal{X} = \{x^{(1)}, x^{(2)}\}$. Three actions are shown as dots, where each dot represents the point $\big(L(a, x^{(1)}), L(a, x^{(2)})\big)$. The dashed lines in the lefthand panel show that a_1 is the pure minimax action. In the righthand panel, every point in the shaded polygon represents a mixed action; the mixed action $(0.75, 0.25, 0)$ is the minimax action, which has a lower minimax loss than any pure action.

Figure 3.1 also illustrates a major issue with the minimax rule, which is that pure rules are almost always bettered by mixed rules (see Sec. 3.4.1), as shown in the righthand panel. So, from this point of view alone, anyone prepared to choose the minimax action has to accept that he would derive a better minimax action than the pure minimax action by specifying a $\boldsymbol{w}$ and then tossing coins.[7] This has been discussed in standard references,[8] but I venture that many non-statisticians who advocate minimax for challenging decisions do not appreciate that minimax actions are typically mixed actions.[9]

There are other reasons to dislike the minimax action as well: why would I want to choose an action based on a worst-case for the predictands that had only a tiny probability of occurring? In order for minimax to suggest sensible actions, it seems as though I would have to set a probability threshold and screen out low-probability elements in $\mathcal{X}$ before I started (e.g. super-volcano eruptions), which is rather self-defeating if the purpose of minimax is to evade the specification of probabilities for $\mathcal{X}$.

Savage (1951, 1954) tried hard to find a justification for the minimax action, including refining the idea of minimax loss to consider instead *minimax regret*, in which the loss function is first transformed using

$$L(a, x) \leftarrow L(a, x) - \min_{a' \in \mathcal{A}} L(a', x).$$

Thinking of L as a matrix, the effect of the transformation is to rebase each column to have a minimum value of zero. The use of regret rather than loss dampens the 'ultra-pessimistic' aspect of minimax loss. For example, using regret can be effective in screening out high-impact low-probability elements of $\mathcal{X}$, for which losses are uniformly high across all actions in $\mathcal{A}$. Savage had a different motivation for using regret rather than loss (Savage, 1954, sec. 9.8).

But, as Savage himself stated, "It seems clear that no categorical justification for the minimax [regret] rule can be given." (Savage,

[7] Note that trying to distinguish between 'pure' loss and expected loss is futile, as explained in Sec. 3.3.

[8] E.g. DeGroot (1970, sec. 8.7). Hacking (1965, ch. 4) is also interesting.

[9] See Kunreuther *et al.* (2013) for a recent example.

1951, p. 61). His tentative suggestion was that it might be appropriate when a person was required to choose an action on behalf of a group whose membership was unknown to him. In his introduction to the 1972 reissue of Savage (1954), he reflected on his previous optimism:

> Freud alone could explain how the rash and unfulfilled promise on page 4 [which promised a justification of minimax and other approaches] went unamended through so many revisions of the manuscript. ... Among the ill-found frequentist devices are minimax rules. (p. iv).

* * *

I should also clarify, since it seems to come up frequently, that minimax is not equivalent to the *precautionary principle*. There are many interpretations of this principle, but none of them states "act as though nature is out to get you". One widely accepted definition is Principle 15 of the Rio Earth Summit, which states

> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation. (Rio Declaration on Environment and Development, 1992)

I would interpret this principle as recommending Bayes actions which are *robust* to a range of specifications for f_X. As already discussed in Sec. 3.2, such robustness is often inherent in the Bayes action for a particular f_X. Where there is scientific uncertainty, it seems much more sensible to look for actions whose near-optimality is robust over competing choices for f_X, rather than actions which pay no attention at all to f_X; see Rubin (1984, sec. 2).

3.5 Decision rules

Suppose that between now and the time where I must choose an action, I will learn the value of some *observations* $Y \in \mathcal{Y}$. Insofar as these observations are informative about the predictands X, I will be able to use them to improve my choice of action over the choice I would make now, pre-observations.

One thing I can do is wait until I learn Y, and then choose my action. In this case, I simply incorporate Y into my beliefs (see Chapter 2), and minimise my expected loss across the set of actions, according to the EUM (Sec. 3.2). But I might also be interested in assessing right now how learning Y will affect my expected loss. Two obvious motivations are

1. To price the observations Y, if I have to pay for them before learning their value, and

2. To choose between different possible sets of observations, $Y, Y', \ldots$ termed *experimental design*.

More generally, we might want to provide simple rules for how to use Y, a 'play book' for harassed risk managers and statisticians.

Observations are incorporated into decision theory by introducing a *decision rule*,

$$\delta : \mathcal{Y} \to \mathcal{A},$$

where '$\delta(y) = a$' is a rule which states 'choose action a if $Y \to y$'. According to EUM, the best decision rule is a *Bayes rule* satisfying

$$\delta^*(y) = \operatorname*{argmin}_{\delta \in \mathcal{D}} R(\delta),$$

where $\mathcal{D}$ is the set of all decision rules, and R is the *risk function*, defined as the expected loss based on the choice of decision rule:

$$R(\delta) := \mathrm{E}\left\{L(\delta(Y), X); \delta\right\}.$$

My joint distribution over (X, Y) depends on δ, because my distribution for X depends on Y and my choice of action a. But my marginal distribution for Y cannot depend on δ, because Y is observed before a is enacted. Hence my joint distribution must factorise as

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

where $f_{X|Y}(x \mid y; a)$ is my conditional distribution for X given Y if I enact action a. This expression is the basis for a remarkable result.

Theorem 3.3 (Bayes Rule theorem, BRT)**.**

$$\delta^*(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\left\{L(a, X) \mid Y \doteq y; a\right\}. \tag{3.3}$$

Proof. Let δ be an arbitrary rule. Then, starting with the definition of the risk function,

$$\begin{aligned}
R(\delta) &= \sum_x \sum_y L(\delta(y), x) \cdot f_{X,Y}(x, y; \delta) && \text{by the FTP, Thm 1.5} \\
&= \sum_y \sum_x L(\delta(y), x) \cdot f_{X|Y}(x \mid y; \delta(y)) \cdot f_Y(y) && \text{by (†)} \\
&\geq \sum_y \min_a \sum_x L(a, x) \cdot f_{X|Y}(x \mid y; a) \cdot f_Y(y) && \\
&= \sum_y \min_a \mathrm{E}\left\{L(a, X) \mid Y \doteq y; a\right\} \cdot f_Y(y) && \text{by the CFTP, (1.6)} \\
&= \sum_y \sum_x L(\delta^*(y), x) \cdot f_{X|Y}(x \mid y; \delta^*(y)) \cdot f_Y(y) && \text{by (3.3)} \\
&= \sum_y \sum_x L(\delta^*(y), x) \cdot f_{X,Y}(x, y; \delta^*) && \text{(†) again} \\
&= R(\delta^*) && \text{FTP again}
\end{aligned}$$

as required. □

The BRT is remarkable because minimisation of the risk function R over the set of all functions from $\mathcal{Y}$ to $\mathcal{A}$ seems to be a formidably difficult task; and yet it can be accomplished relatively easily by a set of discrete minimisations over $\mathcal{A}$, one for each element in $\mathcal{Y}$.

* * *

Once I have δ^* I can compute my expected loss with and without access to the observations Y, and in this way I can price the observations in units of loss, to discover the maximum that I would pay for Y. The obvious but easily missed point is that the value of the observations to me depends not only on my loss function but also my action set: a larger action set allows me to extract more value from the same set of observations.

There is also a calculation I can do to determine the very maximum I would pay for any observations. For we can suppose that the observations are perfectly informative about X. In this case I get to choose a knowing X, and my expected loss is

$$R^{**} := \mathrm{E}\{\min_{a \in \mathcal{A}} L(a, X)\}.$$

Intuitively, but also mathematically, this can never exceed my Bayes risk $R(a^*)$. The difference is termed the *Expected Value of Perfect Information (EVPI)*,

$$\mathrm{EVPI} := R(a^*) - R^{**}.$$

Clearly, if anyone is offering to sell me some observations for more than my EVPI (converted from loss to currency units) then I know that I will be better off not buying the observations.

3.6 Point prediction

Point prediction is providing a single value for a random quantity. It is best thought of as a decision problem, because the nature of the prediction should depend on what it is being used for. Sometimes it is better to err on one side than the other, sometimes it is better to be more accurate for small values than for large ones, and so on. These issues can be encoded into a loss function $L(x', x)$, the loss incurred from predicting X as x' when it is actually x. The point prediction is then the solution

$$\tilde{x}(y) = \operatorname*{argmin}_{x' \in \mathcal{X}} \mathrm{E}\left\{L(x', X) \mid Y \doteq y\right\}$$

according to the Bayes Rule theorem (Thm 3.3), where y is the value of the observations. Clearly, $\tilde{x}(y)$ depends on the loss function and the conditional PMF $f_{X|Y}$. In this type of decision problem, $f_{X|Y}$ is not affected by the choice of x'.

Milner and Rougier (2014) provide a short example of how a statistician and client can work together to choose a loss function for prediction. But how might a statistician proceed when the client has no clear idea about what the loss function might be? There seems to be an accepted answer to this question, based on the view that the loss function for prediction will often be a convex function of $x' - x$, minimised at 0. In this case, the *quadratic loss function*

$$\tilde{L}(x', x) = (x' - x)^2$$

is a reasonable approximation to L. This follows from a second-order Taylor Series expansion of $L(x' - x)$ around $x' - x = 0$,

$$L(x' - x) = L(0) + (x' - x)L'(0) + \tfrac{1}{2}(x' - x)^2 L''(0) + \cdots$$

where $L(0) = 0$ and $L'(0) = 0$, and $L''(0) > 0$ by convexity. An intuitive feature of quadratic loss is that the sum of losses over a set of predictands is the squared Euclidean distance between the predictions and the actual values.

As an alternative route to quadratic loss, we can following the same reasoning of Sec. 1.7.1. The ideal for a prediction would be that $\tilde{x}(Y)$ was effectively equivalent to X, in which case $\tilde{x}(Y)$ and X should be mean-square equivalent. Thus our objective should be to minimise the expected quadratic loss. From this point of view, the following result will be entirely unsurprising.

Theorem 3.4. *The Bayes rule for prediction under quadratic loss is*

$$\tilde{x}(y) = \mathrm{E}(X \mid Y \doteq y),$$

provided that $\Pr(Y \doteq y) > 0$.

Proof. Here is a direct proof.

$$\begin{aligned}\tilde{L}(x' - x) &= (x' - x)^2 \\ &= \big(x' - \psi(y) + \psi(y) - x\big)^2 \\ &= \big(x' - \psi(y)\big)^2 + 2\big(x' - \psi(y)\big)\big(\psi(y) - x\big) + \big(\psi(y) - x\big)^2,\end{aligned}$$

where $\psi \in \mathcal{E}(X \mid Y)$; see Sec. 1.7.2. Now take expectations of $L(x' - X)$ conditional on $Y \doteq y$ to give

$$\mathrm{E}\{L(x' - X) \mid Y \doteq y\} = \big(x' - \psi(y)\big)^2 + \mathrm{E}\{\big(\psi(y) - X\big)^2 \mid Y \doteq y\},$$

the middle term having expectation zero because $\psi(y) = \mathrm{E}(X \mid Y \doteq y)$ by Thm 1.18. This is minimised over x' at $x' \leftarrow \psi(y)$, as required. □

There are other 'off-the-shelf' loss functions for prediction as well. See, for example, Bernardo and Smith (2000, sec. 5.1), who explain how other natural location summaries such as the median and the mode of $f_{X|Y}(\cdot \mid y)$ arise as Bayes rules for simple loss function.

* * *

Prediction is everywhere in Data Science. To give a well-known and much-discussed example, consider the Netflix prize.[10] A huge matrix $\mathbb{X} := \{X_{ij}\}$ encodes the ratings of users (down the rows) for films (along the columns). So $X_{ij} \in \{1, \dots, 5\}$ is user i's rating for film j, in 'stars'. Most of the cells of $\mathbb{X}$ are empty, and the filled-in

[10] https://en.wikipedia.org/wiki/Netflix_Prize

cells are the observations, Y. The challenge is to predict the X_{ij}'s in the empty cells, so that Netflix will know what films to recommend to user i. Interestingly, predictions outside $\mathcal{X}$ are allowed; e.g., a prediction of '4.15 stars' is valid. So we have

$$\mathcal{A} = \mathbb{R}, \quad L : \mathbb{R} \times \mathcal{X} \to \mathbb{R}, \quad \delta : \{\mathcal{X}_{ij}\}_A \to \mathbb{R},$$

where A is the subset of observations.

The Netflix prize is actually a decision problem about model choice, which will be discussed in Sec. 6.2. But for now, suppose that a team entering the competition have decided on their f_X. In this case, f_X and the Netflix prize loss function completely determine the prediction rule, according to the Bayes Rule theorem. The Netflix loss function is root mean squared error, which is the square root of the sum of quadratic losses over a set of predictions. This choice strikes me as rather artificial in this context, an opinion which is apparently shared by others. Nevertheless, it makes the team's task easier, because now they know that their Bayes rule for predicting X_{ij} using Y is simply

$$\delta^*_{ij}(y) = \mathrm{E}(X_{ij} \mid Y \doteq y).$$

The challenge is to compute this expectation over thousands of users and thousands of films.

Data scientists have developed some very cunning solutions to this type of application, whose essential feature is that they scale well to huge $\mathbb{X}$.[11] Typically, computation trumps everything else, and the prediction rule is expressed as an algorithm which takes y as input, and outputs a prediction for a specified set of X's. Whether, in the case of the Netflix prize, this prediction is the conditional expectation with respect to a coherent f_X is moot. It is probably better to conceive of good Data Science algorithms as getting as close to the Bayes rule as is computationally possible.

3.7 *Reflections on utility and Decision Theory*

This chapter is somewhat abstract, compared to the previous two, and says less about the day-to-day practice of statistical inference. Perhaps it would be helpful for me to summarise why I think it is helpful.

First, it is useful to know that there is a calculus for making decisions under uncertainty. This calculus is based on preferences, making it far more widely applicable than a calculus based purely on monetary values (although these can be used to define preferences, if that is appropriate). Furthermore, under the assumptions of Sec. 3.1, this calculus can be implemented with a transparent assessment of outcomes and their utilities. Which is not to say, of course, that the assessment process will not be arduous.[12] But at least there is a clear prescription of what must be done. The outcomes must be listed, ranked, and then scored according to their

[11] An effective Data Science algorithm can be broken down into a sequence of steps, where each step involves many operations which can operate in parallel on small subsets of the dataset. See, for example, O'Neil and Schutt (2014, chs. 8 and 14).

[12] See Aspinall and Cooke (2013) for a survey of expert elicitation, or O'Hagan *et al.* (2006) for a book-length treatment.

utilities. Or, in the terms more directly associated with Decision Theory (Sec. 3.2), the client must identify her action set and loss function.

Second, Expected Utility Maximisation (EUM) provides a justification for the orientation of Chapter 1 and Chapter 2, which was expressed in rather general terms as computing expectations of functions of the random quantities. The risk of action a is exactly this: the expectation of the random quantity $L(a, X)$. According to EUM, I must compute the risk for each of possible action, in order to identify the best one. Chapter 1 and Sec. 2.2 recognised our limitations as uncertainty assessors, and explained why my $\mathrm{E}\{L(a, X)\}$ would often be undecided, and represented as an interval. Sec. 2.3 *et seq.* explored the bolder practice of collapsing every expectation to a point value, which is modern statistical practice. In this case, EUM gives an unambiguous answer about which action to choose. And it has the attractive property of being somewhat robust to the precise specification of my uncertainty about X, as discussed at the end of Sec. 3.2.

Third, the Bayes Rule theorem (BRT, Thm 3.3) provides a justification for the practice of incorporating knowledge through conditioning. One argument for this was given in Sec. 2.1. If Q is a random proposition which is believed to be true and for which $\Pr(Q) > 0$, then $\mathrm{E}^*(\cdot) := \mathrm{E}(\cdot \mid Q)$ has several very attractive features. I find this argument fairly compelling. But the BRT provides a different justification, based on EUM. If I accept EUM then my optimal rule for how I should behave if I find that Q is true is found by solving the no-data problem (Sec. 3.2) conditional on Q. This suggests that the truth of Q should be incorporated into my beliefs by conditioning.

However, we must acknowledge the decisive effect of the linearity property of utility, Def. 3.2. Without it, the EUM is not implied, although it may still be adopted as a reasonable way to proceed to a choice among actions. In particular, when the client is making choices on behalf of her stakeholders, she may welcome the simplicity and transparency of EUM, even if she cannot establish the linearity property of her stakeholders' preferences. A pragmatic adoption of EUM does not commit the client to choosing the action with the smallest expected loss. Instead, she may use it simply to rule out actions that are obviously deficient, in order to focus attention on a more manageable subset. Within this subset, she may attempt to refine the loss function, or she may introduce other aspects of the loss that she was unable to quantify, and use these to make her choice.

In really difficult applications with huge uncertainty, such as regulation, the role of the statistician is not simply to help the client to chose an action, or a subset of actions. The statistician's 'meta-role' is to improve the transparency of the client's assessment, in order that it is both defensible and 'auditable'. It is not easy being a regulator. I have the 2010 OECD publication *Risk and Regulatory Policy:*

Improving the Governance of Risk[13] on my to-read list (251 pages).

[13] http://dx.doi.org/10.1787/9789264082939-en

4 Exchangeability

This is the first of two chapters about statistical modelling. This chapter studies a model for random quantities that are 'similar but not identical'—exchangeability. The detailed study of exchangeability in statistics originated with de Finetti (1937). Kingman (1978), Aldous (1985), and Schervish (1995, ch. 1) provide excellent and complementary surveys.

Sec. 4.1 defines exchangeability and discusses the situations where it is a useful description of beliefs. Sec. 4.2 and Sec. 4.3 consider the main implications of exchangeability for expectations and for probability mass functions (PMFs). Sec. 4.4 considers the most popular model in statistics: the independent and identically distributed (IID) model; Sec. 4.5 considers how a generalisation of exchangeability underpins linear and multilevel modelling. Finally, Sec. 4.6 considers the close relationship between relative frequency histograms and probabilities, to clarify under what conditions the first can stand in for the second.

4.1 Definition

In keeping with the general tenor of these notes, I will define exchangeability in terms of expectations.

Definition 4.1 (Exchangeable random quantities)**.** *A collection of random quantities* $\boldsymbol{X} := (X_1, \dots, X_m)$ *is* m-exchangeable *exactly when*

$$\mathrm{E}\{g(X_1, \dots, X_m)\} = \mathrm{E}\{g(X_{\pi_1}, \dots, X_{\pi_m})\}$$

for all g*, where* $(\pi_1, \dots, \pi_m)$ *is any permutation of* $(1, \dots, m)$.[1]

Typically we would just say 'exchangeable' rather than 'm-exchangeable', but it is sometimes useful, and indeed necessary, to stress the length of the vector (see Sec. 4.4).

This definition asserts that the labels of the random quantities, the i's on the X_i's, convey no information that is relevant to my beliefs about $\boldsymbol{X}$. This might be because some information has been withheld from me. For example, the X_i's might be heights of children, so that each i is a name. In this case, for older children, I might make a distinction between the heights of boys and girls. So,

[1] If you are bold enough to consider random quantities with unbounded realms, then this definition should be qualified with 'for all bounded continuous g'. An equivalent statement is that $(X_1, \dots, X_m)$ is m-exchangeable exactly when $(X_1, \dots, X_m)$ and $(X_{\pi_1}, \dots, X_{\pi_m})$ are *equal in distribution* for all permutations π.

for example, I would specify $\mathrm{E}(X_i) > \mathrm{E}(X_j)$ if i was a boy's name and j was a girl's name. This would violate exchangeability for $g(\boldsymbol{x}) \leftarrow x_1$.

This example illustrates how exchangeability captures the notion of 'similar but not identical'. I do not want to assert that X_i tells me nothing about X_j (because I know children have many features in common), and yet I do not want to assert that $X_i = X_j$ (because I know that children are different one from another). The compromise is to express my beliefs about $\boldsymbol{X}$ in terms of equality of expectations.

Exchangeability is at the heart of statistical inference, precisely because 'similar but not identical' is at the heart of statistical modelling of collections of random quantities. The most obvious example is when we collect a sample to study a population: it is because we believe that the population is similar but not identical that the sample can be informative about it. But exchangeability is used in many other places as well, such as regression modelling, time-series and spatial modelling. Often it lies concealed 'below' the X_i's, occupying the gap under our well-formed beliefs about each X_i; see Sec. 4.5.

Sometimes, we find it convenient to treat $\boldsymbol{X}$ as an exchangeable sequence despite our beliefs; this is usually implemented by deliberately ignoring information. For example, I might treat all of the boys in a class as exchangeable, even if I knew additional information about each one, such as his date of birth, or ethnicity; I'll refer to this as *anonymisation*. This strategy offers both advantages and disadvantages.

The first advantage is that my inference is protected against biases in my beliefs. These biases are ubiquitous, and have now been very extensively studied; see, e.g., Kahneman (2011).[2] I should be wary about situations where I may not know as much as I think I know. Second, I make the analysis much simpler to implement. It is less effort for me, as I do not have to model the effect of the additional information on my beliefs, and it leads to a much more rapid calculation. Third, my inference is likely to be much more acceptable to others, if it reduces the opportunities for my subjective beliefs to influence the result.

But this also highlights the main disadvantage of anonymisation. If my beliefs are non-exchangeable and non-biased, then I would do better to incorporate them into my inference than to suppress them. And if I were an expert, others should value my beliefs and favour an inference in which they are incorporated. By-and-large, though, the advantages of anonymisation seem to have prevailed over the disadvantages, and this is why most statistical inferences contain exchangeable components following anonymisation.

[2] If I had to select just one nugget from this excellent book, it would be the idea of a 'pre-mortem', ch. 24.

4.2 Properties of exchangeable random quantities

One implication of exchangeability is immediate. if $\boldsymbol{X}$ is exchangeable then

$$\operatorname{supp}(\boldsymbol{X}) = \operatorname{supp}(X_1)^m \tag{4.1}$$

(recollect that 'supp' is the support of a vector of random quantities, see Sec. 1.8.3). For if the support of X_i and X_j were not the same, then there would be an (x, x') pair for which

$$\Pr(X_i \doteq x, X_j \doteq x') \neq \Pr(X_i \doteq x', X_j \doteq x).$$

This would violate the definition of exchangeability for $g(\boldsymbol{x}) \leftarrow \mathbb{1}_{x_i \doteq x} \mathbb{1}_{x_j \doteq x'}$. Possibly this is too obvious to mention, except that it plays a role in Thm 4.4.

In this section it is convenient to represent permutations in matrix form. Any permutation $(\pi_1, \dots, \pi_m)$ has an equivalent representation as a $(m \times m)$ matrix P for which

$$P_{ij} = \begin{cases} 1 & j = \pi_i \\ 0 & \text{otherwise.} \end{cases}$$

Note that P^T is another permutation matrix, and $P^T P = I$, the identity matrix. Then the definition of exchangeability is that $\mathrm{E}\{g(\boldsymbol{X})\} = \mathrm{E}\{g(P\boldsymbol{X})\}$ for all g and all P.

I use the conventional notation that if $A := (a_1, \dots, a_n)$ then

$$\boldsymbol{X}_A := (X_{a_1}, \dots, X_{a_n}).$$

In what follows, A is always a permutation of a subsequence of $(1, \dots, m)$, and B likewise. The sequence $\boldsymbol{X}_A$ can also be represented in matrix form, using the $(n \times m)$ selection matrix S_A, defined as

$$(S_A)_{ij} := \begin{cases} 1 & j = a_i \\ 0 & \text{otherwise,} \end{cases}$$

Here are two results about expectations of sequences of an exchangeable $\boldsymbol{X}$. The first states that if $\boldsymbol{X}$ is m-exchangeable then its subsets are n-exchangeable for all $1 < n < m$. In Sec. 4.4 we will see that the converse is *not* true.

Theorem 4.1. *If $\boldsymbol{X}$ is exchangeable then $\boldsymbol{X}_A$ is exchangeable.*

Proof. Denote the length of A as n and let g be an arbitrary n-ary function. We need to show that $\mathrm{E}\{g(S_A\boldsymbol{X})\} = \mathrm{E}\{g(P_n S_A \boldsymbol{X})\}$ where P_n is an arbitrary permutation matrix for $(1, \dots, n)$. Let

$$S := \begin{bmatrix} S_A \\ S_{\bar{A}} \end{bmatrix}$$

where $S_{\bar{A}}$ is a selection matrix for the complement of A, in any order. Note that S is a permutation matrix. Now fix P_n and consider the matrix equation

$$\begin{bmatrix} S_A \\ S_{\bar{A}} \end{bmatrix} P = \begin{bmatrix} P_n S_A \\ S_{\bar{A}} \end{bmatrix}. \tag{†}$$

Note that P is uniquely defined, because S is non-singular with inverse S^T; and P is a permutation matrix, because the righthand-side is a permutation matrix, S^T is a permutation matrix, and the permutation of a permutation is a permutation. Crucially, P satisfies $S_A P = P_n\, S_A$ from the top half of (†). Hence

$$\mathrm{E}\{g(S_A \boldsymbol{X})\} = \mathrm{E}\{g(S_A P \boldsymbol{X})\} = \mathrm{E}\{g(P_n S_A \boldsymbol{X})\}$$

as required, the first equality following because $\boldsymbol{X}$ is exchangeable and $g(S_A\,\cdot)$ is a function of $\boldsymbol{x}$. □

The next result states that any length-n subset of $\boldsymbol{X}$ has exactly the same expectations as any other length-n subset.

Theorem 4.2. *Let A and B be length n subsets of $(1, \dots, m)$. If $\boldsymbol{X}$ is exchangeable then*

$$\mathrm{E}\{g(\boldsymbol{X}_A)\} = \mathrm{E}\{g(\boldsymbol{X}_B)\}$$

for all n-ary functions g.

Proof. Follows the same reasoning as the proof of Thm 4.1, except now A and B are specified and P satisfies

$$\begin{bmatrix} S_A \\ S_{\bar{A}} \end{bmatrix} P = \begin{bmatrix} S_B \\ S_{\bar{B}} \end{bmatrix}$$

so that $S_A P = S_B$ and then

$$\mathrm{E}\{g(S_A \boldsymbol{X})\} = \mathrm{E}\{g(S_A P \boldsymbol{X})\} = \mathrm{E}\{g(S_B \boldsymbol{X})\}$$

as required. □

The next result extends exchangeability to hypothetical expectations.

Theorem 4.3. *Let $\boldsymbol{X} = [\boldsymbol{X}_A, \boldsymbol{X}_B]$ and let $q(\boldsymbol{x}_B)$ be a first-order sentence. If $\boldsymbol{X}$ is m-exchangeable and A is of length n, then*

$$\mathrm{E}\left\{g(\boldsymbol{X}_A) \mid q(\boldsymbol{X}_B)\right\} = \mathrm{E}\left\{g(P_n \boldsymbol{X}_A) \mid q(\boldsymbol{X}_B)\right\}$$

for all n-ary functions g, all P_n, and all q for which $\Pr\{q(\boldsymbol{X}_B)\} > 0$.

Proof. Follows the same reasoning as Thm 4.1, except now A, B, and P_n are specified, and P satisfies

$$\begin{bmatrix} S_A \\ S_B \end{bmatrix} P = \begin{bmatrix} P_n S_A \\ S_B \end{bmatrix}. \tag{†}$$

Then, because $\Pr\{q(\boldsymbol{X}_B)\} > 0$,

$$\begin{aligned}
&\mathrm{E}\left\{g(\boldsymbol{X}_A) \mid q(\boldsymbol{X}_B)\right\} \\
&\quad = \frac{\mathrm{E}\left\{g(S_A \boldsymbol{X})\, \mathbb{1}_{q(S_B \boldsymbol{X})}\right\}}{\mathrm{E}\left\{\mathbb{1}_{q(S_B \boldsymbol{X})}\right\}} && \text{by (1.3)} \\
&\quad = \frac{\mathrm{E}\left\{g(S_A P \boldsymbol{X})\, \mathbb{1}_{q(S_B P \boldsymbol{X})}\right\}}{\mathrm{E}\left\{\mathbb{1}_{q(S_B P \boldsymbol{X})}\right\}} && \text{by exchangeability} \\
&\quad = \frac{\mathrm{E}\left\{g(P_n S_A \boldsymbol{X})\, \mathbb{1}_{q(S_B \boldsymbol{X})}\right\}}{\mathrm{E}\left\{\mathbb{1}_{q(S_B \boldsymbol{X})}\right\}} && \text{by (†)} \\
&\quad = \mathrm{E}\left\{g(P_n \boldsymbol{X}_A) \mid q(\boldsymbol{X}_B)\right\} && \text{(1.3) again}
\end{aligned}$$

as required. □

There is a lot going on in this result, but we can simplify by noting that $g(P_n \,\cdot)$ is a way of defining a function of any length-ℓ subset of A if g is an n-ary function which only depends on its first ℓ arguments. So that one implication of the theorem is that if A_1 and A_2 are both length-ℓ subsets of A, $A \cap B = \emptyset$, and g is an ℓ-ary function, then

$$\mathrm{E}\left\{g(\boldsymbol{X}_{A_1}) \mid q(\boldsymbol{X}_B)\right\} = \mathrm{E}\left\{g(\boldsymbol{X}_{A_2}) \mid q(\boldsymbol{X}_B)\right\}, \tag{4.2}$$

for all q for which $\Pr\{q(\boldsymbol{X}_B)\} > 0$. Eq. (4.2) is just the hypothetical expectation analogue to Thm 4.3.

Another useful result for exchangeable sequences is given in Thm 6.4.

4.3 *Exchangeability and probability*

Exchangeability is conveniently defined in terms of the properties of my probability distribution for $\boldsymbol{X}$, according to the following result.

Theorem 4.4. $\boldsymbol{X}$ *is exchangeable if and only if*

$$f_{\boldsymbol{X}}(x_1, \dots, x_m) = f_{\boldsymbol{X}}(x_{\pi_1}, \dots, x_{\pi_m}) \tag{4.3}$$

where $(\pi_1, \dots, \pi_m)$ *is any permutation of* $(1, \dots, m)$*. In other words, if and only if* $f_{\boldsymbol{X}}$ *is a* symmetric function*.*

Proof. In what follows, fix P to be an arbitrary permutation matrix—equivalently, P^T is an arbitrary permutation matrix, because $P \longleftrightarrow P^T$ is a bijection.

($\Leftarrow$). This follows from the FTP (Thm 1.5):

$$\begin{aligned}
\mathrm{E}\{g(\boldsymbol{X})\} &= \textstyle\sum_{\boldsymbol{x}} g(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x}) && \text{by the FTP} \\
&= \textstyle\sum_{\boldsymbol{x}} g(\boldsymbol{x})\, f_{\boldsymbol{X}}(P\boldsymbol{x}) && \text{by symmetry of } f_{\boldsymbol{X}} \\
&= \textstyle\sum_{\boldsymbol{x}} g(P^T P\boldsymbol{x})\, f_{\boldsymbol{X}}(P\boldsymbol{x}) && \text{as } P^T P = I \\
&= \textstyle\sum_{\boldsymbol{y}} g(P^T \boldsymbol{y})\, f_{\boldsymbol{X}}(\boldsymbol{y}) && \text{letting } \boldsymbol{y} := P\boldsymbol{x}\text{, see below} \\
&= \mathrm{E}\{g(P^T \boldsymbol{X})\} && \text{FTP again}
\end{aligned}$$

as required, because P^T is an arbitrary permutation. For the 'see below' line, note that $\boldsymbol{x}$ ranges over the support of $\boldsymbol{X}$, and that this set has a product form, from (4.1). Hence if the support of $\boldsymbol{X}$ is

$$\boldsymbol{\mathcal{X}} := \left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(s)}\right\} \quad \boldsymbol{x}^{(j)} \in \mathcal{X}^m$$

then this is the same set as

$$\boldsymbol{\mathcal{Y}} := \left\{P\boldsymbol{x}^{(1)}, P\boldsymbol{x}^{(2)}, \dots, P\boldsymbol{x}^{(s)}\right\}$$

since permuting each element simply changes the order of the elements in the set.

($\Rightarrow$). This follows from choosing $g(x) \leftarrow \mathbb{1}_{x \doteq x'}$ for any $x' \in \mathcal{X}$, for then

$$\begin{aligned} f_X(x') &= \mathrm{E}\{\mathbb{1}_{X \doteq x'}\} && \text{by definition} \\ &= \mathrm{E}\{\mathbb{1}_{PX \doteq x'}\} && \text{by exchangeability} \\ &= \mathrm{E}\{\mathbb{1}_{X \doteq P^T x'}\} && \text{pre-multiply by } P^T \\ &= f_X(P^T x') \end{aligned}$$

as required. □

The characterisation in Thm 4.4 can be more convenient to work with than the original definition in Def. 4.1, because no reference is made to arbitrary functions g, but only to the symmetry of a single function f_X. This will be exploited in Sec. 4.6. Before that, however, we can clarify some points about f_X. These follow directly from the results in Sec. 4.2 and do not need to be reproved.

Theorem 4.5. *If X is exchangeable, then*

1. *f_{X_A} is a symmetric function;*
2. *$f_{X_A} = f_{X_B}$ if A and B are the same size; and*
3. *$f_{X_A \mid X_B}(\cdot \mid x_B)$ is a symmetric function if* $\Pr(X_B \doteq x_B) > 0$.

The symmetry of f_X is preserved under marginalisation and conditioning. This is another reason why exchangeability is conveniently characterised in terms of the symmetry of f_X, rather than equality of expectations of arbitrary functions of X. But I think that 'equality of expectations' is a better way to understand the 'similar but not identical' nature of exchangeability.

4.4* *Extendability*

Suppose $(X_1, \dots, X_n)$ is a sequence of random quantities that are n-exchangeable for me. Is this sequence necessarily the margin of a longer exchangeable sequence $(X_1, \dots, X_n, X_{n+1}, \dots, X_m)$? We might ask 'Are my beliefs about $(X_1, \dots, X_n)$ *m-extendable*?' The intriguing answer is No, although the converse is true according to Thm 4.1. This asymmetry introduces a subtle element to exchangeability modelling. It completely constrains my choice of functional form for f_X if I would like my beliefs to be extendable to arbitrarily long exchangeable sequences.

There is an elegant proof of non-extendability, found in Aldous (1985, ch. 1). It is based on the following inequality.[3]

Theorem 4.6. *Let $(X_1, \dots, X_n)$ be an exchangeable sequence. Then the correlation between X_i and X_j, denoted ρ, satisfies*

$$\rho \geq \frac{-1}{n-1}$$

with equality if and only if $\sum_i X_i$ is constant.

[3] See Sec. 1.3.1 for the definition and properties of the correlation.

Proof. The correlation is invariant to linear scalings, so let $Z_i := (X_i - \mu)/\sigma$, where μ and σ are the expectation and standard deviation of each X_i, so that

$$\rho := \operatorname{Corr}(X_i, X_j) = \operatorname{Corr}(Z_i, Z_j) = \operatorname{E}(Z_i\, Z_j).$$

Then

$$\begin{aligned} 0 &\leq \operatorname{E}\left\{ \left(\textstyle\sum_i Z_i\right)^2 \right\} \\ &= n \textstyle\sum_i \operatorname{E}\{(Z_i)^2\} + n(n-1)\operatorname{E}(Z_i\, Z_j) \qquad\qquad i \neq j \\ &= n + n(n-1)\rho \end{aligned}$$

and the displayed result follows after re-arranging. Finally, $\sum_i Z_i = 0$ if and only if $\sum_i X_i$ is constant. □

The lower bound in Thm 4.6 is increasing in n, which implies that an n-exchangeable $(X_1, \dots, X_n)$ with $\sum_i X_i$ constant cannot be $(n+1)$-extendable, because in this case

$$\operatorname{Corr}(X_i, X_j) = \frac{-1}{n-1} < \frac{-1}{(n+1)-1}$$

which is the lower bound for the correlation of two quantities from an $(n+1)$-exchangeable sequence. And the condition that $\sum_i X_i$ is constant is not vacuuous. For example, the distribution of a complete random draw without replacement from an urn with n quantities is exchangeable, and it has constant $\sum_i X_i$. It is also easy to construct examples of 2-exchangeable sequences that are not 3-extendable. For example, if U has a uniform distribution, then consider $(U, 1-U)$.

* * *

This result is not disturbing if I know m, the length of the sequence. But there are many situations where m is unknown to me: for example, I might be interested in some feature of the population of elephants in Etosha National Park in Namibia. Let M be the number of adult male elephants and X_i be some measurement that I make on such an elephant; then I might want $f_{X_1,\dots,X_M|M}(\cdot \mid m)$ to be exchangeable for all feasible m. Of course, I can definitely provide an upper bound on M, but I may prefer not to. A very strong result states that there is only one functional form for the probability distribution of $(X_1, \dots, X_m)$ which is m'-extendable for all finite m'.

Theorem 4.7 (De Finetti's representation theorem)**.** *An m-exchangeable $\boldsymbol{X} := (X_1, \dots, X_m)$ is m'-extendable for all finite m' if and only if its probability distribution has the form*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int \prod_{i=1}^{m} f_X(x_i; t)\, dF_\theta(t)$$

for some parameter θ with distribution function F_θ, and PMF $f_X(\cdot\,; \theta)$.

($\Leftarrow$) is immediate, but ($\Rightarrow$) is hard; see, e.g., Kingman (1978).

De Finetti's representation theorem is about arbitrary extendability of an exchangeable sequence. It does not insist that an m-exchangeable $\boldsymbol{X}$ be represented by the parametric model

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{m} f_X(x_i;\theta).$$

In applied statistics, though, this is a very common choice, often written as

$$X_1, \dots, X_m \overset{\text{iid}}{\sim} f_X(\cdot;\theta),$$

where '$\overset{\text{iid}}{\sim}$' denotes *independent and identically distributed (IID)*. Sec. 4.5 and Sec. 5.4 show how flexible this IID model is for representing beliefs.

If the origin of the IID model is exchangeable beliefs, then it must be augmented with a parameter distribution F_θ, in order that the $\boldsymbol{X}$ margin of $(\boldsymbol{X},\theta)$ is exchangeable (see Sec. 5.3). Usually θ is given an uncountably infinite realm, and F_θ is specified as a probability density function (PDF). There is no foundational reason for the realm of θ to be finite, because it is a random variable, not a random quantity, a distinction I made in Sec. 2.6. Having said that, though, I have preferred to treat θ's realm as finite for simplicity, to avoid highly technical issues concerned with the interpretation of $f_X(\cdot;\theta)$ as a conditional probability. Thus, in these notes π_θ is a PMF, not a PDF. But no errors will arise from replacing each '$\sum_t \cdot$' with '$\int \cdot \, \mathrm{d}t$'.

This derivation of the IID model as a representation of exchangeability will not appeal to Frequentists, because it implies the existence of a parameter distribution π_θ (see Sec. 2.4). Another derivation is available, based on the notion of a *random sample*. In a population of size m, a sample of size n is a random sample exactly when it is selected in such a way that every one of the $\binom{m}{n}$ possible samples has an equal probability of being chosen. This is sampling without replacement. Suppose a Frequentist believes that the proportion of $x^{(j)}$'s in the population is $f_X(x;\theta)$ for each $x \in \mathcal{X}$. If $n \ll m$, then sampling without replacement can be approximated by sampling with replacement, and then

$$f_{\boldsymbol{X}_{1:n}}(\boldsymbol{x}_{1:n};\theta) \approx \prod_{i=1}^{n} f_X(x_i;\theta).$$

Freedman (1977) gives a bound on the accuracy of this approximation. This random sampling model allows us to make inferences about the population based on the sample (because the sample can tell us about θ), and the substitution of with-replacement for without-replacement ensures that the sample distribution $f_{\boldsymbol{X}_{1:n}}$ has the simple and tractable IID form.

Beyond the framework of random sampling for $n \ll m$, however, no simple justification of the IID model presents itself, and yet the IID model is ubiquitous. Thus exchangeable beliefs seems to be the approporiate justification for most of the IID models in statistical applications. This strongly supports the Bayesian position of providing a prior distribution for the parameters in the case of IID components in the statistical model (see Sec. 2.6).

4.5 *Partial exchangeability and linear modelling*

Exchangeability of a population $\boldsymbol{X} := (X_1, \dots, X_m)$ is a very strong belief, asserting that I choose to ignore any information I have that X_i and X_j belong to the same group, or that X_i and X_k belong to different groups. And yet it is common for this information to be available, and for me to want to incorporate it into my beliefs.

Suppose that each member of the population can be assign to a group. The variable that identifies the group is termed a *factor*, and the different groups are identified by different *levels* of the factor. For humans, for example, 'Biological sex' is a factor, with levels 'male' and 'female'. A factor and its levels can be compound; for example, the factor could be 'Sex × Age', and the levels could be selected from

$$\{\text{male}, \text{female}\} \times \{\text{<5}, \text{5–10}, \text{10–20}, \dots, \text{>90}\}.$$

Here I am following a minor convention in statistical programming that the factor is capitalised and the levels are not.

The natural belief specification in this situation is that two members of the population with the same level of the factor are more like each other than two members of the population with different levels. So I do not want to treat the population as exchangeable, because I may want to assign different beliefs to $h(X_i, X_j)$ and $h(X_i, X_k)$, in the case where X_i and X_j have the same level and X_i and X_k have different levels; e.g. for $h(x, x') \leftarrow (x - x')^2$.

I introduce a slightly more general notation, so that

$$\boldsymbol{X} := \left(X_{11}, \dots, X_{1m_1}, X_{21}, \dots, X_{2m_2}, \dots, X_{g1}, \dots, X_{gm_g}\right)$$

where there are g groups altogether, group i has m_i members, and X_{ij} is the jth member of group i. I assume that the group structure exhausts my ability to discriminate between X's, so that $(X_{i1}, \dots, X_{im_i})$ are exchangeable for me for each i. This is an example of *partial exchangeability*. The simple and flexible way to represent these beliefs is to decompose each X_{ij} as

$$X_{ij} \sim \alpha_i + \varepsilon_{ij} \qquad \begin{cases} i = 1, \dots, g \\ j = 1, \dots, m_i, \end{cases}$$

where $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_g)$ is a parameter and the ε's are random variables, for which

$$\boldsymbol{\varepsilon}_i := (\varepsilon_{i1}, \dots, \varepsilon_{im_i})$$

is exchangeable for each i, and $(\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_g)$ are mutually independent (to be defined in Sec. 5.2).

The simplest exchangeable model for $\boldsymbol{\varepsilon}_i$ is the IID model

$$f_{\varepsilon}(\boldsymbol{e}_i; \sigma_i) = \prod_{j=1}^{m_i} f_{\varepsilon}(e_{ij}; \sigma_i)$$

for some statistical model $f_{\varepsilon}(\cdot; \sigma_i)$ with parameter σ_i. The attraction of this model is its simplicity of specification and its adaptability across groups. In the absence of strong beliefs, I may choose to use the same σ_i for all i, or for i's from similar groups. For example, in a compound factor I may choose to have σ_i depend on the level of Sex but not Age. Whenever I assign the same σ_i across two or more different levels I am representing additional exchangeability beliefs. For example, if $\sigma_i = \sigma_j$ then

$$\big(X_{i1} - \alpha_i, \dots, X_{im_i} - \alpha_i, X_{j1} - \alpha_j, \dots, X_{jm_j} - \alpha_j\big)$$

is exchangeable.

My beliefs about the differences between groups are incorporated into my probability distribution for $\boldsymbol{\alpha}$, as well as in my σ_i's. The strongest belief I can have is that all of the α_i's are the same, or

$$\alpha_1 = \cdots = \alpha_g = \beta$$

for some scalar parameter β. A slightly less strong version with more parameters is to let α_i depend on quantifiable features of the group. Let $\boldsymbol{v}_i := (v_{i1}, \dots, v_{ip})$ denote a vector of values I can attach to level i of the factor, such as 0 or 1 for Sex or the centre of an interval for Age. Then I might choose to represent $\boldsymbol{\alpha}$ as

$$\alpha_i = \beta_0 + \sum_{k=1}^{p} \beta_k v_{ik} \qquad i = 1, \dots, g \tag{†}$$

where $\boldsymbol{\beta} := (\beta_0, \beta_1, \dots, \beta_p)$ is a vector parameter. There is no end to the different choices I might make for the functional relationship between $\boldsymbol{v}_i$, $\boldsymbol{\beta}$, and α_i, but the linear relationship is the default choice because of its tractability. Readers will recognise that this choice for $\boldsymbol{\alpha}$ and a common value for the σ_i's is the classical *linear regression model*: undoubtedly the most popular statistical model of all time.

What about weaker beliefs? Well, I could have $\alpha_i = \beta_i$ as a special case of the above, giving me a parameter for each group. In this case α_i tells me nothing at all about α_j. But I might feel that this is too strong, and that 'similar but not identical' might better summarise my beliefs about the $\boldsymbol{\alpha}$. In other words, I might want $\boldsymbol{\alpha}$ to be exchangeable. This could be handled by treating $\boldsymbol{\alpha}$ as a random variable. The simplest exchangeable model for $\boldsymbol{\alpha}$ would be the IID model

$$f_{\alpha}(\boldsymbol{a}; \beta) = \prod_{i=1}^{g} f_{\alpha}(a_i; \beta) \tag{‡}$$

for some statistical model $f_{\alpha}(\cdot; \beta)$ with parameter β. With slightly weaker beliefs I could go further, and arrange for $\boldsymbol{\alpha}$ to have a partially exchangeable specification, notably for compound groups. For example, if group i corresponded to level j of the first factor and k of the second, then I might have

$$\alpha_i = \gamma_j + \delta_k + \xi_i$$

where the ξ's might be IID, and the γ's and δ's might be parameters, or they might themselves be exchangeable. At this point, we have a *multilevel model*.

The gradient from 'strong' to 'weak' beliefs is represented in increasing numbers of parameters, on the basis that stronger beliefs arise from constraining the parameters of weaker beliefs. If this strikes you as unintuitive, then possibly you have confused 'simple' with 'weak'. The simple statistical models used in many applied areas, such as economics, medical science, or experimental psychology, encode extremely strong beliefs.

Partial exchangeability for $\boldsymbol{X}$ illustrates two very important features of the general process of statistical modelling. First, sometimes it is better to treat parameters as random variables, in order to incorporate beliefs about $\boldsymbol{X}$ through exchangeability in the parameters. This blurs the distinction between Frequentist and Bayesian approaches, and does not allow the Frequentist to maintain a hard-line position that it is unacceptable to attach probability distributions to parameters (Sec. 4.4 already challenged that position for IID models). Frequentists have introduced the notion of a *random effect* to finesse this issue, but a random effect passes the Duck Test for being an exchangeable random variable.[4]

Second, for practical reasons (see Sec. 2.7) the statistical modeller is usually trying to reduce the number of parameters, relative to the number of observations. It is easy to construct a statistical model with at least as many parameters as observations (e.g. where each element of $\boldsymbol{X}$ has its own group), but in such a model the observations will fail to constrain the parameters, and the tenability of the inference will be extremely low (see Sec. 2.8). For maximum tenability, a parameter should be present in the marginal PMF of many observations, but differently for each parameter. This is only possible if the number of parameters is much less than the number of observations. So the statistician is always aiming for fewer distinct groups, and simpler specifications for $\boldsymbol{\alpha}$. The practical attraction of an exchangeable model for $\boldsymbol{\alpha}$ such as (‡) is that with f_α specified only a single parameter β is required, to model all g groups. Likewise, linear models such as (†) can allow every observation to have its own group, but with just a few parameters.

4.6* Probabilities and histograms

Let $\boldsymbol{X} := (X_1, \dots, X_m)$ be exchangable, and let the realm of each X be finite, $X_i \in \mathcal{X} := \{x^{(1)}, \dots, x^{(s)}\}$. For future reference, note that this may be the original realm of each X, or it may be a coarsened realm in which similar elements have been grouped together.[5]

[4] "If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck."

[5] It may well be that $\mathcal{X}$ is a notionally infinite realm coursened into a finite one.

$\boldsymbol{X}$ with a finite realm is exchangeable if and only if the PMF $f_{\boldsymbol{X}}$ depends only on the histogram of $\boldsymbol{X}$,

$$\boldsymbol{h}(\boldsymbol{x}) := \big(h_1(\boldsymbol{x}), \dots, h_s(\boldsymbol{x})\big)$$

where $h_j(\boldsymbol{x})$ is the number of elements of $\boldsymbol{x}$ which are equal to $x^{(j)}$. There are

$$\mathcal{M}_{\boldsymbol{h}(\boldsymbol{x})} := \frac{m!}{h_1(\boldsymbol{x})! \cdots h_s(\boldsymbol{x})!}$$

different ways of arranging the elements of the histogram $\boldsymbol{h}(\boldsymbol{x})$ into a sequence, termed the *multinomial coefficient*. Hence every exchangeable distribution for $\boldsymbol{X}$ can be written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{x})}}{\mathcal{M}_{\boldsymbol{h}(\boldsymbol{x})}}$$

where $\mathcal{P}_{\boldsymbol{h}(\boldsymbol{x})}$ is the probability assigned to histogram $\boldsymbol{h}(\boldsymbol{x})$. There is a $\mathcal{P}_{\boldsymbol{h}}$ for each possible histogram, and there are

$$r := \binom{m+s-1}{s-1},$$

histograms altogether; there is a cute proof of this result in Sec. 4.6.1. Hence the set of $\mathcal{P}_{\boldsymbol{h}}$'s for a given $f_{\boldsymbol{X}}$ is a point in the unit simplex $\mathbb{S}^{r-1}$.

For example, suppose that I am going to spin the same coin three times, with $X_i = 0$ if the ith spin is a tail, and $X_i = 1$ if it is a head. So $\mathcal{X} = \{0, 1\}$, $s = 2$, and $m = 3$. I believe that the spins are exchangeable. This means that I would attach the same probability to each element within each of the following four sets of outcomes:

$$\begin{aligned}
&\{(0,0,0)\} && \text{or } \boldsymbol{h}(\boldsymbol{x}) = (3,0) \\
&\{(1,0,0), (0,1,0), (0,0,1)\} && \boldsymbol{h}(\boldsymbol{x}) = (2,1) \\
&\{(0,1,1), (1,0,1), (1,1,0)\} && \boldsymbol{h}(\boldsymbol{x}) = (1,2) \\
&\{(1,1,1)\} && \boldsymbol{h}(\boldsymbol{x}) = (0,3)
\end{aligned}$$

The number of triples in each set is the multinomial coefficient of the histogram. There are $r = 4$ possible histograms in total, which satisifes the formula,

$$\binom{3+2-1}{2-1} = 4.$$

Hence my PMF for $\boldsymbol{X}$ is specified by a point in $\mathbb{S}^3$, such as

$$\mathcal{P} = (0.2, 0.5, 0.1, 0.2)$$

in the same order as above. If I want to know the value of $f_{\boldsymbol{X}}$ at $\boldsymbol{x} = (1, 1, 0)$ I compute

$$f_{\boldsymbol{X}}(1,1,0) = \frac{\mathcal{P}_{(1,2)}}{\mathcal{M}_{(1,2)}} = \frac{0.1}{3} = 0.033.$$

So when X has a finite realm there is a bijective relationship between the exchangeable distributions for $\boldsymbol{X}$ and the space $\mathbb{S}^{r-1}$. The cardinality of the unit simplex is the same as that of the continuum. Thus there are an uncountably infinite number of exchangeable distributions for $\boldsymbol{X}$, notwithstanding that exchangeability is a very strong condition.

Now suppose that we have observed the first n X's in the sequence,

$$X_1 \to y_1, \dots, X_n \to y_n,$$

and I would like to compute my conditional probabilities for X_{n+1} (or any other unobserved X), i.e. to find

$$\Pr(X_{n+1} \doteq x^{(j)} \mid \boldsymbol{X}_{1:n} \doteq \boldsymbol{y}) \quad j = 1, \dots, s.$$

We can derive this value indirectly, following the approach outlined by David Blackwell, in his discussion of Diaconis (1988). Define

$$\beta_j := \frac{\Pr(X_{n+1} \doteq x^{(j)} \mid \boldsymbol{X}_{1:n} \doteq \boldsymbol{y})}{\Pr(X_{n+1} \doteq x^{(1)} \mid \boldsymbol{X}_{1:n} \doteq \boldsymbol{y})} \quad j = 1, \dots, s$$

where of course $\beta_1 = 1$. But note that

$$\sum_{j=1}^{s} \beta_j = \frac{1}{\Pr(X_{n+1} \doteq x^{(1)} \mid \boldsymbol{X}_{1:n} \doteq \boldsymbol{y})}.$$

Hence

$$\Pr(X_{n+1} \doteq x^{(j)} \mid \boldsymbol{X}_{1:n} \doteq \boldsymbol{y}) = \frac{\beta_j}{\sum_{j'} \beta_{j'}} \quad j = 1, \dots, s. \tag{4.4}$$

So the probabilities we seek are just the normalised values of the β's. It turns out that the β's have a fairly simple representation.

Theorem 4.8. *Define $\boldsymbol{e}_j$ as the unit vector with a 1 in the jth position. Then, using the definitions above,*

$$\beta_j = \frac{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j}}{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_1}} \frac{h_j(\boldsymbol{y}) + 1}{h_1(\boldsymbol{y}) + 1} \quad j = 1, \dots, s.$$

Proof. $\boldsymbol{X}$ is exchangeable, and hence $\boldsymbol{X}_{1:(n+1)}$ is exchangeable by Thm 4.1. The probabilities $\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j}$ correspond to histograms defined on the first $(n+1)$ terms in the sequence. They can be found from f_X by marginalising out $\boldsymbol{X}_{(n+2):m}$. Then

$$\begin{aligned}
\beta_j &= \frac{\Pr(X_{n+1} \doteq x^{(j)} \mid \boldsymbol{X}_{1:n} \doteq \boldsymbol{y})}{\Pr(X_{n+1} \doteq x^{(1)} \mid \boldsymbol{X}_{1:n} \doteq \boldsymbol{y})} \\
&= \frac{\Pr(\boldsymbol{X}_{1:n} \doteq \boldsymbol{y}, X_{n+1} \doteq x^{(j)})}{\Pr(\boldsymbol{X}_{1:n} \doteq \boldsymbol{y}, X_{n+1} \doteq x^{(1)})} \\
&= \frac{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j} / \mathcal{M}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j}}{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_1} / \mathcal{M}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_1}} \\
&= \frac{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j}}{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_1}} \frac{\mathcal{M}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_1}}{\mathcal{M}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j}}.
\end{aligned}$$

And then

$$\frac{\mathcal{M}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_1}}{\mathcal{M}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j}} = \frac{\frac{(n+1)!}{(h_1(\boldsymbol{y})+1)!\cdots h_s(\boldsymbol{y})!}}{\frac{(n+1)!}{h_1(\boldsymbol{y})!\cdots(h_j(\boldsymbol{y})+1)!\cdots h_s(\boldsymbol{y})!}} = \frac{h_j(\boldsymbol{y})+1}{h_1(\boldsymbol{y})+1},$$

as required. □

This is an exciting result, because (4.4) and Thm 4.8 have exactly the form that was analysed in the Stable estimation theorem (Sec. 2.7 and Thm 2.4). Each β_j is the product of two terms, which we can represent as

$$u_j = \frac{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j}}{\mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_1}} \quad \text{and} \quad v_j = \frac{h_j(\boldsymbol{y})+1}{h_1(\boldsymbol{y})+1}.$$

So we are able to state and assess the *stability conditions*, under which my probabilities will be well-approximated by the normalised v's—which we should think of here as the likelihoods. The normalised v's have the form

$$\frac{v_j}{\sum_{j'} v_{j'}} = \frac{h_j(\boldsymbol{y})+1}{n+s} \qquad j = 1, \dots, s.$$

So if the stability conditions were appropriate for me, I would have a simple rule for turning frequencies into probabilities. Note that this rule *does not* simply normalise the frequencies, although this would be a good approximation if $h_j(\boldsymbol{y}) \gg 1$ for each j: I'll return to this point below.

For the stability conditions we need to identify a subset $B \subset \{1, \dots, s\}$ for which $\alpha \ll 1$, and for which my $\mathcal{P}$ implies a small β and γ. But actually this is rather simple, because all of the relevant histograms are in the add-one-ball neighbourhood of the observed histogram $\boldsymbol{h}(\boldsymbol{y})$. If my $\mathcal{P}$ is smooth in the neighbourhood of $\boldsymbol{h}(\boldsymbol{y})$ then β will be small even if I take $B = \{1, \dots, s\}$, and this choice for B implies that $\alpha = 0$ and $\gamma = 0$. So the stability conditions can be reduced to the single sufficient condition:

1. My histogram probabilities $\mathcal{P}$ are smooth in the neighbourhood of the observed histogram $\boldsymbol{h}(\boldsymbol{y})$.

The interesting point about this sufficient condition is that it does not always hold. Take the extreme case, when n is zero. In this case,

$$u_j \propto \mathcal{P}_{\boldsymbol{h}(\boldsymbol{y})+\boldsymbol{e}_j} = \mathcal{P}_{\boldsymbol{e}_j} = f_{X_1}(x^{(j)}).$$

If I believe that some values of $\mathcal{X}$ are far more probable than others, then there will be large differences between the u_j's, and stability condition 2 will be violated if I select $B = \{1, \dots, s\}$. Or, to belabour the point, $1/s$ would be a poor approximation to $f_{X_1}(x^{(j)})$. But if I try a smaller B, I will run into trouble with a large α or a large γ.

On the other hand, when $h_j(\boldsymbol{y}) \gg 1$ for each j it seems very natural that my $\mathcal{P}$ would be smooth in the neighbourhood of $\boldsymbol{h}(\boldsymbol{y})$.

Consider the case where $s = 4$ and $n = 40$. Now in some applications my beliefs might distinguish between the two histograms

$$(28, 10, 1, 1) \quad \text{and} \quad (29, 10, 1, 0),$$

if $x^{(1)}$ was a very different outcome to $x^{(4)}$. But I doubt they would ever distinguish meaningfully between the two histograms

$$(11, 12, 9, 8) \quad \text{and} \quad (10, 12, 9, 9).$$

So in many applications it seems appropriate to replace the above condition with

1′. $h_j(\boldsymbol{y}) \gg 1$ for $j = 1, \dots, s$.

This condition also implies that

$$\frac{v_j}{\sum_{j'} v_{j'}} \approx \frac{h_j(\boldsymbol{y})}{n} \quad j = 1, \dots, s$$

although there is never any need to make this approximation. But at least we now have a clear description of when it is that relative frequencies can approximate probabilities, expressed as the following result.

Theorem 4.9 (Histogram theorem). *Relative frequencies approximate probabilities when the underlying sequence is exchangeable, and the number of elements in each bin is much larger than one.*

The fly in the ointment of the Histogram theorem is a large s, because the sample size n needs to be a multiple of s if condition 1′ is to be satisfied. Unfortunately, n might not be under the control of the statistician or the client; perhaps the sample has already been collected, or perhaps each element is very expensive to collect. But s is always adjustable, because we can coarsen $\mathcal{X}$ to reduce s, if necessary. This may not suit the client, who would prefer her prediction for X_{n+1} (or any other unobserved X) to have high resolution. So she must make up her mind, on the basis of her sample $\boldsymbol{y}$. If this sample does not satisfy condition 1′, then either she provides prior probabilities, in terms of values for the $\mathfrak{P}$'s in the neighbourhood of $\boldsymbol{h}(\boldsymbol{y})$, or else she coarsens $\mathcal{X}$, so that condition 1′ is satisfied.

4.6.1 Stars and bars*

There is a very cute proof of $r = \binom{m+s-1}{s-1}$, which originated with William Feller (see, e.g., Feller, 1968). To divide m balls between s buckets with at least one ball per bucket, represent each ball as a star and arrange them in a line:

$$\star \star \star \star \star \star \star \star \star \star$$

say, for $m = 10$. These m stars define $m - 1$ gaps, of which we choose $s - 1$ gaps to define a particular arrangement, and indicate them with bars:

$$\star \star \mid \star \mid \star \star \star \mid \star \star \mid \star \star$$

say, with $s = 5$. So there must be $\binom{m-1}{s-1}$ possible arrangements of m balls in s buckets, where each bucket has at least one ball.

Now to allow for some buckets to have no balls, do the same thing with $m + s$ balls, and then take one ball out of each bucket to leave m balls in total. Then the total number of possible arrangements is

$$r = \binom{m+s-1}{s-1} = \binom{m+s-1}{m}$$

as required.

5 Conditional Independence

This is the second of two chapters on statistical modelling. This chapter defines the notion of conditional independence, and explains how it is used to induce rich dependencies across a set of random quantities. Sec. 5.1 defines the notion of conditional independence, and Sec. 5.2 derives some familiar representations in terms of probabilities. Sec. 5.3 outlines the general practice of using conditional independence to construct statistical models. This marks the point at which the Frequentist and Bayesian approaches to statistical modelling properly separate (from which the synthesis in Sec. 2.6 is seen to be an unhappy compromise). Sec. 5.4 and Sec. 5.5 present the two dominant statistical modelling strategies based on conditional independence.

5.1 Conditional independence

Informally, two sets of random quantities are probabilistically independent for me, if knowledge of one set has no implications for my beliefs about the other set. Such independence is very strong—too strong to be useful, because I can only improve my inferences about X using the observations Y if I believe that there is some dependence between them. On the other hand, a situation where every random quantity directly affects my beliefs about every other random quantity seems too complicated to be elicited, for real-world analyses. Somewhere in the middle we have the very useful notion of *conditional independence*.[1]

Because independence is a special case of conditional independence, I will just explore conditional independence in this section. In the material below and in Sec. 5.2, independence results can be recovered simply by dropping the conditioning on $\boldsymbol{Z}$ (this follows from Thm 1.19). Here is a formal definition of conditional independence.[2]

Definition 5.1 (Conditional independence). *Let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ be collections of random quantities. Then $\boldsymbol{X}$ is conditionally independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ exactly when, for all g, there is a $\psi_g \in \mathcal{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Y}, \boldsymbol{Z}\}$ which is invariant to $\boldsymbol{y}$. This is denoted $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$.*

An immediate consequence of this definition is that if $\boldsymbol{U} = h(\boldsymbol{X})$

[1] Now may be a good time to review Sec. 1.7, on conditional expectation.

[2] As usual, I treat all realms as finite. If you are bold enough to use non-finite and possibility non-bounded realms, then g, and similar functions below, should be bounded.

then $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \implies \boldsymbol{U} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$.

Recollect from Thm 1.13 that if ψ_g and ψ'_g are any two elements of $\mathcal{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Y}, \boldsymbol{Z}\}$, then $\psi_g(\boldsymbol{Y}, \boldsymbol{Z}) \overset{\text{ms}}{=} \psi'_g(\boldsymbol{Y}, \boldsymbol{Z})$. So if there is an element of $\mathcal{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Y}, \boldsymbol{Z}\}$ which is invariant to $\boldsymbol{y}$, then all versions of my conditional expectation are mean-square invariant to $\boldsymbol{Y}$. A conditional expectation solves an optimisation problem (Sec. 1.7.2), and hence $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ asserts that my conditional expectation of any function of $\boldsymbol{X}$ given both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is no better than that based on $\boldsymbol{Z}$ alone. That is not to say that $\boldsymbol{Y}$ is uninformative about $\boldsymbol{X}$, but simply that it does not bring me any information about $\boldsymbol{X}$ which is not already present in $\boldsymbol{Z}$.

Causal chains provide a very intuitive illustration of conditional independence. My beliefs about the power generated at a hydroelectric plant, X, are strongly influenced by the depth of the reservoir, Z. So much so that, given Z, knowledge of the previous rainfall on the reservoir catchment, Y, has no further impact on my beliefs about X. Hence, for me, $X \perp\!\!\!\perp Y \mid Z$. This illustration also shows that $X \perp\!\!\!\perp Y \mid Z \not\Longrightarrow X \perp\!\!\!\perp Y$. For if I did not know the depth of the water, then the previous rainfall would be highly informative about power generated.

We can also clarify that $X \perp\!\!\!\perp Y \not\Longrightarrow X \perp\!\!\!\perp Y \mid Z$. Suppose that X and Y are the points from two rolls of a die believed by me to be fair. In this case, I might reasonably believe that $X \perp\!\!\!\perp Y$, if I had shaken the die extensively inside a cup before each roll. Note that it is important that I know the probabilities (the die being believed fair, they are $1/6$ for each score), because otherwise I would want to treat (X, Y) as exchangeable, not independent. But if Z is the sum of the points in the two rolls, then I can predict X exactly knowing Y and Z, but only approximately using Z alone. So Y brings information about X that augments the information in Z, and I do not believe that $X \perp\!\!\!\perp Y \mid Z$.

These two illustrations show that conditional independence is its own thing, not simply a necessary or sufficient condition for independence. My belief that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ is something I accept or reject after reflecting on how my beliefs about $\boldsymbol{X}$ in the presence of $\boldsymbol{Z}$ change on the further presence of $\boldsymbol{Y}$. The asymmetry of $\boldsymbol{X}$ and $\boldsymbol{Y}$ turns out to be an illusion—a fascinating and deep result, which I will demonstrate in Sec. 5.2 (eq. 5.1). The relationship between conditional independence (symmetric) and causality (asymmetric) is very subtle; see Pearl (2000) and Dawid (2002, 2010) for discussions.

5.2 Conditional independence and probabilities

In Sec. 5.1, I presented conditional independence in its most natural guise, which is as a statement about my conditional expectations. Practically, however, conditional independence gets implemented by statisticians in terms of conditional PMFs. This section presents some equivalences for this purpose, which form the basis of the statistical modelling presented in the following sections.

Now would be a good time to revisit Sec. 1.8.3 on probability mass functions (PMFs), functional equalities, and the support of a set of random quantities, denoted 'supp'. Recall that

$$\operatorname{supp}(\boldsymbol{Y}, \boldsymbol{Z}) \subset \operatorname{supp}(\boldsymbol{Y}) \times \operatorname{supp}(\boldsymbol{Z}).$$

The first two results below require a stronger *positivity condition* in which the subset is replaced by an equality.

Theorem 5.1. *If* $\operatorname{supp}(\boldsymbol{Y}, \boldsymbol{Z}) = \operatorname{supp}(\boldsymbol{Y}) \times \operatorname{supp}(\boldsymbol{Z})$, *then the following two statements are equivalent:*

A. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$

B. $f(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{x} \mid \boldsymbol{z}), \quad (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X} \times \operatorname{supp}(\boldsymbol{Y}) \times \operatorname{supp}(\boldsymbol{Z}).$

Proof. I will treat $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ as scalar, simply to reduce the amount of ink on the page (and similarly for the proofs that follow).

(A ⇒ B). From the definition of conditional independence, there is an element of $\mathcal{E}\{g(X) \mid Y, Z\}$ which is invariant to y and hence also an element of $\mathcal{E}\{g(X) \mid Z\}$, from Thm 1.19. Call the first element ψ_g and the second ϕ_g, so we have $\psi_g \in \mathcal{E}\{g(X) \mid Y, Z\}$ and

$$\psi_g(y, z) = \phi_g(z)$$

for some $\phi_g \in \mathcal{E}\{g(X) \mid Z\}$.

Now consider $(y, z) \in \operatorname{supp}(Y) \times \operatorname{supp}(Z)$. According to Thm 1.18, $\phi_g(z) = \mathrm{E}\{g(X) \mid Z \doteq z\}$. But since $f(y, z) > 0$ by the positivity condition, Thm 1.18 also implies that $\psi_g(y, z) = \mathrm{E}\{g(X) \mid Y \doteq y, Z \doteq z\}$. Setting $g(X) \leftarrow \mathbb{1}_{X \doteq x}$ then completes this branch of the proof.

(B ⇐ A). We need to show that there is a $\psi_g(y, z) \in \mathcal{E}\{g(X) \mid Y, Z\}$ which is invariant to y. So we must consider the expression

$$\mathrm{E}\left[\{g(X) - \psi_g(Y, Z)\}k(Y, Z)\right]$$

which must be zero for all k, in order for $\psi_g \in \mathcal{E}\{g(X) \mid Y, Z\}$. Assuming (B), I will construct a ψ_g which is invariant to y:

$$\begin{aligned}
&\mathrm{E}\left[\{g(X) - \psi_g(Y, Z)\}k(Y, Z)\right] \\
&\quad = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \{g(x) - \psi_g(y, z)\}k(y, z) \cdot f(x, y, z) && \text{by the FTP} \\
&\quad = \sum_{x \in \mathcal{X}} \sum_{y \in \operatorname{supp}(Y)} \sum_{z \in \operatorname{supp}(Z)} \{g(x) - \psi_g(y, z)\}k(y, z) \cdot f(x, y, z) && \text{as the missing terms are each zero} \\
&\quad = \sum_{x \in \mathcal{X}} \sum_{y \in \operatorname{supp}(Y)} \sum_{z \in \operatorname{supp}(Z)} \{g(x) - \psi_g(y, z)\}k(y, z) \cdot f(x \mid y, z)\, f(y, z) && \text{by the positivity condition} \\
&\quad = \sum_{x \in \mathcal{X}} \sum_{y \in \operatorname{supp}(Y)} \sum_{z \in \operatorname{supp}(Z)} \{g(x) - \psi_g(y, z)\}k(y, z) \cdot f(x \mid z)\, f(y, z) && \text{by hypothesis (B)} \\
&\quad = \sum_{y \in \operatorname{supp}(Y)} \sum_{z \in \operatorname{supp}(Z)} \Big\{ \sum_{x \in \mathcal{X}} g(x)\, f(x \mid z) - \psi_g(y, z) \Big\} k(y, z) \cdot f(y, z) && \text{as } \textstyle\sum_x f(x \mid z) = 1.
\end{aligned}$$

Hence the choice

$$\psi_g(y,z) \leftarrow \begin{cases} \sum_{x \in \mathcal{X}} g(x)\, f(x \mid z) & z \in \operatorname{supp}(Z) \\ 0 & \text{otherwise} \end{cases}$$

ensures that this expression is zero for all k. But this choice is invariant to y which is the result we seek. □

This next result is a stepping-stone to Thm 5.3.

Theorem 5.2. *Under the same positivity condition as Thm 5.1, the following two statements are equivalent:*

A. $f(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = f(\boldsymbol{x} \mid \boldsymbol{z})$

B. $f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = f(\boldsymbol{x} \mid \boldsymbol{z})\, f(\boldsymbol{y} \mid \boldsymbol{z})$

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

Proof. Follows immediately from Thm 1.24, which asserts

$$f(x,y \mid z) = f(x \mid y,z)\, f(y \mid z),$$

noting that $f(y,z) > 0$ implies that $f(y \mid z) > 0$. □

This next—very important—result is the one taken to define conditional independence in terms of probabilities. It has a deep corollary (eq. 5.1).

Theorem 5.3. *The following two statements are equivalent:*

A. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$

B. $f(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = f(\boldsymbol{x} \mid \boldsymbol{z})\, f(\boldsymbol{y} \mid \boldsymbol{z}), \quad (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{Y}} \times \operatorname{supp}(\boldsymbol{Z})$.

Proof. We have already established this result under the positivity condition $f(y,z) > 0$, concatenating Thm 5.2 and Thm 5.1, so we only need to consider the case where $f(z) > 0$ and $f(y,z) = 0$. $f(z) > 0$ implies that $f(x \mid z)$ and $f(x,y \mid z)$ are well-defined, and $f(y,z) = 0$ implies that

$$f(y \mid z) = 0, \text{ and } f(x,y \mid z) = 0 \text{ for all } x.$$

Thus B is always true, and $A \Rightarrow B$ is proved.

Now we show that under the same conditions, A is also always true, to prove $B \Rightarrow A$. Let ψ_g be an element of $\mathcal{E}\{g(X) \mid Y, Z\}$, i.e. any function satisfying

$$\mathrm{E}\left[\{g(X) - \psi_g(Y,Z)\}\, k(Y,Z)\right] = 0 \qquad \text{for all } k.$$

Applying the FTP (Thm 1.5), this condition can also be written

$$\sum_{x,y,z} \{g(x) - \psi_g(y,z)\}\, k(y,z) \cdot f(x,y,z) = 0 \qquad \text{for all } k.$$

Because $f(z) > 0$ we can write $f(x,y,z) = f(x,y \mid z)\, f(z)$, but with $f(x,y \mid z) = 0$ for all x, we have

$$\sum_{x,y,z} \big\{g(x) - \psi_g(y,z)\big\}\, k(y,z) \cdot 0 = 0 \qquad \text{for all } k.$$

This is satisfied by all well-defined ψ_g, including $\psi_g(y,z) = \phi_g(z)$, a function invariant to y. □

As an immediate corollary we have the symmetry result

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \iff \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z}, \tag{5.1}$$

which follows from the symmetry of the righthand side of B in Thm 5.3. See the references given at the end of Sec. 5.1 for more details.

Following Thm 5.3, there is a simple way to read conditional independence relationships from the factorisation of a conditional PMF.

Theorem 5.4. *Under the same positivity condition as Thm 5.3,*

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \iff f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{z}) = a(\boldsymbol{x},\boldsymbol{z})\, b(\boldsymbol{y},\boldsymbol{z})$$

for some non-negative functions a and b.

A similar result holds for the 'unconditional' $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$, because $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = f(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{z})\, f(\boldsymbol{z})$, and the $f(\boldsymbol{z})$ can be incorporated into either a or b.

Proof.
($\Rightarrow$). This follows immediately from Thm 5.3, with $a(x,z) \leftarrow f(x \mid z)$ and $b(y,z) \leftarrow f(y \mid z)$.

($\Leftarrow$). Let $a(z) := \sum_x a(x,z)$ and $b(z) := \sum_y b(y,z)$. Summing over y, x, and y and x, $f(x \mid z) = a(x,z)\, b(z)$, $f(y \mid z) = a(z)\, b(y,z)$, and $1 = a(z)\, b(z)$. Then

$$f(x,y \mid z) = a(x,z)\, b(y,z) = \frac{f(x \mid z)}{b(z)}\, \frac{f(y \mid z)}{a(z)} = f(x \mid z)\, f(y \mid z)$$

showing that $X \perp\!\!\!\perp Y \mid Z$, by Thm 5.3. □

The following special case of conditional independence is frequently used, see Sec. 5.4.

Definition 5.2 (Mutual conditional independence, MCI)**.**
Let $\boldsymbol{X} := (X_1, \dots, X_m)$, and let A and B be disjoint subsets of $(1, \dots, m)$. Then $\boldsymbol{X}$ is mutually independent given $\boldsymbol{Z}$ *exactly when $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{Z}$ for all A and B. I denote this as $\models \boldsymbol{X} \mid \boldsymbol{Z}$.*[3]

MCI has a very distinctive representation in terms of products.

Theorem 5.5.

$$\models \boldsymbol{X} \mid \boldsymbol{Z} \iff f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{i=1}^{m} f_{X_i|\boldsymbol{Z}}(x_i \mid \boldsymbol{z}).$$

for all $(\boldsymbol{x}, \boldsymbol{z}) \in \mathfrak{X} \times \operatorname{supp}(\boldsymbol{Z})$.

Proof. ($\Rightarrow$) follows by recursion from Thm 5.3. ($\Leftarrow$) follows by recursion from Thm 5.4. □

[3] '$\models$' is my invention. I have also seen mutually conditional independence written as $\perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z}$ but this is unintuitive to me, because I recognise $\perp\!\!\!\perp$ as the first separator of a ternary operator, as shown in Def. 5.1.

5.3 Modelling using conditional independence

To return to the main subject of this chapter: the benefits of conditional independence for constructing statistical models of a set of random quantities. Sec. 2.4 introduced the notion of a family of models, indexed by a parameter θ; and Sec. 2.6 explained how one route into a Bayesian approach was to take this parameter and specify a PMF for it. But now we are going to look at parameters in a different and more radical way, according to the following principle.

Definition 5.3 (Basic principle of modelling)**.** *To achieve the complex dependency structure you want across the random quantities of interest by implemented a much simpler structure over a larger collection of random quantities.*

My interest is in the set of random quantities $\boldsymbol{X}$, and I need to specify my beliefs in the form of a PMF, $f_{\boldsymbol{X}}$. One way to specify such a PMF is to introduce additional random quantities θ, specify the joint PMF of $[\boldsymbol{X}, \theta]$, and then (notionally) marginalise out θ, i.e.

$$f_{\boldsymbol{X}}(\boldsymbol{x}) \leftarrow \sum_t f_{\boldsymbol{X},\theta}(\boldsymbol{x}, t),$$

according to Thm 1.6. At the moment this looks harder than specifying $f_{\boldsymbol{X}}$ directly. But there are two ameliorating features. First, the joint distribution over $[\boldsymbol{X}, \theta]$ can be specified in two parts,

$$f_{\boldsymbol{X},\theta}(\boldsymbol{x}, t) \leftarrow f_{\boldsymbol{X}|\theta}(\boldsymbol{x} \mid t)\, f_\theta(t),$$

and I may then be able to use conditional independence to substantially simplify the form of $f_{\boldsymbol{X}|\theta}$, in particular.[4] Second, marginalising out θ has become a much cheaper calculation since the development of MCMC algorithms and high-performance computers (see Sec. 2.3). Without this second feature, exploiting conditional independence as a modelling strategy would be of only formal interest.

Exchangeability provides the canonical example of the power of this modelling strategy. My beliefs about $\boldsymbol{X}$ are exchangeable if and only if $f_{\boldsymbol{X}}$ is a symmetric function (Thm 4.4). Of the infinity of different symmetric functions, the one almost always favoured is the IID model

$$f_{\boldsymbol{X}}(\boldsymbol{x}) \leftarrow \sum_t \prod_{i=1}^m f_{X|\theta}(x_i \mid t)\, f_\theta(t)$$

(see Sec. 4.4). From Thm 5.5, we now recognise this as a special case of the statement that $\boldsymbol{X}$ is mutually conditionally independent (MCI) given θ, written $\vDash \boldsymbol{X} \mid \theta$. Or, in words, that $\{X_j\}_{j \neq i}$ does not bring any information about X_i which is not already present in θ. In this context θ is not really a parameter, as defined in Sec. 2.4, although it still retains that name. Instead, it is an auxiliary random quantity, introduced to implement a dramatic simplification in my specification of $f_{\boldsymbol{X}}$. Instead of specifying a symmetric $f_{\boldsymbol{X}}$ directly, I can now approach it indirectly by specifying a $f_{X|\theta}$ and a f_θ.

[4] In this chapter I am not using π_θ for the PMF of the parameters, as I did in Chapter 2, because the more general notion of parameter implied by Def. 5.3 extends way beyond an index of a family of distributions.

The key thing to appreciate is that although a construction such as $\models \boldsymbol{X} \mid \theta$ is a dramatic simplification of a more general $f_{\boldsymbol{X}\mid\theta}$, the step of marginalising out θ re-introduces relationships across the X's. We can see this from the marginal distribution $f_{\boldsymbol{X}}$, which does not have a product structure over the X's, indicating that the X's are *not* mutually independent, even though they are MCI given θ. So X_j *does* bring information about X_i. This is a very general point. $f_{\boldsymbol{X}\mid\theta}$ and f_θ may be constructed out of lots of rather simple products, indicating lots of conditional independencies, but none of these survive marginalising out θ.

So, to paraphrase Def. 5.3, we introduce additional random variables precisely to exploit the conditional independence structures that we can construct with them. Sec. 5.4 and Sec. 5.5 discuss two general strategies which between them cover most statistical modelling approaches.

5.4 *Hierarchical modelling*

Hierarchical modelling is the general strategy that is implemented in the approach to partial exchangeability given in Sec. 4.5.

Suppose that $\boldsymbol{X}$ can be divided into g groups,

$$\boldsymbol{X} := \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_g\},$$

including the special case where every element of $\boldsymbol{X}$ has its own group. The standard conditional independence structure would be

$$\models \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_g\} \mid \theta, \tag{†}$$

for some parameter θ. In this case, $\{\boldsymbol{X}_j\}_{j \neq i}$ brings no information about $\boldsymbol{X}_i$ which is not already present in θ. If I want to use (†), I have to choose θ to make this an appropriate representation of my beliefs.

Because θ is a parameter of my own invention, I might allow some of its elements to be group-specific,

$$\theta := (\phi_1, \dots, \phi_g, \psi).$$

Then I can use the stronger conditional independence structure

$$\models \{(\boldsymbol{X}_1, \phi_1), \dots, (\boldsymbol{X}_g, \phi_g)\} \mid \psi. \tag{‡}$$

In this case, $\{(\boldsymbol{X}_j, \phi_j)\}_{j \neq i}$ brings no information about $(\boldsymbol{X}_i, \phi_i)$ which is not already present in ψ. We can tell that (‡) is stronger than (†) because the ϕ's that were previously to the right of the bar in θ have jumped over to the left, leaving only ψ—see the interpretation of conditional independence given in Sec. 5.1. For the statistician, stronger = simpler.

According to Thm 5.5, (‡) is equivalent to

$$f(\boldsymbol{x}, \boldsymbol{\phi} \mid \psi) = \prod_{i=1}^{g} f_i(\boldsymbol{x}_i, \phi_i \mid \psi) \tag{5.2a}$$

where once again I am suppressing the subscripts on f, which can be inferred from the function arguments (the i on f_i is a reminder that this function can vary by i). The distributions in the product would usually be factorised using Thm 1.24 to give

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\phi} \mid \psi) &= \prod_{i=1}^{g} f_i(\boldsymbol{x}_i \mid \phi_i, \psi)\, f_i(\phi_i \mid \psi) \\ &= \prod_{i=1}^{g} f_i(\boldsymbol{x}_i \mid \phi_i, \psi) \cdot \prod_{i=1}^{g} f_i(\phi_i \mid \psi). \end{aligned} \tag{5.2b}$$

The attraction of this modelling approach is that it is an excellent platform for further simplifications. Recall the discussion at the end of Sec. 4.5 (and before that in Sec. 2.7), in which the statistician is aiming to capture the joint structure of $\boldsymbol{X}$ with a small number of parameters.

A tempting simplification is to make $f_i(x_i \mid \phi_i, \psi)$ the same for all i, but this is only possible if each $\boldsymbol{X}_i$ has the same realm. In the case where each X_{ij} has the same realm, though, $\boldsymbol{X}_i := (X_{i1}, \dots, X_{im_i})$ can be exchangeable for each i, with

$$f_i(\boldsymbol{x}_i \mid \phi_i, \psi) = \prod_{j=1}^{m_i} f_i(x_{ij}; \phi_i, \psi). \tag{5.2c}$$

Note that while it might be possible to replace f_i with a common model for all i, it is not necessary, and would not be appropriate if there was information in addition to ϕ_i which distinguished one group from another. Eq. (5.2) with a common model is the partially exchangeable model proposed in Sec. 4.5, with $\phi_i \leftarrow (\alpha_i, \sigma_i)$.

The next simplification would be to make $f_i(\phi_i \mid \psi)$ the same for all i, requiring just a single model $f(\cdot; \psi)$. In this case, ϕ is exchangeable.[5] Note, however, that making both simplifications does not make $\boldsymbol{X}$ exchangeable, e.g. because the marginal distribution of (X_{i1}, X_{i2}) is not the same as that of (X_{i1}, X_{j2}) when $j \neq i$. To make $\boldsymbol{X}$ exchangeable would require the restriction that $\phi_i \leftarrow \phi$ for all i. For most statistical modellers this would be a step too far: exchangeable $\boldsymbol{X}_i$ for each i plus exchangeable $\boldsymbol{\phi}$ is a happy compromise.

These types of models are known as *hierarchical models*, because they are structured hierarchically, from the top down, starting with the random quantities and then passing through one or more layers of parameters. They have a fairly conventional syntax, which I illustrate here.[6] In the general case, (5.2b),

$$\begin{aligned} \boldsymbol{X}_i \mid \phi_i, \psi &\overset{\text{ind}}{\sim} f_i(\cdot; \phi_i, \psi) && i = 1, \dots, g \\ \phi_i \mid \psi &\overset{\text{ind}}{\sim} g_i(\cdot; \psi) && i = 1, \dots, g \\ \psi &\sim h(\cdot). \end{aligned}$$

[5] To clarify: marginalising out $\boldsymbol{X}$ from (5.2b) shows that when $f_i \leftarrow f$ for all i, $f_{\boldsymbol{\phi}|\psi}$ has the IID exchangeable form $\prod_i f(\phi_i \mid \psi)$.

[6] Their conditional independence structure can also be represented graphically as a *Directed Acyclic Graph (DAG)*; see Cowell *et al.* (1999).

Thus '$\overset{\text{ind}}{\sim}$' and an index set indicates a product over the righthand side of the row. In the case where $\boldsymbol{X}_i$ is exchangeable, (5.2c),

$$\begin{aligned} X_{i1}, \dots, X_{im_i} \mid \phi_i, \psi &\overset{\text{iid}}{\sim} f_i(\cdot\,; \phi_i, \psi) && i = 1, \dots, g \\ \phi_i \mid \psi &\overset{\text{ind}}{\sim} g_i(\cdot\,; \psi) && i = 1, \dots, g \\ \psi &\sim h(\cdot). \end{aligned}$$

Thus '$\overset{\text{iid}}{\sim}$' indicates a product within the row, over the comma-separated list to the left of '|'. So the first row of this hierarchical model indicates a double product, the inside product over $j = 1, \dots, m_i$, and the outside product over $i = 1, \dots, g$. In the case where $\boldsymbol{\phi}$ is exchangeable, the second row can also be written as

$$\phi_1, \dots, \phi_g \mid \psi \overset{\text{iid}}{\sim} g(\cdot\,; \psi),$$

which I slightly prefer. It is quite common, where the distribution on the righthand side has a well-known form, to suppress the slot for the argument, and just write, for example, $\phi_1, \dots, \phi_g \mid \psi \overset{\text{iid}}{\sim} \text{Exp}(\psi)$ if g is an Exponential distribution with rate ψ, or $\psi \sim \text{Ga}(0.1, 0.1)$ if h is a Gamma distribution with shape 0.1 and rate 0.1.

This arrangement, of $\boldsymbol{X}$ on the top level, group-specific parameters on the middle level, and 'global' parameters on the bottom level, is very common. In the case where $\boldsymbol{X}_i \perp\!\!\!\perp \psi \mid \phi_i$ for all i it would be usual to suppress the conditioning on ψ at the top level; in this case ψ would often be referred to as a *hyperparameter*, being a parameter of the model for $\boldsymbol{\phi}$. All three levels can be expanded to allow for richer models, incorporating a more detailed group structure, or allowing for random quantities of different types (i.e. with different realms).

Hierarchical models are ideally adapted for MCMC by *Gibbs sampling*, as discussed in Lunn *et al.* (2013) and Gelman *et al.* (2014). It is worth stressing again that a hierarchical modelling strategy is effective precisely because the introduction of additional random quantities, the parameters, does not impose a large computational cost.

* * *

To illustrate a hierarchical model, suppose that each i is an active Japanese stratovolcano, and $\boldsymbol{X}_i$ represents its recent large-eruption history. There is an accepted statistical model for $\boldsymbol{X}_i$ which depends on three parameters, (μ_i, σ_i, ξ_i). The client—a vulcanologist—wants to treat these volcanoes as similar but not identical, which I plan to implement by making the parameter vectors exchangeable across volcanoes, using the IID model

$$f(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi}) = \prod_{i=1}^{g} f(\mu_i, \sigma_i, \xi_i).$$

Unfortunately, the eruption histories are rather meagre, and a model with three parameters per volcano will have low tenability (see Sec. 2.8). So I take the most difficult of the three parameters, ξ_i, representing 'shape', and treat it as common across volcanoes, giving

$$\begin{aligned} \boldsymbol{X}_i \mid \mu_i, \sigma_i, \xi &\overset{\text{ind}}{\sim} f_i(\cdot\,; \mu_i, \sigma_i, \xi) \qquad i = 1, \dots, g \\ (\mu_1, \sigma_1), \dots, (\mu_g, \sigma_g) \mid \xi &\overset{\text{iid}}{\sim} g(\cdot\,; \xi) \\ \xi &\sim h(\cdot). \end{aligned}$$

While I could go further and treat, say, σ_i as common as well, I prefer not to, because the client believes that volcanoes can vary in at least two properties: their maximum eruption magnitude and their repose time (time between eruptions). Both of these quantities are important for a risk assessment, and it seems better not to constrain them to lie on a 1D manifold for given ξ, which would be the consequence of having only one volcano-specific parameter.

5.5 Markov Random Fields

In Sec. 5.4, conditional independence was used to represent the relationship among the X's according to their group structure: whether elements i and j were in the same group, or in similar groups (for compound groups). This involved the introduction of parameters, possibly in two or more levels, in order to use IID representations of exchangeability.

In this section, conditional independence is used to factorise the PMF of $\boldsymbol{X}$ directly. This approach has proved to be very powerful in specifying PMFs for a large number of random quantities of the same type, which are indexed in some way with respect to an underlying domain. The pixels in an image is a classic example, where each X_i has a coordinate in $\mathbb{R}^2$, but the concept is very general. Details can be found in Cressie and Wikle (2011, ch. 4), Rue and Held (2005), or Banerjee *et al.* (2004).

So consider $\boldsymbol{X} := (X_1, \dots, X_m)$, a collection of random quantities of similar type, in some kind of relationship with each other. According to Factorisation theorem (Thm 1.23), we can always write

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) \prod_{i=2}^{m} f_{X_i \mid \boldsymbol{X}_{1:(i-1)}}(x_i \mid \boldsymbol{x}_{1:(i-1)}),$$

in an obvious notation. This is a recipe for constructing $f_{\boldsymbol{X}}$, by specifying each of the terms on the righthand side. This might be m times as hard as specifying $f_{\boldsymbol{X}}$ directly, except for the possibility of conditional independence.

For example, suppose that i corresponds to a time t_i, with $t_i < t_j$ whenever $i < j$; thus $\boldsymbol{X}$ is a time-series. In this case I might believe that the present was conditionally independent of the distant past conditional on the recent past, for example

$$X_i \perp\!\!\!\perp \boldsymbol{X}_{1:(i-2)} \mid X_{i-1} \quad i = 3, \dots, m.$$

Then Thm 5.1 implies that

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) \prod_{i=2}^{m} f_{X_i \mid X_{i-1}}(x_i \mid x_{i-1})$$

This is known as a *Markov process*.[7] As a further simplification, if

$$f_{X_i \mid X_{i-1}}(x \mid x') = f_{X_2 \mid X_1}(x \mid x') \quad i = 2, \dots, m$$

then this is a *homogeneous* Markov process. A homogeneous Markov process is a very attractive model when $\boldsymbol{X}$ is a time-series, because the whole of $f_{\boldsymbol{X}}$ is induced by specifying f_{X_1} and $f_{X_2 \mid X_1}$. Indeed, there is a further simplification, because $f_{X_2 \mid X_1}$ may have a stationary distribution f^* satisyfing

$$f^*(x) = \sum_{x'} f_{X_2 \mid X_1}(x \mid x') \, f^*(x')$$

in which case f_1 can be replaced by f^*. An entire $f_{\boldsymbol{X}}$ for a single $f_{X_2 \mid X_1}$ is a real bargain, and it is not surprising that homogeneous Markov processes with stationary distributions are popular in time-series modelling; see Chatfield (2004) for more details.

And, just to be absolutely clear, note that the conditional independence in a first order Markov process does not imply independence between non-neighbours. Just taking the first three terms,

$$\begin{aligned} f_{X_1, X_3}(x_1, x_3) &= \sum_{x_2} f_{X_1, X_2, X_3}(x_1, x_2, x_3) \\ &= \sum_{x_2} f_{X_1}(x_1) \, f_{X_2 \mid X_1}(x_2 \mid x_1) \, f_{X_3 \mid X_2}(x_3 \mid x_2) \\ &= f_{X_1}(x_1) \sum_{x_2} f_{X_2 \mid X_1}(x_2 \mid x_1) \, f_{X_3 \mid X_2}(x_3 \mid x_2) \\ &= a(x_1) \, b(x_1, x_3), \text{ say,} \end{aligned}$$

which does not factorise in the form $a(x_1)\, b(x_3)$, as would be necessary and sufficient for X_1 and X_3 to be independent (Thm 5.4). Therefore a Markov process can still have lots of long-range dependence.

This approach to time-series modelling works because there is an ordering on $\boldsymbol{X}$ which is compatible with my beliefs about $\boldsymbol{X}$. Unfortunately, however, many interesting $\boldsymbol{X}$'s do not admit such an ordering. For example, suppose that each X_i corresponds to a measurement at location $s_i \in \mathcal{S} \subset \mathbb{R}^2$. I could easily order the $\boldsymbol{X}$ according to their s's. For example, I could use a *lexicographic order* on s_i.[8] But this will sometimes put X_i and X_j nearby in the ordering, even though s_i and s_j are far apart in $\mathcal{S}$. This is an unavoidable difficulty once the domain of the elements of $\boldsymbol{X}$ has more than one dimension.

Happily, there is another factorisation of $f_{\boldsymbol{X}}$, which generalises the Markov factorisation of a time-series. Let $\boldsymbol{X}_{-i}$ denote every random quantity in $\boldsymbol{X}$ bar X_i. The PMF $f_{X_i \mid \boldsymbol{X}_{-i}}$ is termed the *full conditional* PMF of X_i. The following result states that $f_{\boldsymbol{X}}$ can be factorised in terms of its full conditionals.[9]

[7] This is a *first order* process. Higher-order processes have more of the history to the right of the bar; they are richer in the types of belief they can represent, but also more expensive to work with.

[8] Order by the first component of s_i, and then break ties by ordering by the second component.

[9] Cressie and Wikle (2011, p. 177) dispute that this should be called 'Brook's Lemma', which is the name given by Rue and Held (2005, sec. 2.2); they would prefer 'Besag's Lemma', from Besag (1974).

Theorem 5.6 (Brook's lemma)**.** *Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be two points in the support of $f_{\boldsymbol{X}}$. Then*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x}') \prod_{i=1}^{m} \frac{f_{X_i \mid \boldsymbol{X}_{-i}}(x_i \mid x_1, \dots, x_{i-1}, x'_{i+1}, \dots, x'_m)}{f_{X_i \mid \boldsymbol{X}_{-i}}(x'_i \mid x_1, \dots, x_{i-1}, x'_{i+1}, \dots, x'_m)}.$$

Proof. It suffices to give this proof for $m = 3$.

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= f_{X_1,X_2,X_3}(x_1, x_2, x_3) \\ &= f_{X_3 \mid X_1,X_2}(x_3 \mid x_1, x_2)\, f_{X_1,X_2}(x_1, x_2) \\ &= \frac{f_{X_3 \mid X_1,X_2}(x_3 \mid x_1, x_2)}{f_{X_3 \mid X_1,X_2}(x'_3 \mid x_1, x_2)}\, f_{X_1,X_2}(x_1, x_2)\, f_{X_3 \mid X_1,X_2}(x'_3 \mid x_1, x_2) \\ &= \frac{f_{X_3 \mid X_1,X_2}(x_3 \mid x_1, x_2)}{f_{X_3 \mid X_1,X_2}(x'_3 \mid x_1, x_2)}\, f_{X_1,X_2,X_3}(x_1, x_2, x'_3). \end{aligned}$$

Now iterate on the final term. □

To see how this factorisation works, fix on a value $\boldsymbol{x}'$ in the support of $f_{\boldsymbol{X}}$, say $\boldsymbol{x}' = \boldsymbol{0}$. Then

$$f_{\boldsymbol{X}}(\boldsymbol{x}) \propto \prod_{i=1}^{m} \frac{f_{X_i \mid \boldsymbol{X}_{-i}}(x_i \mid x_1, \dots, x_{i-1}, 0, \dots, 0)}{f_{X_i \mid \boldsymbol{X}_{-i}}(0 \mid x_1, \dots, x_{i-1}, 0, \dots, 0)}.$$

The only terms on the righthand side are the full conditionals, with the missing normalisation constant being computable by summing over all $\boldsymbol{x} \in \mathcal{X}$.

Now each of these full conditionals can be simplified using beliefs about conditional independence, as in Thm 5.1. But there is a catch. Although every coherent $f_{\boldsymbol{X}}$ can be decomposed into its full conditionals, not every set of candidates for $\{f_{X_i \mid \boldsymbol{X}_{-i}}\}$ can be reassembled into a coherent $f_{\boldsymbol{X}}$. But there is a sufficient condition under which this can be done, and which also provides a convenient alterntive way to construct an $f_{\boldsymbol{X}}$ which respects a rich set of conditional independence relations.

When I consider each X_i in turn, I can identify the set of *neighbours* of X_i,

$$\mathrm{ne}_i \subset \{1, \dots, m\} \setminus \{i\},$$

which a subset of the other X's for which my beliefs satisfy

$$X_i \perp\!\!\!\perp \text{everything else} \mid \boldsymbol{X}_{\mathrm{ne}_i}. \tag{†}$$

But this is where incoherence can arise, because my marginal choices for ne_i for each i might be jointly incompatible over $i = 1, \dots, m$. The solution is to ensure that the entire set of neighbours can be represented by an *undirected graph* $\mathcal{G} := \{V, E\}$ with vertices $V := \{1, \dots, m\}$, and with an edge $(i, j) \in E$ exactly when $j \in \mathrm{ne}_i$. This restriction to undirected graphs ensures that

$$i \in \mathrm{ne}_j \iff j \in \mathrm{ne}_i,$$

because an edge from i to j is the same as an edge from j to i. If I was specifying the ne_i's marginally, this is the kind of restriction that I might not impose. A graph satisfying (†) is said to respect the *local Markov property*.

From any graph $\mathcal{G}$ we can extract the set of *cliques*, $\mathcal{C}$. $C \subset V$ is a clique of $\mathcal{G}$ exactly when it is a singleton, or E contains a complete set of edges for C. But we can also go backwards: from a set of cliques $\mathcal{C}$ we can reconstruct the vertices V and the edges E. So there are two equivalent ways of representing the local Markov property, as the graph $\mathcal{G}$, or as a set of cliques $\mathcal{C}$. This equivalence is at the heart of the following result, proved in Besag (1974), but named after its originators. Lauritzen (1996, ch. 3) has more details and an alternative proof.[10]

Theorem 5.7 (Hammersley-Clifford theorem)**.**
If $\text{supp}(X_1, \dots, X_m) = \prod_{i=1}^m \text{supp}(X_i)$*, then* f_X *respects the local Markov property of graph* $\mathcal{G}$ *if and only if it factorises as*

$$f_X(x) \propto \prod_{C \in \mathcal{C}} \phi_C(x_C), \tag{5.3}$$

where $\mathcal{C}$ *are the cliques of* $\mathcal{G}$ *and the* ϕ_C*'s are positive functions which depend only on* x_C*.*

Thus the tractable approach to incorporating conditional independence structure into f_X is to construct an undirected graph with the local Markov property, identify its cliques, and use these and choices for the ϕ_C's to construct f_X using (5.3). A PMF that is specified in this way is termed a *Markov Random Field*.

There are some conventional approaches to defining $\mathcal{G}$. Suppose that each X_i corresponds to a value s_i in some domain for which we can define a *dissimilarity measure d*.[11] Then a common choice is

$$\text{ne}_i = \{j : d(s_i, s_j) \le c\} \setminus \{i\}.$$

I.e., join i and j in $\mathcal{G}$ whenever they are not more than c apart. This includes the case where the domain is Euclidean and the dissimilarity measure is distance. Or each i might be a person, and $-d(s_i, s_j)$ might be a measure of consanguinity, or the average number of daily contacts.[12] Another approach is useful when each i represents an area, like a pixel or a region: ne_i might be all of the other elements which share a boundary with i.

We can also bootstrap our way to more complicated neighbourhoods by iterating simple ones. For example, creating a new neighbourhood structure $\mathcal{N}^1$ from $\mathcal{N}^0$ in which i and j are neighbours in $\mathcal{N}^1$ if i and j are neighbours of neighbours in $\mathcal{N}^0$.

Incorporating parameters into f_X is simple in theory, although it does pose difficulties in computation. Take the clique representation, and add in a parameter θ to give

$$f_X(x; \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \phi_C(x_C; \theta),$$

[10] My statement is not 'the' Hammersley-Clifford theorem, but it is equivalent to it.

[11] A measure with the properties $d(s_i, s_j) \ge 0$, $d(s_i, s_i) = 0$, and $d(s_i, s_j) = d(s_j, s_i)$.

[12] These might be suitable choices for epidemiology.

where

$$Z(\theta) := \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{C \in \mathcal{C}} \phi_C(\boldsymbol{x}_C; \theta).$$

Z is known as the *partition function*. If $\mathcal{X}$ is large then $Z(\theta)$ is extremely expensive to compute, but it needs to be recomputed in every iteration of a MCMC simulation from the PMF of $\boldsymbol{X}$ conditional on the dataset, in which θ is treated as uncertain. The upshot is that for really large applications, choices for $\mathcal{G}$ and the clique functions are limited to cases for which there is a closed-form expression for Z, or a good approximation. These tend to be *Gauss Markov Random Fields*, see Rue and Held (2005). Some well-studied applications, such as the Ising Model, have efficient bespoke simulation algorithms, see Liu (2001).

6
Model diagnostics and model choice

It is a complicated business to estabish whether a statistical inference is appropriate, and we should beware of facile answers to complicated questions. This chapter covers model checking (Sec. 6.1), model choice (Sec. 6.2), hypothesis testing (Sec. 6.3), significance levels (Sec. 6.4), and confidence sets (Sec. 6.5). Some strong opinions will be expressed.

6.1 Turing tests

There is a precise sense in which an inference is appropriate. Refer back to Chapter 2, notably Sec. 2.4 and Sec. 2.6. An inference $\mathrm{E}^*(\cdot)$ or $\hat{\mathrm{E}}^*(\cdot)$ is appropriate exactly when it reflects the beliefs of the client.[1] The statistician should not take this for granted: the client ought to insist on a demonstration.

We might contemplate comparing the client's beliefs with E^*. For example, the client might have quite strong beliefs about some random quantity $h(\boldsymbol{X})$, and then it is easy to see whether $\mathrm{E}^*\{h(\boldsymbol{X})\}$ is similar to them. But strong beliefs ought already to have been incorporated into E^*, and so this simple criterion might be vacuous. If we keep some strong beliefs back for checking, then we have not addressed the question of whether, once those beliefs are incorporated, E^* is appropriate.

Recollect that E^* is constructed by conditioning on the random proposition Q, where the truth of Q represents the dataset, or a portion of it. Another superficially attractive approach is to compare E^* with the values of data not used in Q. But this does not answer the question of whether E^* represents the client's beliefs; it answers a different question, which will be explored in Sec. 6.2.4 and Sec. 6.4. It also runs into the same difficulty as above: if these held-out data have an expectation, then why were they not included in Q?

What we would really like to do is to construct E^* using as many of the clients beliefs and as much of the dataset as possible, and *then* demonstrate to the client that E^* is appropriate. To set the scene, consider this observation from McWilliams (2007):

> [Atmospheric and Ocean Simulation] models yield space-time patterns remeniscent of nature (e.g., visible in semiquantitative, high-resolution satellite images), thus passing a meaningful kind of *Turing*

[1] I will continue to say 'the client', even though it may be the client's experts, or, indeed, the statistician may be his own client. In this section I will focus on E^*; the same issues apply to $\hat{\mathrm{E}}^*$.

> *test* between the artifical and the actual. (p. 8709, emphasis added)

This is not a vacuous comparison in climate modelling, because the ocean simulation is not conditioned on the satellite observations. Rubin (1984, sec. 5) proposed a method for making the same kind of comparison in statistical inference. This involves 'cloning' the observations, so that the comparison can be between the actual observations and the clones.[2] The inference passes its Turing test if the client cannot spot the observations among the clones.

Consider the simple observation model (eq. 2.6), for which we can compute f_Y, the PMF of the observations. We can create a clone of $\boldsymbol{Y}$ using just this and the prior distribution π_θ, by imposing the two properties:

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^* \mid \theta \quad \text{and} \quad f_{\boldsymbol{Y}^*}(\boldsymbol{y}; \theta) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta). \tag{†}$$

The conditional PMF of the clone given the actual observations $\boldsymbol{y}^{\text{obs}}$ is the marginal distribution of

$$\begin{aligned} &\Pr(\boldsymbol{Y}^* \doteq \boldsymbol{y}', \theta \doteq t \mid \boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}) \\ &\qquad = \Pr(\boldsymbol{Y}^* \doteq \boldsymbol{y}' \mid \theta \doteq t, \boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}) \Pr(\theta \doteq t \mid \boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}) && \text{seq. cond., Thm 1.24} \\ &\qquad = f_{\boldsymbol{Y}}(\boldsymbol{y}'; t)\, \pi^*_\theta(t) && \text{by (†),} \end{aligned}$$

where π^*_θ is the posterior distribution (see Sec. 2.6). To simulate a realisation of $\boldsymbol{Y}^*$ conditional on $\boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}$, simply simulate a t^* from the posterior distribution, and then simulate a $\boldsymbol{Y}^*$ from the model $f_{\boldsymbol{Y}}(\cdot; t^*)$.

To run the Turing test, the statistician generates, say, eight clones, and then presents these eight and the actual observations $\boldsymbol{y}^{\text{obs}}$ to the client, to see if she can spot the actual observations among the clones. The more clones the better of course, in terms of the power of the test, but the client's time and patience are also an issue. Several observations are in order.

First, it may not be necessary to involve the client at all. If the statistician can spot the observations among the clones then no doubt the client can too. So early in the model development process the statistician alone runs the Turing test, saving the client for later in the process, when more nuanced beliefs come into play.

Second, the client may know the observations well enough to spot them, in the same way that we might be able to spot the Plough[3] among other less-well-known constellations, just through familiarity. In this case she is assessing whether the clones are different from the observations in ways which are decision-relevant. Another possibility is to present the client with dimensionally-reduced summaries of the observations and the clones. These summaries might conceal familiar features.

Third, and following on from the previous point, I would resist attempting to 'score' the observations in terms of the distribution of the clones, although this was Rubin's original proposal, and now goes by the name of *posterior predictive P-values*; see, e.g., Gelman *et al.* (2014, ch. 6). In common with most statisticians, I have some

[2] Prof. Rubin did not write 'clones', of course.

[3] AKA the Big Dipper.

serious reservations about P-values, to be discussed in Sec. 6.4. Putting those aside, I would prefer to keep the client in the loop, unless she is very clear that she can be replaced with a test statistic. The statistician needs to know how it was that the client was able to spot the observations among the clones, because this directs the further development of the model. I also think it is important to keep the client engaged in the modelling process. Otherwise she plays a somewhat oracular role, and it is the poor statistician who has to fill in the gaps.

How insightful is the client? We should bear in mind that the client may not be very knowledgeable about $\boldsymbol{X}$, compared to some other expert. That is to say, E^* might pass the Turing test with the client, but not with some other expert. This is the client's concern, not the statistician's.

6.2 Model choice

Now consider the case where there are competing proposals for f_X. This would usually arise where the client has several groups of experts, with incompatible beliefs. For example, the client might be the Intergovernmental Panel on Climate Change (IPCC) and the experts might be the different climate modelling groups, where each group has one or more simulators of future weather, which can be used to induce a PMF (see Rougier and Goldstein, 2014). Or else the client might be the State of California, and the experts might be different earthquake modelling groups.[4]

These are large-scale examples, but there are many smaller-scale ones as well, such as a catastrophe modelling company which requires a storm model, or an engineering consultancy which requires a fatigue model, or a pharmaceutical company which requires a model for metabolic processes, and so on. In each case a search of the literature will reveal a number of alternatives. For reasons of cost the client would like to choose a single one, but it is important to appreciate that she does not have to.

6.2.1 Belief pooling and model averaging

Let the competing models for $\boldsymbol{X}$ be represented as the PMFs

$$\{f_m\}_{m \in \mathcal{M}}$$

where I have suppressed the subscript $\boldsymbol{X}$ on f_X because I need space for the subscript m. The client has an application, and for this application she would like to select a subset of these models—say just one for simplicity—because it is cheaper than maintaining and using all of the models. We will analyse this decision as though she could proceed with all of the models, in order to decide on the appropriate criterion for selecting just one.

[4] See `http://www.cseptesting.org/` for an example of a large experiment to choose between different models for earthquakes.

The client always has the option to combine her models into a single 'super-model'. There are two main approaches:

$$
\begin{aligned}
f_X(x) &= \sum_m f_m(x)\, w_m && \text{linear pooling} \\
f_X(x) &\propto \prod_m f_m(x)^{w_m} && \text{logarithmic pooling,}
\end{aligned}
$$

For linear pooling it is necessary and sufficient that $w \in \mathbb{S}^{|\mathcal{M}|-1}$, in order that f_X is always a PMF. For logarithmic pooling it is necessary and sufficient that $w_m \geq 0$. One advantage of linear pooling is immediately apparent: there is no need to compute a normalising constant for f_X.

The reason for two approaches is that both approaches have attractive and unattractive theoretical properties.[5] Linear pooling has the attractive property that it preserves a common ordering across the probabilities of random propositions. In other words, if every model in $\mathcal{M}$ implies that $\Pr_m(P) \leq \Pr_m(Q)$, then $\Pr(P) \leq \Pr(Q)$ in the pooled model as well. Another attractive property is that $f_X(x)$ depends only on the probabilities assigned to x. A third is that the support of f_X is the union of the supports of the f_m's. Logarithmic pooling does not have any of these properties.

On the other hand, logarithmic pooling is invariant to the order of pooling and conditioning. With linear pooling the result will typically be different if we pool first and then condition, or if we condition first and then pool—as shown below. This implies that every model in $\mathcal{M}$ might treat X_1 and X_2 as probabilistically independent, and yet the pooled model might not.[6] Logarithmic pooling has the very unattractive property that the support of f_X is the intersection of the supports of the f_m's; in other words it takes only one model to assert $f_m(x) = 0$ to ensure that $f_X(x) = 0$.

Overall, linear pooling seems to have won out due to its practical simplicity, and its intuitive form when pooling first and then conditioning. The default position would be to take the weights equal, but it is useful to have the flexibility to go further. For example, two similar models could share the weight of one model, or a model that was apparently deficient (e.g. missing a process) could be down-weighted. For the climate simulators used by the IPCC, the default position of the IPCC is to give all of the models equal weight. But most climate scientists would definitely have a view about non-equal weights, reflecting all sorts of things like simulator genealogies, and the accumulated experience of the research groups.

As in Chapter 2, represent the dataset as the truth of the proposition $Q := q(X)$, where q is a first-order sentence. Now consider the

[5] More details are available in Cooke (1991, ch. 11); note the typo on p. 172, item 6, where the inequalities go the wrong way. See also the recent survey by French (2011).

[6] Cooke (1991, p. 174) notes that this is not necessarily an attractive property if the models themselves disagree one with another about the marginal probabilities of X_1 and X_2.

effect of conditioning the linearly pooled model:

$$\begin{aligned} f_X^*(x) &:= \Pr(X \doteq x \mid Q) \\ &\propto \mathbb{1}_{q(x)}\, f_X(x) && \text{Muddy table theorem, Thm 2.1} \\ &= \mathbb{1}_{q(x)} \textstyle\sum_m f_m(x) \cdot w_m && \text{linear pooling} \\ &= \textstyle\sum_m \mathbb{1}_{q(x)} f_m(x) \cdot w_m \\ &= \textstyle\sum_m f_m^*(x) \cdot \Pr_m(Q)\, w_m \end{aligned}$$

where

$$f_m^*(x) := \Pr_m(X \doteq x \mid Q) = \frac{\mathbb{1}_{q(x)}\, f_m(x)}{\Pr_m(Q)}$$

by the Muddy table theorem again. Reincorporating the normalising constant $\Pr(Q)^{-1}$ then gives

$$f_X^*(x) = \textstyle\sum_m f_m^*(x) \cdot w_m^*$$

where

$$w_m^* := \frac{\Pr_m(Q)\, w_m}{\Pr(Q)} = \frac{\Pr_m(Q)\, w_m}{\sum_{m'} \Pr_{m'}(Q)\, w_{m'}}.$$

So conditioning the linear pooled model on Q has two parts: conditioning each model on Q, and updating the weights from w to w^*. Combining multiple models into an inference in this way is termed *Bayesian Model Averaging (BMA)*; see Hoeting *et al.* (1999) for a review.

This update of the weights is the reason that linear pooling is sensitive to the order of pooling and conditioning. But the update of the weights is also one of the most attractive features of this procedure. Observe that the expression for w^* has the form analysed in Sec. 2.7. Hence we can consider the stability conditions in order to determine whether, for the dataset represented by Q, the simpler expression

$$\tilde{w}_m := \frac{\Pr_m(Q)}{\sum_{m'} \Pr_{m'}(Q)}$$

provides a good approximation to w^*. If so, the precise values of the linear combination w can be neglected (without needing to be all the same), and the weights and the conditional PMF are determined entirely by the set of models and the dataset Q.

The value $\Pr_m(Q)$ is termed the *evidence* of model m. The suitability of the evidence in updating the weights depends on $\Pr_m(Q)$ being a reasonable representation of modelling group m's beliefs about the dataset. As discussed at the end of Sec. 2.7, and also touched on in Sec. 6.1 above, the statistician in group m may decide that the stability conditions hold, so that he can replace a carefully-considered prior distribution π_θ with something flat and tractable. In this case the X margin of the joint distribution (X, θ) may *not* be a reasonable representation of group m's beliefs about X, and hence $\Pr_m(Q)$ may not be a reasonable representation of group m's beliefs about the dataset. This is a serious impediment to the model choice methods discussed in the next two subsections, and a reason to favour cross-validation methods, discussed in Sec. 6.2.4.

6.2.2 *Model choice as a decision problem*

As shown in the previous subsection, the client could have pooled her models and derived a BMA representation of her beliefs, $f^*_{\boldsymbol{X}}$. Even if she chooses not to do this, but to proceed with one model alone, she still has the option to assess each of the models according to how well it matches $f^*_{\boldsymbol{X}}$ in her application. And this last point is crucial: she needs a model for a reason, and this reason will help to determine which model she should choose.

Now would be a good time to refer back to Sec. 3.2 and Sec. 3.5 for a review of decision theory and decision rules. The client has an action set $\mathcal{A}$ and a loss function $L(a, \boldsymbol{x})$. Her best choice of action is the one that minimises her expected loss $\mathrm{E}^*\{L(a, \boldsymbol{X})\}$.[7] Both $\mathcal{A}$ and L may be somewhat simplified compared to the client's actual application, but the idea is that they are representative, in the sense that conclusions she draws from the simplified analysis are informative about how she should proceed in her actual application.

Using her BMA representation, the client can identify her best choice of action,

$$a^* := \operatorname*{argmin}_{a \in \mathcal{A}} E^*\{L(a, \boldsymbol{X})\},$$

for which her Bayes risk is defined as $R^* := E^*\{L(a^*, \boldsymbol{X})\}$. She can also identify her best choice of action using each model on its own,

$$a_m := \operatorname*{argmin}_{a \in \mathcal{A}} E^*_m\{L(a, \boldsymbol{X})\} \quad m \in \mathcal{M}.$$

When she uses the action optimal for model m in place of her optimal action a^* her risk is

$$R(m) := \mathrm{E}^*\{L(a_m, \boldsymbol{X})\}$$

where, necessarily, $R^* \leq R(m)$. The different $R(m) - R^*$ is the expected additional loss she incurs from using just model m. The optimal single model is therefore

$$\widetilde{m} := \operatorname*{argmin}_{m \in \mathcal{M}} R(m).$$

This analysis makes it clear that the choice of a single model from $\mathcal{M}$ depends on the dataset, but also on the action set and loss function; i.e. on the client's application.

To illustrate, consider two different applications. In one, the loss is approximately linear in $\boldsymbol{X}$ for each action; for example, the action is the amount of advertising spending and the loss is the negative revenue. In this case, a good model has $\mathrm{E}^*_m(\boldsymbol{X}) \approx \mathrm{E}^*(\boldsymbol{X})$. In another application, the loss is highly non-linear in $\boldsymbol{X}$ for some actions. This is typically the case for hazards, such as flooding; for example, the

[7] I assume, purely for simplicity, that the client's choice of action has no impact on her beliefs about $\boldsymbol{X}$.

action is the height of the flood defenses, and the loss is the amount of land inundated. In this case, a good model has $f_m^* \approx f_{\boldsymbol{X}}^*$ for the extreme values of $\boldsymbol{X}$.

There are just two conditions in which the choice of model does not depend sensitively on the action set and loss function. If the relative likelihood is almost entirely concentrated on one model, the maximum likelihood model $\hat{m}$, then the Stable estimation theorem (Thm 2.4) implies that

$$w_m^* \approx \begin{cases} 1 & m = \hat{m} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $f_{\hat{m}}^* \approx f_{\boldsymbol{X}}^*$, and consequently $a_{\hat{m}} \approx a^*$ no matter what the action set and loss function. Hence there is no additional expected loss from replacing $\mathcal{M}$ with the single model $\hat{m}$. Second, if the action set is really small—perhaps having only two elements—then there is a chance that the maximum likelihood action will be the same as the optimal action; but this is not something one could just assume.

Except under these conditions, though, the action set and loss function *should* play a role in selecting a single model from a set of competing models. Milner and Rougier (2014) provide a short example of model choice using simple assessments of loss.

6.2.3 Where does this going wrong?

There is a huge amount of confusion in statistical model choice, and many competing criteria. This reflects an unwillingness on the part of the statistician and/or the client to think explicitly about the underlying application, which I have represented above as an action set and a loss function. It goes without saying that the client is choosing between models for a reason; we should not be surprised that neglecting the reason leads to disarray.

If we strike out both the action set and the loss function, then the main thing left to judge the model on is $\Pr_m(Q)$, the 'evidence' of model m. As outlined at the end of the previous subsection, if the evidence of one of the models is a lot larger than the sum of the evidences of the other models, and if the action set is small, then selecting the model with the largest evidence is a defensible choice. Otherwise, further checks are required.

At the very least, the statistician would need to check that the model with the largest evidence was a member of a subset of the models which together made up most of the relative likelihood, and for which f_m^* was similar across the subset. Checking f_m^* is important, because different models will connect Q and $\boldsymbol{X}$ in different ways. 'Similar' is tough to quantify, because what is really needed is an assessment in terms of the action set and the loss function. But one could imagine a metric on PMFs for $\boldsymbol{X}$ which tried to reflect aspects of the PMF of $\boldsymbol{X}$ which are important in the client's application. In some cases this could be as simple as checking the each of the f_m^*'s had similar expectations (linear loss function) or expectations and variances (quadratic or convex loss function, see Sec. 3.6).

But at this point another difficulty looms. In a parametric approach, $f_m(x)$ is constructed as the $\boldsymbol{X}$ margin of the joint PMF

$$\Pr_m(\boldsymbol{X} \doteq \boldsymbol{x}, \theta_m \doteq t) = f_m(\boldsymbol{x}; t)\, \pi_m(t) \qquad t \in \Omega_m.$$

There are two difficulties here. First, as already discussed, the Bayesian statistician in group m may be happy (keen!) to replace his considered prior distribution π_m with a flatter more tractable alternative, on the grounds that the stability conditions of Sec. 2.7 hold. So $Pr_m(Q)$ is not representative of his group's beliefs about Q. Second, the Frequentist statistician in group m' may not want to supply a $\pi_{m'}$ at all. So we may have to do model choice without the evidences.

This is where *Information Criteria* come in. There are lots of different ones, originating with the Akaike Information Criterion (AIC). All Information Criteria have the same form, comprising a goodness-of-fit term for the dataset and a penalty for model complexity. A penalty is required, because more complex models will tend to fit the same dataset better than simpler ones; e.g. a quadratic will fit a time-series better than a straight line. The penalty guards against *over-fitting*, in which a good fit within the dataset can lead to bad fits outside it. Information Criteria are presented in a negative orientation, so that smaller is better.

For example, the Bayes Information Criterion (BIC) under the simple observation model (eq. 2.6) is

$$BIC_m(\boldsymbol{y}) := -2 \log f_m(\boldsymbol{y}; \hat{\theta}_m(\boldsymbol{y})) + k \log n$$

where $\hat{\theta}_m$ is the Maximum Likelihood Estimator of model m, and $k = \dim \Omega_m$. The first term measures goodness-of-fit by the maximum of the log-likelihood, and the second term penalises the number of model parameters by the log of the number of observations. The difference in the BICs of two models is a first-order approximation to the difference in the log-evidences; see Kass and Raftery (1995, sec. 4). This may be preferred to the actual difference in the log-evidences if the actual prior distributions are too flat, but in this case the BIC is being preferred because it is *not* a good approximation, which is a delicate argument.

Information Criteria are discussed in Gelman *et al.* (2014, ch. 7). The most popular one at the moment is the DIC (Spiegelhalter *et al.*, 2002).[8] This is for a couple of reasons:

1. It is easy to compute on the back of a Monte Carlo simulation from the posterior distribution of the model parameters; indeed DIC is built-in to software tools such as BUGS (Lunn *et al.*, 2013, sec. 8.6).

2. It has a sophisticated assessment of model complexity which goes beyond simply counting the parameters. This is important

[8] Be sure to read the discussion and rejoinder of this paper.

for hierarchical models in which exchangeability modelling creates large numbers of highly dependent parameters (Sec. 5.4).

As I hope I have made clear, I doubt that Information Criteria are appropriate for helping a client to choose a model in an important application. But by all means use them if you—the statistician—are your own client, and you are working on something not important enough to devote much thought to.

6.2.4 *Cross-validation*

Two difficulties were identified in Sec. 6.2.3. First, the relationship between Q and $\boldsymbol{X}$ can vary by model, so selection on the basis of Q is not necessarily helpful when the client's loss depends on $\boldsymbol{X}$. Second, the evidence $\Pr_m(Q)$ might not be available, because of a reluctance to provide a carefully-considered prior distribution for the parameters of model m, or even any prior distribution at all. *Cross validation* can address both of these issues. It relies on the simple observation model (eq. 2.6), and is most effective if the observations $\boldsymbol{Y}$ are all the same type, and n, the number of observations, is large.

There are several variants on cross validation, of which I describe the simplest, *leave-one-out (LOO)*.[9] Recollect the notation of Sec. 2.8, and let

$$\mu^*_{mj} := \begin{cases} \mathrm{E}_m\{Y_j \mid \boldsymbol{Y}_{-j} \doteq \boldsymbol{y}_{-j}\} & \text{Bayesian} \\ \mathrm{E}_m\{Y_j \mid \boldsymbol{Y}_{-j} \doteq \boldsymbol{y}_{-j}; \hat{\theta}_{m,-j}(\boldsymbol{y})\} & \text{Frequentist} \end{cases}$$

where in the first case the model parameter θ_m has been integrated out, and in the second it has been plugged in. In both cases, μ^*_{mj} is model m's expectation of Y_j based on all of the other observations. As explained in Sec. 3.6, μ^*_{mj} is model m's point prediction of Y_j under quadratic loss. The LOO *mean squared error (MSE)* of model m is defined as

$$\mathrm{MSE}_m := \frac{1}{n} \sum_{j=1}^{n} \big(y_j - \mu^*_{mj}\big)^2.$$

If n is large, then the MSE approximates the expected quadratic loss for predicting a new Y like those in $\boldsymbol{Y}$. Take the square root to map this into the same scale as the Y's, giving the *RMSE*,

$$\mathrm{RMSE}_m := \sqrt{\mathrm{MSE}_m}.$$

Hence we can evaluate the accuracy of model m in terms of its RMSE, and we favour models with smaller RSME's.

When compared to methods based on the evidence or on Information Criteria, cross-validation is n times as expensive, as well as being more restrictive regarding the datasets for which it can be applied. But it addresses two widespread difficulties with those methods, is intuitive, and provides a simple measure of model accuracy.

[9] See Gelman *et al.* (2014, ch. 7) for a discussion of cross validation in the context of model choice.

6.3 Hypothesis testing

Hypothesis testing is a special case of model choice, in which all of the models for $\boldsymbol{X}$ are contained within the same parametric family. Usually the choice is between two distinct subsets of the parameter space. Thus we start with the family of distributions

$$f_{\boldsymbol{X}}(\cdot;\theta) \quad \theta \in \Omega$$

and then specify the competing subsets as

$$\begin{aligned} H_0 &: \theta \in \Omega_0 \\ H_1 &: \theta \in \Omega_1 \end{aligned}$$

where $\Omega_0 \cap \Omega_1 = \varnothing$ and, ideally, Ω_0 and Ω_1 are well-separated. H_0 and H_1 are termed *hypotheses*. If a hypothesis contains only a single element of Ω it is termed a *simple hypothesis*, otherwise it is a *composite hypothesis*. We adopt the simple observation model (eq. 2.6). The objective of hypothesis testing is to choose in favour of H_0 or H_1 using $f_{\boldsymbol{Y}}$ and $\boldsymbol{y}$, and to quantify the strength of evidence on which that choice is based.

The most studied hypothesis test, and the one for which the strongest theoretical results are available, is the case of two simple hypotheses, written

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1. \end{aligned}$$

In this case there is agreement among statisticians that the appropriate test is based on the *likelihood ratio*,

$$B_{01}(\boldsymbol{y}) := \frac{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta_0)}{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta_1)},$$

and the test must have the form

$$B_{01}(\boldsymbol{y}) \begin{cases} < k_1 & \text{choose } H_1 \\ \text{in the middle} & \text{undecided} \\ > k_2 & \text{choose } H_0 \end{cases} \tag{6.1}$$

for some values $0 < k_1 \leq k_2$. If the client dislikes 'undecided' then k_1 and k_2 will need to be close together, maybe even the same value. But the client should understand that 'undecided' is a perfectly acceptable category for evidential support, and that suppressing it can lead to wrong choices.[10]

Appropriate values for k_1 and k_2 are tricky to decide—they ought to depend on the consequences of each choice, but in hypothesis testing we are discouraged from taking explicit account of the client's action set and loss function.

For the Bayesian, the following equality is a direct consequence of Bayes's theorem in odds form (after Thm 1.26):

$$\frac{\Pr^*(\theta \doteq \theta_0)}{\Pr^*(\theta \doteq \theta_1)} = B_{01}(\boldsymbol{y}) \, \frac{\Pr(\theta \doteq \theta_0)}{\Pr(\theta \doteq \theta_1)} \tag{6.2}$$

[10] In Scots law, for example, the judge or jury can return a verdict of 'not proven', lying somewhere between 'proven' and 'not guilty'.

where $\Pr^*$ is the probability conditional on $\boldsymbol{Y} \doteq \boldsymbol{y}$, as usual; B_{01} in this context is termed the *Bayes factor* for H_0 versus H_1. This equality is also expressed as the mantra

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

where 'odds' is a generic term for a ratio of probabilities. So the Bayesian concerned with selecting the hypothesis with the highest posterior probability needs only to be satisfied that the Bayes factor outweighs the prior odds. Unless he starts with strong beliefs about θ, the Bayesian will find a Bayes factor of 20 or greater fairly compelling evidence for favouring H_0. Jeffreys (1961, Appendix B) termed this 'strong' evidence in favour of H_0.[11] So the Bayesian might well have $k_1 \leftarrow 1/20$ and $k_2 \leftarrow 20$, or something similar.

Things are more complicated for the Frequentist, because he will not admit a prior PMF for the parameter. The traditional approach to this problem is the *Neyman-Pearson* approach, in which the sampling distributions of the random quantity $B_{01}(\boldsymbol{Y})$ under H_0 and H_1 are used to set k_1, with $k_2 \leftarrow k_1$. I am not going to describe the Neyman-Pearson approach, because I believe it is obsolete.[12] Instead, I will give '*Barnard's rule*'. I base this on a suggestion of George Barnard, in the discussion of Lindley (2000). It is necessary to suspend our disbelief and to assume that exactly one of H_0 or H_1 is 'true'.[13]

Theorem 6.1 (Barnard's rule)**.** *Suppose that exactly one of H_0 or H_1 is 'true'. Define an incorrect choice as choosing H_0 when H_1 is 'true', or choosing H_1 when H_0 is 'true'. Then the probability of an incorrect choice when using* (6.1) *with $k_1 \leftarrow 1/c$ and $k_2 \leftarrow c$ is never more than* $1/c$.

Proof. This is a double application of Markov's inequality (Thm 1.4). Suppose that H_0 is 'true'. Then, in an obvious notation,

$$\Pr_0\{\text{incorrect}\} = \Pr_0\{B_{01}(\boldsymbol{Y}) \le 1/c\} = \Pr_0\{1/B_{01}(\boldsymbol{Y}) \ge c\} \le 1/c$$

by Markov's inequality, because $1/B_{01}(\boldsymbol{Y})$ has expectation 1 under H_0. On the other hand, suppose that H_1 is 'true'. Then

$$\Pr_1\{\text{incorrect}\} = \Pr_1\{B_{01}(\boldsymbol{Y}) \ge c\} \le 1/c$$

by the same reasoning. Since the probability of an incorrect choice is $\le 1/c$ under both H_0 and H_1, it must be never more than $1/c$, because of the condition that exactly one of H_0 or H_1 is 'true'. □

With Barnard's rule, the client states, "I don't like being wrong, so I am going to set $c \leftarrow 20$ whenever I have to make a choice between two mutually exclusive simple hypotheses." This implies that the client will not be wrong more than 5% of the time.[14] Because Markov's inequality is generous (i.e. seldom tight) the client can be fairly sure that her actual probability is a lot less than 0.05, and so she might be happier with a more relaxed value for c, say $c \leftarrow 10$. In many applications, she will find that the observations leave her undecided: that's just the way it is—sometimes the evidence in $\boldsymbol{y}$ is not very compelling.

[11] See also the scale given in Kass and Raftery (1995).

[12] Casella and Berger (2002, ch. 8) and then Lehmann and Romano (2005, ch. 3) are good references.

[13] The same suspension of disbelief is necessary in the Neyman-Pearson approach. Nowhere else in these notes do we need to make this totally bogus assertion. I am compelled to put 'true' in scare quotes.

[14] The comments on confidence intervals at the end of Sec. 2.A.2 are appropriate here. It is not obvious that the client wants to control her lifetime probability of being wrong, rather than her probability of being wrong in this particular application.

Composite hypotheses. Composite hypotheses provide no additional challenges for the Bayesian, who simply sums over subsets of the posterior distribution to compute $\Pr^*(\theta \in \Omega_0) / \Pr^*(\theta \in \Omega_1)$.

Composite hypotheses are a major challenge for the Frequentist, except for one special case which is massively overused in both teaching and research (the Normal distribution). The general theory is a veritable blizzard of 'adhockery'.[15] You can take my work for this, or you can read chapters 3–10 of Lehmann and Romano (2005, pp. 56–415) and decide for yourself.

6.4 Significance levels (P-values)

Significance levels were the invention of the great statistician R.A. Fisher. Savage (1976) and Efron (1998) give fascinating commentaries on Fisher's statistical work. Fisher was also a founder of the modern theory of evolution, and nominated by Richard Dawkins as the greatest biologist since Darwin.[16]

6.4.1 Motivation and definition

The distinguishing feature of a significance level is the absence of an explicit alternative hypothesis. In other words, a significance level attaches a score to

$$H_0 : \theta \in \Omega_0$$

directly. Initially, consider simple hypotheses of the form $\Omega_0 \leftarrow \{\theta_0\}$; composite hypotheses will be covered in Sec. 6.4.4. Thus H_0 corresponds to

$$H_0 : \boldsymbol{Y} \sim f_{\boldsymbol{Y}}(\cdot\,; \theta_0)$$

which I write as $\boldsymbol{Y} \sim f_0$, where f_0 is the *null distribution*. This type of H_0 simply describes a PMF for $\boldsymbol{Y}$. There is no particular reason for it to be one member of a parametric family: it could equally well be the $\boldsymbol{Y}$-margin of a fully-probabilistic $f_{\boldsymbol{X}}$ under the simple observation model (eq. 2.6), although this is less common in practice (see Box, 1980, and the discussion and rejoinder).

There is some interesting theory about how to score an observation $\boldsymbol{y}$ with respect to a distribution f_0. Any score can be written as $s(f_0, \boldsymbol{y})$, which I will take in the positive orientation, so that larger scores indicate a better match. A sensible constraint on scoring rules is that they are *proper*:

$$\mathrm{E}_0\{s(f_0, \boldsymbol{Y})\} \geq \mathrm{E}_0\{s(f', \boldsymbol{Y})\} \qquad \text{for all } f',$$

where E_0 is the expectation with respect to f_0.[17] Among the proper scoring rules, perhaps the simplest is the *logarithmic scoring rule*

$$s(f_0, \boldsymbol{y}) \leftarrow \log f_0(\boldsymbol{y}).$$

It is easy to prove that this scoring rule is proper, using *Gibbs's inequality*.[18] Bernardo and Smith (2000, sec. 2.7) make a case for

[15] 'Adhockery' has gone mainstream, but I believe it was coined by Bruno de Finetti, although I have not tracked his first usage yet.

[16] `http://edge.org/conversation/who-is-the-greatest-biologist-of-all-time`

[17] See Gneiting and Raftery (2007) for more details about proper scoring rules.

[18] Which states that $\sum_i p_i \log(p_i/q_i) \geq 0$ for probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, with equality if and only if $\boldsymbol{p} = \boldsymbol{q}$. Follows immediately from $\log(x) \leq x - 1$ with equality if and only if $x = 1$. And this latter result follows from $\log(1) = 0$ and $\log(\cdot)$ strictly concave.

favouring the logarithmic scoring rule, on the basis that it uniquely satisfies properties of smoothness and locality.

Proper scoring rules are useful for comparing two PMFs for $\boldsymbol{y}$. If

$$s(f_0, \boldsymbol{y}) - s(f_1, \boldsymbol{y}) > 0$$

then the evidence in $\boldsymbol{y}$ favours f_0 over the alternative f_1. This conclusion would be much less compelling if the scoring rule was not proper.[19] But the score $s(f_0, \boldsymbol{y})$ on its own is much harder to interpret. Is a value such as $s(f_0, \boldsymbol{y}) = -23.982635$ large or small? Even with a logarithmic scoring rule, finding that

$$\log f_0(\boldsymbol{y}) \approx -13.9$$

is not very helpful. We infer that $f_0(\boldsymbol{y}) \approx 0.000000919$, but this small value might just reflect a huge $\mathcal{Y}$, in which any value for $\boldsymbol{y}$ has only a small probability of occurring. For example, I toss a coin 20 times and get

$$H, H, T, T, T, H, H, H, T, H, T, T, T, H, T, H, T, H, T, H.$$

The probability of this outcome under the hypothesis

$$H_0 : \text{the coin tosses are independent and fair}$$

is $2^{-20} = 0.000000954$, but clearly the smallness of this value on its own cannot convince me that H_0 is false, since every outcome has the same probability under H_0.[20]

This is the basic problem that the significance level seeks to address: to construct a score $s(f_0, \boldsymbol{y})$ which occupies a meaningful scale, so that we can identify small values which cause us to question the model f_0 as an appropriate representation for $\boldsymbol{Y}$. The P-value is the result.[21] In the following definition, the scalar random quantity X has a *subuniform distribution* exactly when

$$\Pr(X \stackrel{.}{\leq} u) \leq u \quad \text{for all} \quad 0 \leq u \leq 1.$$

The uniform distribution is a special case of the subuniform distribution, with $\Pr(X \stackrel{.}{\leq} u) = u$.

Definition 6.1 (P-value)**.** *The statistic $p_0 : \mathcal{Y} \to [0, 1]$ is a P-value for the simple hypothesis H_0 exactly when $p_0(\boldsymbol{Y})$ has a subuniform distribution under H_0.*

In this definition, p_0 appears to be a function of $\boldsymbol{y}$ alone, but the construction of p_0 must involve f_0 as well, in order to ensure that the subuniformity property holds: hence the inclusion of the '0' subscript. A subuniform distribution is a weaker condition that uniform distribution, but without it there would not be P-values for random quantities with finite or countably-infinite realms (Sec. 6.4.3), P-values computed by sampling (also Sec. 6.4.3), or P-values for composite hypotheses (Sec. 6.4.4).

[19] With a logarithmic scoring rule, the difference in the scores is greater than zero if and only if the Bayes factor B_{01} is greater than one.

[20] This is something that Fisher got wrong. See, for example, Hacking's critique (Hacking, 1965, p. 80) of Fisher's exposition of significance levels in his final book (Fisher, 1956).

[21] Here I am presenting a modern definition of a P-value, not the one advanced by Fisher.

6.4.2 *Difficulties with interpretation*

The basic idea with a P-value is that a value of $p_0(\boldsymbol{y})$ close to zero indicates an event in the lefthand tail of the distribution that would be implied by the truth of H_0. For example, since

$$\Pr_0\{p_0(\boldsymbol{Y}) \dot{\le} 0.005\} \le 0.005$$

we conclude that the outcome $p_0(\boldsymbol{y}) = 0.005$ is no larger than the 0.5th percentile of the distribution of $p_0(\boldsymbol{Y})$ under H_0. An outcome this far into the tail of a distribution is unusual, and leads us to consider whether in fact H_0 is 'true', or even adequate. But this apparently simply story is full of subtlety.

Sub-uniformity. The subuniformity of P-values limits the conclusions that we can draw. Suppose that $p_0(\boldsymbol{Y})$ were uniform under H_0, rather than subuniform. In this case if $p_0(\boldsymbol{y}) = 0.005$ we could conclude that we were in the tail of the distribution of $p_0(\boldsymbol{Y})$ under H_0, while if $p_0(\boldsymbol{y}) = 0.35$ we could conclude that we were near the middle of the distribution. But with subuniformity we can no longer conclude the latter, because $\Pr_0\{p_0(\boldsymbol{Y}) \dot{\le} 0.35\}$ is no longer equal to 0.35, but only no larger than 0.35. So subuniformity prevents us from interpreting middling P-values as indicating we are near the centre of the distribution of $p_0(\boldsymbol{Y})$ under H_0. In fact, with $p_0(\boldsymbol{y}) = 0.35$ we may actually be in the lefthand tail, but not know it. Similarly, $p_0(\boldsymbol{y}) = 0.09$ looks a bit improbable under H_0, being in the lefthand tail, but with subuniformity we do not know whether we are a little into the lefthand tail, or way into the lefthand tail.

So wherever possible, we construct P-values which are uniform or nearly uniform under H_0, rather than subuniform. Sec. 6.4.4 shows that a necessary (but not sufficient) condition is that the realm of $\boldsymbol{Y}$ is large; ideally uncountably infinite.

The 'truth' of H_0. Computing P-values is, from the outset, a forlorn exercise, because we know that f_0 is not the 'true' PMF for $\boldsymbol{Y}$. It is a representation of my beliefs about $\boldsymbol{Y}$. The randomness I perceive in $\boldsymbol{Y}$ is a symptom of my lack of knowledge. So I should respond to a small P-value without any surprise: what did I expect, that nature herself would choose $\boldsymbol{Y} \to \boldsymbol{y}$ according to *my* PMF?! Likewise, even in the uniform case a large P-value does not indicate that H_0 is 'true': it simply shows that $\boldsymbol{Y}$ is not very informative, since it has failed to alert me to a claim which I know to be false.

It is tempting to try to finesse this difficulty by focusing on adequacy rather than 'truth'. But this does not work. If n, the number of observations is small, then the P-value will be large because $\boldsymbol{Y}$ is not very informative. But if n is large, then the P-value will be small because H_0 is not 'true'. So one might argue that in the middle there is a 'sweet spot' for n for which the P-value is informative about the adequacy of H_0 as a model for $\boldsymbol{Y}$. But there is no logic for this claim. If something is not useful for small n or for large n, there is no basis to claim that it will nevertheless be useful for middling n.[22] We should call this the *no sweet spot* argument.

Many P-values. For any H_0, there is an infinity of P-values (see Sec. 6.4.3). In fact, one can find a P-value to take any value in $(0, 1)$, for any $\boldsymbol{y}$. So if you have a P-value $p_0(\boldsymbol{y}) = 0.005$, which you think is quite interesting, I can counter with a $p_0'(\boldsymbol{y}) = 0.35$, which is not very interesting at all. Who is right?

Here is a recipe for a completely meaningless P-value. Use $\boldsymbol{y}$ to seed a uniform random number generator $u_1, u_2, \dots$, and let $p_0(\boldsymbol{y})$ be the value of the trillionth term in the sequence. By any reasonable criterion this value has a uniform distribution, and hence is a valid P-value according to Def. 6.1.

DeGroot (1973) identified an additional condition which was necessary for p_0 to be a sensible P-value: $p_0(\boldsymbol{Y})$ under H_0 has to *stochastically dominate* $p_0(\boldsymbol{Y})$ under a decision-relevant alternative to H_0. A random quantity X stochastically dominates Y exactly when

$$\begin{aligned} &\Pr(X \leq v) \leq \Pr(Y \leq v) \text{ for all } v, \\ \text{and } &\Pr(X \leq v) < \Pr(Y \leq v) \text{ for some } v. \end{aligned}$$

The stochastic dominance property implies that the distribution of $p_0(\boldsymbol{Y})$ is pushed to the left under a decision-relevant alternative to H_0, and in this sense small P-values favour the alternative over H_0.

The stochastic dominance condition eliminates P-values like the uniform random number generator, because its distribution is the same under all models. But the condition reintroduces a competitor model through the back door—the decision-relevant alternative to H_0—and thus compromises the 'purity' of the P-value as an assessment of H_0 alone. If users of P-values want to demonstrate that their particular choice of p_0 is an appropriate one, then they need to establish that it has both the subuniformity property for H_0 *and* the stochastic dominance property for some alternative to H_0. But, having produced an alternative, these users might as well be doing a hypothesis test.

Not a hypothesis test. Many people confuse P-values and hypothesis tests. A typical symptom is to compute a $p_0(\boldsymbol{y})$ less than 0.05 and then report that "the null hypothesis is rejected at a Type 1 error of 5%." The characteristic of a hypothesis test is that two hypotheses compete, and *both* of them get tested by the observations $\boldsymbol{y}$. The result is a test statistic $B_{01}(\boldsymbol{y})$ which, to the Bayesian at least, is directly meaningful, and which is used by statisticians of all tribes to construct a rule for choosing between H_0 and H_1, or remaining undecided. This does not happen with a P-value, and there is no sense in using a P-value to "reject" H_0 if it cannot be demonstrated that an alternative H_1 is better. See Goodman (1999a,b).

Not a proper scoring rule either. The function p_0 takes both f_0 and $\boldsymbol{y}$ as its arguments (as will be seen explicitly in Sec. 6.4.3), and

[22] For a similar observation ("the P-value can be viewed as a crude measure of sample size"), see this blog entry, `http://andrewgelman.com/2009/06/18/the_sample_size/`, and the following comments.

therefore it is a scoring rule. But it is not a proper scoring rule (see Sec. 6.4.1), and therefore comparisons between P-values of different models is not a good way to choose between models.

* * *

There is a huge literature on why P-values do not do what people would like them to do; start at Greenland and Poole (2013) and work backwards. There is also, unfortunately, quite a lot of evidence that it is easy to cheat with P-values, and that people do cheat with P-values; see, for example, Simmons *et al.* (2011) and Masicampo and Lalande (2012).[23] See the discussion on level error at the end of Sec. 6.5 for further comments.

6.4.3 *Constructing and computing P-values*

I stated above that there is an infinity of P-values for any H_0. This subsection presents a recipe for making them. But first, a very useful general result.

Theorem 6.2 (Probability Integral Transform, PIT)**.**
Let $X \in \mathcal{X} \subset \mathbb{R}$ be a scalar random quantity with distribution function $F_X(x) := \Pr(X \leq x)$, and let $Y := F_X(X)$. Then Y has a sub-uniform distribution, *and $F_Y(u) = u$ if there exists an $x \in \mathcal{X}$ such that $u = F_X(x)$.*

Proof. First, consider the case where $u = F_X(x)$ for some $x \in \mathcal{X}$:

$$F_Y(u) = \Pr\{F_X(X) \leq F_X(x)\} = \Pr\{X \leq x\} = F_X(x) = u.$$

The 'cancellation' of F at the second equality occurs because of the bijective relationship between x and $F(x)$ for $x \in \mathcal{X}$.[24] This proves the second part of the claim.

Otherwise, let x and x' be two consecutive values in $\mathcal{X}$, with $u = F_X(x)$ and $u' = F_X(x')$, and let $u + \delta$ be some value in the open interval (u, u'). Then

$$Y \leq u + \delta \implies X \leq x$$

and so $F_Y(u + \delta) \leq F_X(x) = u$. But we must also have $F_Y(u + \delta) \geq F_Y(u) = u$. Therefore we conclude that $F_Y(u + \delta) = u$, and hence $F_Y(u + \delta) < u + \delta$. □

So the distribution function of Y looks like a staircase where each step starts from the 45° line drawn from $(0, 0)$ to $(1, 1)$; see Figure 6.1. If X is a 'continuous' random quantity then the steps will remain infinitesimally close to the 45° line, and $F_X(X)$ will be uniform. Otherwise, and this includes random quantities with countably infinite support like the Poisson in Figure 6.1, the steps can diverge substantially from the 45° line and $F_X(X)$ can be severely subuniform.

Now here is the recipe for making a P-value.

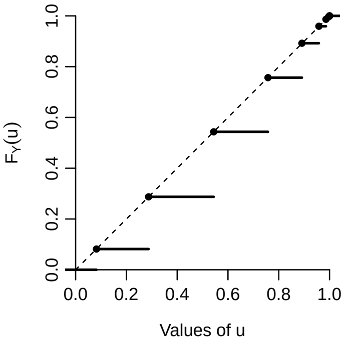

Figure 6.1: Distribution function of $Y := F_X(X)$, where $X \sim \text{Poisson}(\lambda = 2.5)$.

[23] Simmons *et al.* coin the coy euphemism 'researcher degrees of freedom' to describe 'flexibility in data collection, analysis, and reporting'; i.e. ways that researchers can get their P-values lower without 'cheating'. This practice is prevalent enough to have acquired the unsavoury name of 'P-hacking'.

[24] Technical note: here we can ignore points in $\mathcal{X}$ that have zero probability.

Theorem 6.3. *Let $t : \mathcal{Y} \to \mathbb{R}$ be a statistic. Then*

$$p_0(\boldsymbol{y}) := \Pr_0\{t(\boldsymbol{Y}) \geq t(\boldsymbol{y})\}$$

is a P-value satisfying Def. 6.1.

Proof. I use a nifty trick from Casella and Berger (2002, section 8.3.4). Define $T := t(\boldsymbol{Y})$. Let G_0 be the distribution function of $-T$ under H_0. Then

$$p_0(\boldsymbol{y}) = \Pr_0\{T \geq t(\boldsymbol{y})\} = \Pr_0\{-T \leq -t(\boldsymbol{y})\} = G_0(-t(\boldsymbol{y})).$$

Then since $p_0(\boldsymbol{Y}) = G_0(-T)$, subuniformity of $p_0(\boldsymbol{Y})$ under H_0 follows from the PIT (Thm 6.2). □

Hence there is a P-value for every test statistic, and there is an infinity of test statistics. Here is another dodgy P-value: $t(\boldsymbol{y}) = c$ (any constant will do). This does indeed have a subuniform distribution under H_0, with

$$\Pr_0\{p_0(\boldsymbol{Y}) \dot{\leq} 1\} = 1 \quad \text{and} \quad \Pr_0\{p_0(\boldsymbol{Y}) \dot{\leq} u\} = 0 \text{ for } u < 1.$$

What a useless P-value! This makes the point that of the infinity of possible P-values for H_0, many of them will be useless, or nearly so. Clearly, T needs to have a large support under H_0, in order that $p_0(\boldsymbol{Y})$ is even approximately uniform under H_0. But recollect Figure 6.1, which showed that a countably infinite support was not big enough.

* * *

Occasionally it will be possible to choose a test statistic $t(\cdot)$ with a known distribution under H_0, from which an explicit p_0 can be derived.[25] But this puts the cart before the horse—we want to choose our test statistic to reflect our application; in particular, we would like the resulting P-value to satisfy the stochastic dominance property discussed in Sec. 6.4.2. Happily, a P-value for any $t(\cdot)$ can be computed by simulation using following result, which uses exchangeability (Chapter 4).

Theorem 6.4. *For any finite sequence of scalar random quantities $X^0, X^1, \dots, X^m$, define the rank of X^0 in the sequence as*

$$R := \sum_{i=1}^{m} \mathbb{1}_{X^i \dot{\leq} X^0}.$$

If $X^0, X^1, \dots, X^m$ are exchangeable then R has a uniform distribution on the integers $0, 1, \dots, m$, and $(R+1)/(m+1)$ has a subuniform distribution.

Proof. By exchangeability, X^0 has the same probability of having rank r as any of the other X's, for any r, and therefore

$$\Pr(R = r) = \frac{1}{m+1} \quad \text{for } r = 0, 1, \dots, m \tag{†}$$

and zero otherwise, proving the first claim.

[25] Asymptotic results are useful here; see Cox (2006, ch. 6). These give approximate P-values, in which the distribution of $p_0(\boldsymbol{Y})$ is approximately uniform under H_0. There is a level error problem with these P-values, just as in confidence intervals; see Sec. 2.A.2.

To prove the second claim,[26]

$$
\begin{aligned}
\Pr\left\{\frac{R+1}{m+1} \leq u\right\} &= \Pr\left\{R+1 \leq u(m+1)\right\} \\
&= \Pr\left\{R+1 \leq \lfloor u(m+1) \rfloor\right\} \quad \text{as } R \text{ is an integer} \\
&= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \Pr(R = r) \\
&= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \frac{1}{m+1} \qquad \text{from } (\dagger) \\
&= \frac{\lfloor u(m+1) \rfloor}{m+1} \leq u,
\end{aligned}
$$

as required. □

Now take a statistic $t : \mathcal{Y} \to \mathbb{R}$ which has the property that larger values of $t(\boldsymbol{y})$ are suggestive of a decision-relevant alternative from H_0. Define $T := t(\boldsymbol{Y})$ and $T^j := t(\boldsymbol{Y}^j)$ where $\boldsymbol{Y}^1, \dots, \boldsymbol{Y}^m \overset{\text{iid}}{\sim} f_0$. Then $T, T^1, \dots, T^m$ form an exchangeable sequence under H_0. Hence if

$$R(\boldsymbol{y}) := \sum_{j=1}^{m} \mathbb{1}_{-T^j \dot{\leq} -t(\boldsymbol{y})} = \sum_{j=1}^{m} \mathbb{1}_{T^j \dot{\geq} t(\boldsymbol{y})}$$

then Thm 6.4 implies that

$$P(\boldsymbol{y}) := \frac{R(\boldsymbol{y}) + 1}{m+1}$$

has a subuniform distribution under H_0.[27] Furthermore, the Weak Law of Large Numbers (see, e.g. Grimmett and Stirzaker, 2001, sec. 5.10) shows that

$$\lim_{m \to \infty} P(\boldsymbol{y}) = \frac{\lim_m m^{-1}\big(R(\boldsymbol{y}) + 1\big)}{\lim_m m^{-1}(m+1)} = \mathrm{E}_0\{\mathbb{1}_{T \dot{\geq} t(\boldsymbol{y})}\} = \Pr_0\{T \dot{\geq} t(\boldsymbol{y})\}$$

and so the asymptotic limit of $P(\boldsymbol{y})$ is the P-value defined in Thm 6.3.

$P(\boldsymbol{y})$ is subuniform for all m, but it is approximately uniform for large m, because in this case

$$\frac{\lfloor u(m+1) \rfloor}{m+1} \approx u.$$

So a bigger m is preferred, because a more uniform distribution under H_0 is more informative, as discussed in Sec. 6.4.2.

In cases where it is not straightforward to simulate independent realisations from f_0, the value of $P(\boldsymbol{y})$ can be computed from an MCMC sequence from f_0. In order for the $\boldsymbol{Y}^j$'s to be exchangeable it is sufficient that they are independent, and hence these m values must be extracted from well-separated locations in the sequence. Besag and Clifford (1989) described an elegant backwards-and-forwards implementation for MCMC sampling from f_0 which produces exchangeable but not IID $\boldsymbol{Y}$'s under H_0.

[26] Notation: $\lfloor x \rfloor$ is the largest integer no larger than x, termed the 'floor' of x.

[27] Here I write both R and P as capitals, because they are functions of the random quantities $\boldsymbol{Y}^1, \dots, \boldsymbol{Y}^m$.

6.4.4 *Composite hypotheses*

The definition in Def. 6.1 is for a simple null hypothesis. It can be extended to a composite hypothesis, written

$$H_0 : \theta \in \Omega_0$$

where the previous simple hypothesis was just the special case $\Omega_0 = \{\theta_0\}$. Composite hypotheses are common—more common than simple ones in fact. Nuisance parameters were mentioned in Sec. 2.5. Where interest is in a subset of the parameters, the other nuisance parameters remain unconstrained, and the hypothesis is composite. For example, if $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}$, then

$$H_0 : \mu = \mu_0$$

is a composite hypothesis with $\Omega_0 = \{\mu_0\} \times \mathbb{R}_{++}$.

With a composite hypothesis, $p(\cdot; \Omega_0)$ is a P-value for H_0 if it has a subuniform distribution under every element of Ω_0. This is easy to achieve, given a P-value for a simple hypothesis.

Theorem 6.5. *Let $H_0 : \theta \in \Omega_0$ be a composite hypothesis, and $p(\boldsymbol{y}; t)$ be a P-value for the simple hypothesis $\theta = t$. Then*

$$p(\boldsymbol{y}; \Omega_0) := \sup_{t \in \Omega_0} p(\boldsymbol{y}; t)$$

has a subuniform distribution for every $t \in \Omega_0$.

Proof. Follows from the fact that $p(\boldsymbol{y}; \Omega_0) \leq u$ implies $p(\boldsymbol{y}; t) \leq u$ for all $t \in \Omega_0$. Therefore

$$\Pr_t\{p(\boldsymbol{Y}; \Omega_0) \dot{\leq} u\} \leq \Pr_t\{p(\boldsymbol{Y}; t) \dot{\leq} u\} \leq u$$

for all $t \in \Omega_0$, where $\Pr_t$ is the probability under $\theta = t$. □

From this proof it is clear that $p(\boldsymbol{Y}; \Omega_0)$ can be extremely subuniform, even in the case where $p(\boldsymbol{Y}; t)$ is uniform for every $t \in \Omega_0$. As discussed in Sec. 6.4.2, subuniformity reduces the information in a P-value, and hence P-values for composite hypotheses are generally rather uninformative. Berger and Boos (1994) have a clever suggestion to address this, but it has not been taken up in practice.

A much more common approach, in the case of a single parameter of interest plus nuisance parameters, is to compute a confidence interval for the parameter, discussed in the next section.

6.5 *Confidence sets*

Confidence sets are a way of assessing uncertainty about the parameters without treating the parameters as random variables. Some general reflections on confidence sets are given in Sec. 2.A.2. I will say more about *level error* at the end of the section.

Definition 6.2 (Confidence set and coverage)**.** $\mathcal{C}_\beta$ *is a level* β confidence set *for* θ *exactly when* $\mathcal{C}_\beta(\boldsymbol{y}) \subset \Omega$ *and*

$$\Pr_t\left\{t \in \mathcal{C}_\beta(\boldsymbol{Y})\right\} \geq \beta \quad \textit{for all } t \in \Omega.$$

The probability on the lefthand side is defined as the coverage *of* $\mathcal{C}_\beta$ *at* t*. If the coverage is exactly* β *for all* t*, then the confidence set is 'exact'.*

There is a close relationship between confidence sets and P-values; for every P-value, there is a confidence set (and *vice versa*).[28] Thus reservations about P-values hold for confidence sets as well.

Theorem 6.6. *Let* $p(\boldsymbol{y}, t)$ *be a P-value for the hypothesis* $H_0 : \theta = t$*. Then*

$$\mathcal{C}_\beta(\boldsymbol{y}) := \left\{t \in \Omega : p(\boldsymbol{y}, t) > 1 - \beta\right\}$$

is a level β *confidence set for* θ*. If the P-value is exact, then the confidence set is exact as well.*

From this construction it is immediate that, for the same P-value, $\beta \leq \beta'$ implies that $\mathcal{C}_\beta(\boldsymbol{y}) \subset \mathcal{C}_{\beta'}(\boldsymbol{y})$, so that these confidence sets are always nested. While this property is not in the definition of a confidence set, anything else would seem bizarre.

Proof. This proof uses the subuniformity property of P-values.

$$\begin{aligned}\Pr_t\{t \in \mathcal{C}_\beta(\boldsymbol{Y})\} &= \Pr_t\{p(\boldsymbol{Y}, t) > 1 - \beta\} \\ &= 1 - \Pr_t\{p(\boldsymbol{Y}, t) \leq 1 - \beta\} \\ &\geq 1 - (1 - \beta) = \beta,\end{aligned}$$

where the inequality follows from the P-value being subuniform. In the case where the P-value is uniform, the inequality is replaced by an equality, and the confidence set is exact. □

A more general definition of confidence sets holds for any function of θ. If $g : \theta \mapsto \psi$, then $\mathcal{C}^\psi_\beta(\boldsymbol{y})$ is a level β confidence set for ψ exactly when

$$\Pr_t\left\{g(t) \in \mathcal{C}^\psi_\beta(\boldsymbol{Y})\right\} \geq \beta \quad \text{for all } t \in \Omega.$$

If ψ is one-dimensional and $\mathcal{C}^\psi_\beta(y)$ is convex for every $\boldsymbol{y}$, then $\mathcal{C}^\psi_\beta$ is termed a *confidence interval*.

These confidence sets can be contructed directly from a confidence set for θ.

Theorem 6.7 (Marginal confidence sets)**.**
If $\mathcal{C}_\beta$ *is a level* β *confidence set for* θ*, and* $g : \theta \mapsto \psi$*, then* $g\mathcal{C}_\beta$ *is a level* β *confidence set for* ψ*.*[29]

Proof. This follows immediately from

$$t \in \mathcal{C}_\beta(\boldsymbol{y}) \implies g(t) \in g\mathcal{C}_\beta(\boldsymbol{y})$$

for each $\boldsymbol{y}$. Hence

$$\beta \leq \Pr_t\left\{t \in \mathcal{C}_\beta(\boldsymbol{Y})\right\} \leq \Pr_t\left\{g(t) \in g\mathcal{C}_\beta(\boldsymbol{Y})\right\}$$

as required. □

[28] The *vice versa* is that if θ_0 is on the boundary of a level β confidence set, then $1 - \beta$ is a P-value for $H_0 : \theta = \theta_0$.

[29] If A is a set in $\mathcal{A}$, and g is a function with domain $\mathcal{A}$, then gA is the set $\{b : b = g(a) \text{ for some } a \in A\}$.

If g is one-to-one and $\mathcal{C}_\beta$ is an exact level β confidence set for θ, then the proof shows that $g\mathcal{C}_\beta$ is an exact level β confidence set for ψ. Otherwise, though, the coverage of $g\mathcal{C}_\beta$ might be much larger than β for all θ.

As mentioned at the end of Sec. 6.4.4, confidence intervals are an alternative to P-values for composite hypotheses involving nuisance parameters. The P-value for the composite hypothesis

$$H_0 : \mu = \mu_0$$

may be very subuniform, and effectively useless. The nominal 95% confidence set for all of the parameters can be marginalised to derive a nominal 95% confidence set for μ, according to Thm 6.7. If the resulting confidence interval does not contain μ_0 then the P-value for H_0 is less than 0.05, by Thm 6.6.

But there are merits to presenting H_0 in terms of a confidence interval for μ, rather than a P-value, because there is a difference in interpretation between a narrow confidence interval which just misses μ_0 and a wide interval that just includes it. In the former case, μ may be significantly different from μ_0, but not enough to worry about. In the latter case it μ has the potential to be very different from μ_0, even if it is not significantly different.

Level error and calibration. Level error is the different between the nominal coverage β and the actual coverage, both of which depend on θ. Level error typically arises when large-sample theory is used to construct a confidence set which is asymptotically exact. But, for finite n, the coverage is only approximately β, and can vary by θ.

Due to the duality of P-values and confidence sets, level error in a confidence set is equivalent to error in the calculation of a P-value. This error cannot be assumed to be in the direction of larger confidence sets and P-values, which would be consistent with their definitions. It may be that confidence sets are too small, and P-values likewise. For people who make the mistake of confusing P-values and hypothesis tests (see Sec. 6.4.2) this would lead to too many H_0's rejected. This was the basis of John Ioannidis's controversial paper 'Why most published research findings are false' (Ioannidis, 2005). He noted that ambitious scientists had an incentive not to correct this bias, because a small P-value increased their chance of publication in a prestigious journal.

If it is possible to sample cheaply from f_Y then there is a simple way to correct for level error, to first order, which is to adjust the nominal coverage β until the actual coverage at the MLE $\hat{\theta}(\boldsymbol{y})$ is equal to the desired coverage. This is *bootstrap calibration*, see DiCiccio and Efron (1996). I find it surprising that this sensible precaution is not better known (but perhaps I should be more cynical!). My advice would be not to trust a confidence set or P-value unless it has been calibrated to have the desired coverage at the MLE.

In related research the Observational Medical Outcomes Partnership (OMOP) have studied confidence intervals and P-values from observational studies based on medical databases (Schuemie *et al.*, 2014; Madigan *et al.*, 2014). The very alarming conclusion is

> Empirical calibration was found to reduce spurious results to the desired 5% level. Applying these adjustments to literature suggests that at least 54% of findings with $p < 0.05$ are not actually statistically significant and should be reevaluated. (Schuemie *et al.*, 2014, abstract)

Commentators are talking about a 'crisis of reproducibility' in science, and it looks as though uncorrected level error in confidence intervals and P-values is to blame.

7
Bibliography

D.J. Aldous, 1985. Exchangeability and related topics. In *Ecole d'Ete St Flour 1983*, pages 1–198. Springer Lecture Notes in Math. 1117. Available at `http://www.stat.berkeley.edu/~aldous/Papers/me22.pdf`. 79, 84

W.P. Aspinall and R.M. Cooke, 2013. Quantifying scientific uncertainty from expert judgment elicitation. In J.C. Rougier, R.S.J. Sparks, and L.J. Hill, editors, *Risk and Uncertainty Assessment for Natural Hazards*, chapter 4. Cambridge: Cambridge University Press, Cambridge, UK. 76

S. Banerjee, B.P. Carlin, and A.E. Gelfand, 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, Florida: Chapman & Hall/CRC. 104

D. Basu, 1975. Statistical information and likelihood. *Sankhyā*, **37**(1), 1–71. With discussion. 6

J. Berger and R. Wolpert, 1984. *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics, second edition. Available online, `http://projecteuclid.org/euclid.lnms/1215466210`. 6, 60

J.O. Berger and D.D. Boos, 1994. *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, **89**, 1012–1016. 127

J.M. Bernardo and A.F.M. Smith, 1994. *Bayesian Theory*. Chichester, UK: Wiley. 5

J.M. Bernardo and A.F.M. Smith, 2000. *Bayesian Theory*. John Wiley & Sons Ltd, Chichester, UK. (paperback edition, first published 1994). 75, 120

J. Besag, 1974. Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**(2), 192–236. 105, 107

J. Besag, 2004. Markov Chain Monte Carlo methods for statistical inference. Available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.320.9631`. 43

J. Besag and P. Clifford, 1989. Generalized Monte Carlo significance tests. *Biometrika*, **76**(4), 633–642. 126

P. Billingsley, 1995. *Probability and Measure*. John Wiley & Sons, Inc., New York NY, USA, third edition. 5

G.E.P. Box, 1980. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**(4), 383–430. With discussion. 120

A.R. Brazzale, A.C. Davison, and N. Reid, 2007. *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press, Cambridge, UK. 60

L.D. Brown, T.T. Cai, and A. DasGupta, 2001. Interval estimation for a binomial proportion. *Statistical Science*, **16**(2), 101–117. With discussion, pp 117–133. 55

G. Casella and R.L. Berger, 2002. *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 58, 119, 125

C. Chatfield, 2004. *The Analysis of Time Series*. Boca Raton, FL: Chapman & Hall/CRC. 105

R.M. Cooke, 1991. *Experts in Uncertainty; Opinion and Subjective Probability in Science*. New York & Oxford: Oxford University Press. 112

R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter, 1999. *Probabilistic Networks and Expert Systems*. New York: Springer. 102

M. Cowles and C. Davis, 1982. On the origins of the .05 level of statistical significance. *American Psychologist*, **37**(5), 553–558. 59

D.R. Cox, 1990. Role of models in statistical analysis. *Statistical Science*, **5**(2), 169–174. 52

D.R. Cox, 2006. *Principles of Statistical Inference*. Oxford University Press. 6, 58, 125

D.R. Cox and D.V. Hinkley, 1974. *Theoretical Statistics*. London: Chapman and Hall. 47

N. Cressie and C.K. Wikle, 2011. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Inc., Hoboken NJ, USA. 104, 105

A.C. Davison, D.V. Hinkley, and G.A. Young, 2003. Recent developments in bootstrap methodology. *Statistical Science*, **18**(2), 141–157. 47, 60

A.P. Dawid, 2002. Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**(2), 161–190. Corrigenda vol. 70, p. 437. 96

A.P. Dawid. Beware of the DAG! In *JMLR Workshop & Conference Proceedings*, volume 6, pages 59–86, 2010. 96

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). 5, 79

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. 133

B. de Finetti, 1972. *Probability, Induction and Statistics*. London: John Wiley & Sons. 5, 133

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

B. de Finetti, 1974/75. *Theory of Probability*. London: Wiley. Two volumes (2nd vol. 1975); A.F.M. Smith and A. Machi (trs.). 5, 10

B. de Finetti and L.J. Savage, 1962. Sul modo di scegliere le probabilità iniziali. *Biblioteca del Metron, Series C: Note E Commenti*, **1**, 81–154. English summary in ch. 8 of de Finetti (1972). Article and summary reprinted in Ericson *et al.* (1981). 39

M.H. DeGroot, 1970. *Optimal Statistical Decisions*. New York: McGraw-Hill, Inc. 63, 66, 71

M.H. DeGroot, 1973. Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, **68**, 966–969. 123

M.H. DeGroot and M.J. Schervish, 2002. *Probability and Statistics*. Addison-Wesley Publishing Co., Reading MA, 3rd edition. 62

P. Diaconis, 1988. Recent progress on de Finetti's notions of exchangeability. In *Bayesian Statistics 3*, pages 111–125. Oxford University Press, Oxford, UK. With discussion and rejoinder. 91

T.J. DiCiccio and B. Efron, 1996. Bootstrap confidence intervals. *Statistical Science*, **11**(3), 189–212. with discussion and rejoinder, 212–228. 60, 129

L.E. Dubins and L.J. Savage, 1965. *How to Gamble if you Must: Inequalities for Stochastic Processes*. McGraw-Hill. 11

W. Edwards, H. Lindman, and L.J. Savage, 1963. Bayesian statistical inference for psychological research. *Psychological Review*, **70**(3), 193–242. 49, 51

B. Efron, 1982. *The Jackknife, the Bootstrap and Other Resampling Methods*. Society for Industrial and Applied Mathematics, Philadelphia PA, USA. 56

B. Efron, 1998. R.A. Fisher in the 21st century. *Statistical Science*, **13** (2), 95–114. With discussion, 114–122. 120

B. Efron and G Gong, 1983. A leisurely look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, **37**(1), 36–48. 56

B. Efron and C. Stein, 1981. The Jackknife estimate of variance. *The Annals of Statistics*, **9**(3), 586–596. 56

J. Ellenberg, 2014. *How Not to be Wrong: The Hidden Mathematics of Everyday Life*. Allen Lane, Penguin Books Ltd, London, UK. 10

W.A. Ericson *et al.*, editors, 1981. *The Writings of Leonard Jimmie Savage: A Memorial Selection*. The American Statistical Association and The Institute of Mathematical Statistics. 133, 137

W. Feller, 1968. *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, New York NY, USA, 3rd edition. 93

R.A. Fisher, 1956. *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd. 121

D. Freedman, 1977. A remark on the difference between sampling with and without replacement. *Journal of the American Statistical Association*, **72**, 681. 86

S. French, 2011. Aggregating expert judgement. *RACSAM Serie A*, **105**, 181–206. 112

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, 2014. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, USA, 3rd edition. 40, 103, 110, 116, 117

T. Gneiting and A.E. Raftery, 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378. 120

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. John Wiley & Sons, Chichester, UK. 5

S. Goodman, 1999a. Toward evidence-based medical statistics. 1: The *p*-value fallacy. *Annals of Internal Medicine*, **130**, 995–1004. 123

S. Goodman, 1999b. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, **130**, 1005–1013. 123

S. Goodman and S. Greenland, 2007. Why most published research findings are false: Problems in the analysis. *PLoS Medicine*, **4** (4), e168. A longer version of the paper is available at `http://www.bepress.com/jhubiostat/paper135`.

S. Greenland and C. Poole, 2013. Living with *P* values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, **24** (1), 62–68. With discussion and rejoinder, pp. 69–78. 124

G.R. Grimmett and D.R. Stirzaker, 2001. *Probability and Random Processes*. Oxford, UK: Oxford University Press, 3rd edition. 5, 24, 126

I. Hacking, 1965. *The Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK. 71, 121

I. Hacking, 2001. *An Introduction to Probability and Inductive Logic*. Cambridge, UK: Cambridge University Press. 60

D. Hilbert, 1926. Über das unendliche. *Mathematische Annalen (Berlin)*, **95**, 161–190. English translation in van Heijenoort (1967). 7

J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky, 1999. Bayesian model averaging: A tutorial. *Statistical Science*, **14**(4), 382–401. 113

C. Howson and P. Urbach, 2006. *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court Publishing Co., 3rd edition. 38

J.P.A. Ioannidis, 2005. Why most published research findings are false. *PLoS Medicine*, **2**(8), e124. See also Goodman and Greenland (2007) and Ioannidis (2007). 129

J.P.A. Ioannidis, 2007. Why most published research findings are false: Author's reply to Goodman and Greenland. *PLoS Medicine*, **4**(6), e215.

E.T. Jaynes, 2003. *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press. 38

R.C. Jeffrey, 2004. *Subjective Probability: The Real Thing*. Cambridge, UK: Cambridge University Press. Unfortunately this first printing contains quite a large number of typos. 38

H. Jeffreys, 1961. *Theory of Probability*. Oxford, UK: Oxford University Press, 3rd edition. 119

J.B. Kadane, 2011. *Principles of Uncertainty*. Chapman & Hall/CRC Press, Boca Raton FL, USA. 7

D. Kahneman, 2011. *Thinking, Fast and Slow*. Penguin Books Ltd, London, UK. 23, 80

R.E. Kass and A.E. Raftery, 1995. Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795. 116, 119

R.E. Kass and L. Wasserman, 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370. 49

H.J. Keisler, 2007. *Foundations of Infinitesimal Calculus*. Privately published. Available online, `https://www.math.wisc.edu/~keisler/foundations.pdf`. 35

J.F.C. Kingman, 1978. Uses of exchangeability. *The Annals of Probability*, **6**(2), 183–197. 79, 86

H. Kunreuther, G. Heal, M. Allen, O. Edenhofer, C.B. Field, and G. Yohe, 2013. Risk management and climate change. *Nature Climate Change*, **3**, 447–450. 71

F. Lad, 1996. *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 5, 6, 15

J. Ladyman, 2002. *Understanding Philosophy of Science*. Routledge, Abingdon, UK. 7, 52

S.L. Lauritzen, 1996. *Graphical Models*. Oxford: Clarendon. 107

E.L. Lehmann, 1990. Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science*, **5**(2), 160–168. 52

E.L. Lehmann and J.P. Romano, 2005. *Testing Statistical Hypotheses*. New York: Springer, 3rd edition. 119, 120

Z. Levin, N. Halfon, and P. Alpert, 2010. Reassessment of rain enhancement experiments and operations in Israel including synoptic considerations. *Atmospheric Research*, **97**, 513–525. 68

D.V. Lindley, 1980. L.J. Savage—his work in probability and statistics. *The Annals of Statistics*, **8**(1), 1–24. 49, 51

D.V. Lindley, 2000. The philosophy of statistics. *The Statistician*, **49**, 293–337. With discussion. 119

J.S. Liu, 2001. *Monte Carlo Strategies in Scientific Computing*. New York: Springer. 108

D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter, 2013. *The BUGS Book: A Practical introduction to Bayesian Analysis*. CRC Press, Boca Raton, FL, USA. 44, 103, 116

D. Madigan, P.E. Strang, J.A. Berlin, M. Schuemie, J.M. Overhage, M.A. Suchard, B. Dumouchel, A.G. Hartzema, and P.B. Ryan, 2014. A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, **1**, 11–39. 130

E.J. Masicampo and D.R. Lalande, 2012. A peculiar prevalence of *p* values just below .05. *The Quarterly Journal of Experimental Psychology*. 124

J.C. McWilliams, 2007. Irreducible imprecision in atmospheric and oceanic simulations. *Proceedings of the National Academy of Sciences*, **104**(21), 8709–8713. 109

K. Milner and J.C. Rougier, 2014. How to weigh a donkey in the Kenyan countryside. *Significance*, **11**(4), 40–43. 74, 115

J. Nocedal and S.J. Wright, 2006. *Numerical Optimization*. New York: Springer, 2nd edition. 22

A. O'Hagan, 2012. Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modellig & Software*, **36**, 35–48. 70

A. O'Hagan, C. E. Buck, A. Daneshkhah, J.E. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow, 2006. *Uncertain Judgements: Eliciting Expert Probabilities*. Chichester: Wiley. 76

C. O'Neil and R. Schutt, 2014. *Doing Data Science*. O'Reilly Media Inc., Sebastopol CA, USA. 40, 76

J.B. Paris, 1994. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge: Cambridge University Press. 38

J. Pearl, 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. 96

C.P. Robert, 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer. 5, 49

C.P. Robert and G. Casella, 2004. *Monte Carlo Statistical Methods*. Springer, New York NY, 2nd edition. 43

F.S. Roberts, 1984. *Measurement Theory with Applications to Decision making, Utility, and the Social Sciences*. Number 7 in the Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, UK. 64

J.C. Rougier and M. Goldstein, 2014. Climate simulators and climate projections. *Annual Review of Statistics and Its Application*, **1**, 103–123. 111

D.B. Rubin, 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**(4), 1151–1172. 60, 72, 110

H. Rue and L. Held, 2005. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton FL, USA. 104, 105, 108

L.J. Savage, 1951. The theory of statistical decision. *Journal of the American Statistical Association*, **46**, 55–67. In Ericson *et al.* (1981). 71

L.J. Savage, 1954. *The Foundations of Statistics*. Dover, New York, revised 1972 edition. 71, 72

L.J. Savage, 1960. Recent tendencies in the foundations of statistics. In *Proceedings of the 8th International Congress of Mathematicians*, pages 540–544. Cambridge University Press, Cambridge, UK. In Ericson *et al.* (1981). 8

L.J. Savage, 1976. On re-reading R.A. Fisher. *The Annals of Statistics*, **4**(3), 441–483. With discussion, 483–500. In Ericson *et al.* (1981). 120

M.J. Schervish, 1995. *Theory of Statistics*. New York: Springer. Corrected 2nd printing, 1997. 5, 79

M.J. Schuemie, P.B. Ryan, W. DuMouchel, M.A. Suchard, and D. Madigan, 2014. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*, **33**(209–218). 130

J.P. Simmons, L.D. Nelson, and U. Simonsohn, 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, **22**, 1359–1366. 124

J.Q. Smith, 2010. *Bayesian Decision Analysis: Principle and Practice*. Cambridge University Press, Cambridge, UK. 63, 69

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**(4), 583–616. With discussion, pp. 616–639. 116

M.C.M. Troffaes and G. de Cooman, 2014. *Lower Previsions*. John Wiley & Sons, Ltd, Chichester, UK. 21

A.W. van der Vaart, 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press. 58

B. van Fraassen, 1989. *Laws and Symmetry*. Oxford University Press. 38

J. van Heijenoort, editor, 1967. *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard Univeristy Press, Cambridge MA, USA. 135

N.Ya. Vilenkin, 1995. *In Search of Infinity*. Birkhäuser Boston, Cambridge MA, USA. English translation by Abe Shenitzer. Currently available online, `http://yakovenko.files.wordpress.com/2011/11/vilenkin1.pdf`. 7

D.F. Wallace, 2003. *Everything and More: A compact history of* ∞. W.W. Norton & Company, Inc., New York NY, USA. 7

P. Walley, 1991. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman & Hall. 21

L.A. Wasserman, 2004. *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer. 60

P. Whittle, 2000. *Probability via Expectation*. New York: Springer, 4th edition. 5

D. Williams, 1991. *Probability With Martingales*. Cambridge University Press, Cambridge, UK. 5