Jonathan Rougier

Department of Mathematics
University of Bristol

APTS lecture notes on Statistical Inference

Cambridge, December 2013

Contents

1
Expectation and Probability Theory

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

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

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

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

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

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

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

1.0 *Conventions and notation*

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

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

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

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

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

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

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

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

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

[1] Or an underscore in handwritten material.

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

1.1 Random quantities

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

random quantity

operational definition

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

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

realm

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

discrete random quantity

continuous random quantity

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

constant

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

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

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

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

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

1.2 *Expectations*

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

expectation

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

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

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

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

judgement

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

Theorem 1.1 (Immediate implications)**.**

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

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

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

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

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

Proof.

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

□

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

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

Schwarz's inequality

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

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

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

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

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

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

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

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

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

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

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

This is stronger because the triangle inequality implies that

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

1.3 Do all random quantities have finite expectations?*

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

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

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

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

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

[6] Think of the Central Limit Theorem.

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

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

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

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

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

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

This expectation should respect these self-evident rules:

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

Then a weakening of the additivity axiom in Definition 1.1 gives

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

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

integrable

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

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

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

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

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

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

Rearrange this to give

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

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

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

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

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

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

1.4 Probability

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

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

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

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

random proposition

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

and so on.

The probability of a random proposition A is defined as

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

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

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

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

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

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

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

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

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

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

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

1.5 *The 'Fundamental Theorem of Prevision'*

Fundamental Theorem of Prevision (FTP)

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

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

partition

First it is necessary to define a *partition*.

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

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

sufficiently fine partition

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

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

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

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

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

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

1. *E is a valid expectation operator.*

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

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

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

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

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

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

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

3. (Additivity)

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

as required.

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

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

as required.

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

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

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

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

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

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

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

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

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

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

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

as was to be shown. □

1.6 Conditional expectation

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

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

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

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

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

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

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

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

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

1.6.0 Preliminary concepts

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

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

square integrable

Square integrability has implications for other expectations as well.

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

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

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

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

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

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

not materially different

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

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

for all real-valued g and all Z.

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

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

mean-square equivalent

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

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

Proof.

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

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

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

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

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

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

1.6.1 Characterisation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hence

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

is a necessary condition for ψ to be a minimum.

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

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

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

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

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

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

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

1.6.2 Notation and definitions

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

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

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

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

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

Now for the definition of *conditional expectation*.

conditional expectation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.6.3 *Properties of the conditional expectation*

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

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

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

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

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

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

   Then, from (1.3),

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

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

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

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

   and the result follows immediately.

□

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

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

Law of Iterated Expectation (LIE)

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

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

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

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

Then three applications of (1.3) gives

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

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

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

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

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

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

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

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

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

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

1.6.4 The special case of a finite realm

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

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

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

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

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

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

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

Let ϕ be written

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

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

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

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

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

This is possible if and only if

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

Multiplying out and taking expectations then gives

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

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

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

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

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

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

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

and undefined otherwise.

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

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

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

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

and undefined otherwise.

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

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

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

1.7 Conditional probabilities

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

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

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

and undefined otherwise.

Then by Theorem 1.14 we have

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

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

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

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

Theorem 1.16 (Sequential conditioning)**.**

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

and undefined otherwise.

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

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

□

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

Theorem 1.17 (Factorisation theorem)**.**

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

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

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

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

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

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

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

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

Law of Total Probability (LTP)
Partition Theorem

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

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

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

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

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

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

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

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

Bayes's Theorem

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

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

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

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

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

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

from (1.7). Rearranging gives the result. □

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Bayesian Confirmation Theory

1.8 Probability mass functions

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

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

probability mass function (PMF)

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

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

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

support

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

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

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

Then

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

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

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

and

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

for all real-valued g.

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

marginal PMFs

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

which is the standard rule for marginalising out Z.

For conditional probabilities, the definition is

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

and then Theorem 1.14 and (1.7) imply that

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

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

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

statistical parameters

statistical modelling

1.8.1 Non-finite realms*

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

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

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

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

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

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

and hence

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

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

countable additivity

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

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

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

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

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

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

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

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

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

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

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

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

distribution function

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

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

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

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

Formally, the definition of f_X is

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

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

probability density function

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

The marginalisation result is

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

And the conditioning result is

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

and undefined otherwise.

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

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

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

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

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

2
Making decisions

We all appreciate that actions have consequences. What interests statisticians are those situations where the consequences are not known at the point where the action must be chosen. We can formalise this as follows:

- There is set of actions termed the *action set*, denoted $\mathcal{A}$ with typical element a, for 'action'. action set
- The consequence of an action depends on a vector of random quantities $X \in \mathcal{X}$, termed the *state of nature*. This is a conventional label going back over half a century—it does not necessarily connote anything 'natural'. state of nature
- The conjunction of an action and the state of nature is represented as a scalar real-valued *loss function* loss function

  $$L : \mathcal{A} \times \mathcal{X} \to \mathbb{R};$$

  the value $L(a, x)$ is the loss that is incurred on choosing action a if the state turns out to be x.

I will treat both $\mathcal{A}$ and $\mathcal{X}$ (and $\mathcal{Y}$ below) as finite. Many of the following results have analogues for non-finite sets, but these typically require additional technical conditions (Section 2.2.2).

The two main players in a decision analysis are the *statistician* (whom I will take to be male) and the *client* (female—although sometimes the statistician is his own client). A third player, not present but who must be borne in mind, is the client's *auditor*. Together the statistician and the client determine $\mathcal{A}$, $\mathcal{X}$, and L, and other quantities below, to best represent the client's needs. The auditor is typically not present during this process, but may critically evaluate its technical aspects, for example when reviewing the decision/report. In many situations, the client is the agent for a group of *stakeholders*, who appoint an auditor to keep an eye on her.

Smith (2010, chapter 1) provides a good introduction to decision analysis, and to modern developments in applied decision analysis. Cox and Hinkley (1974, ch. 11) is a helpful one-chapter summary of the traditional theory. Berger (1985) has most of the technical material, and there are more recent references in Robert (2007). For foundational aspects, including why it is appropriate to minimise

the expected loss (Section 2.1), Savage (1954) and DeGroot (1986, notably ch. 7) are both classics.

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

2.1 *No-data decision analysis*

In this simple analysis the client needs to choose an action without access to any further information. There are two cases to distinguish. In the first, the client feels able to specify f_X, her distribution for the unknown state of nature; but in the second she does not, or would rather not.

In the first case, it is possible to construct a complete ordering over the elements of $\mathcal{A}$, in which

$$a \preceq a' \text{ exactly when } \mathsf{E}\{L(a, X)\} \geq \mathsf{E}\{L(a', X)\}.$$

In other words, a is no better than a' exactly when the *expected loss* from choosing a is at least as large as that from choosing a'. Thus we can define the *Bayes action*, as the best action according to this ordering.

expected loss

Bayes action

Definition 2.1 (Bayes action). *Action $a^* \in \mathcal{A}$ is a Bayes action exactly when*

$$a^* = \operatorname*{argmin}_{a \in \mathcal{A}} \mathsf{E}\{L(a, X)\}.$$

Note that a Bayes action may not be unique. Here I am treating f_X as invariant to the choice of action, for simplicity; the generalisation is discussed in Section 2.3.

The second case is more tricky, because there is no complete ordering over the actions, and so no way to define an action as 'best'. However, it is at least possible to rule out bad choices, which are actions that are dominated by other actions.

Definition 2.2 (Inadmissible actions). *Action a is dominated by action a' exactly when $L(a, x) \geq L(a', x)$ for all x, and $L(a, x) > L(a', x)$ for at least one x. Action a is* inadmissible *exactly when it is dominated by another action. Action a is* admissible *exactly when it is not inadmissible.*

inadmissible

admissible

If the loss function has been carefully considered, then recommending an inadmissible action is simply a mistake, and cannot be defended. Suppose it was possible to establish whether any action was admissible or inadmissible. In this case, the auditor could quite reasonably insist that the statistician prove that the recommended action is admissible. And this indeed this is possible, as I now show.

First, it is necessary to broaden the set of available actions. An action in $\mathcal{A}$ is termed a *pure action*. But it is also possible to have mixed actions, which are random combinations of the actions in $\mathcal{A}$. Let S^{m-1} be the *unit $(m-1)$-simplex*

pure action

unit $(m-1)$-simplex

$$\mathbb{S}^{m-1} := \{w \in \mathbb{R}^m : w_i \geq 0, \sum_i w_i = 1\}.$$

If $\mathcal{A}$ is of size m and $w \in \mathbb{S}^{m-1}$, then $w \cdot a$ is a *mixed action* where action a_i is randomly selected with probability w_i. The loss for a mixed action is

mixed action

$$L(w \cdot a, x) = \sum_i w_i \, L(a_i, x),$$

which is the expected loss over the random choice of action. With this broader set of actions, pure actions can be dominated by mixed actions even if they are not dominated by other pure actions; see Figure 2.1. For us, mixed actions are just a device. The statistician would not tell the client, "You should randomly select action a_i with probability w_i"—see the end of this section.

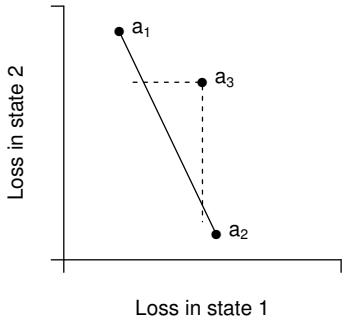


Figure 2.1: The pure action a_3 is dominated by some mixtures of a_1 and a_2.

Now it is possible to prove that Bayes actions are admissible actions, and *vice versa*.

Theorem 2.1 (Bayes actions and admissible actions)**.** *Let $\mathcal{A}$ and $\mathcal{X}$ be finite.*

1. *If the support of f_X is $\mathcal{X}$, then a Bayes action is admissible.*
2. *Allowing for mixed actions, every admissible action is a Bayes action for some f_X with support $\mathcal{X}$.*

Proof. Recollect that the support of f_X is the set $\{x \in \mathcal{X} : f_X(x) > 0\}$. So in both parts of this proof, we have $f_X(x_j) > 0$ for all $x_j \in \mathcal{X}$.

1. This proof holds for general $\mathcal{A}$ and $\mathcal{X}$, subject to the condition that $\mathcal{A}$ is restricted to those elements for which $\mathsf{E}\{L(a, X)\}$ is finite.

Suppose, that a Bayes action a^* is inadmissible, being dominated by some other pure action a'. But in this case

$$\begin{aligned} \mathsf{E}\{L(a', X)\} &= \sum_j L(a', x_j) f_X(x_j) \\ &< \sum_j L(a^*, x_j) f_X(x_j) \qquad \text{because } a' \text{ dominates } a^* \\ &= \mathsf{E}\{L(a^*, X)\}, \end{aligned}$$

where the inequality is strict because the support of f_X is $\mathcal{X}$. Hence a^* could not be a Bayes action because a' has a smaller expected loss: a contradiction. Thus the Bayes action must be admissible.

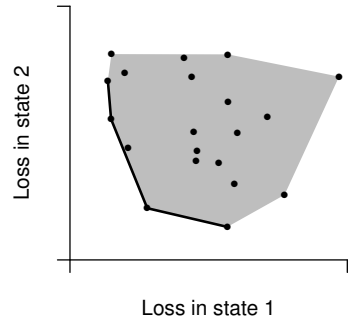


Figure 2.2: The set of pure actions, the convex hull of mixed actions, and the admissible actions.

2. This proof uses concepts from *convex analysis*; see Whittle (2000, section 15.2). In particular, the characterisation of boundary points is the *supporting hyperplane theorem*. Both $\mathcal{A}$ and $\mathcal{X}$ must be finite.

convex analysis

supporting hyperplane theorem

Let L be the $(m \times n)$ matrix with ij component $L(a_i, x_j)$, and denote one row of L as L_i. Here $L_i \in \mathbb{R}^n$ represents the loss vector for action i. Allowing for mixed actions, the set of all possible loss vectors is

$$[L] := \left\{ \ell \in \mathbb{R}^n : \ell = \textstyle\sum_i w_i \, L_i \text{ for some } w \in \mathbb{S}^{m-1} \right\},$$

termed the *convex hull* of $\{L_1, \ldots, L_m\}$. See Figure 2.2. The admis-

convex hull

sible actions are all those actions with losses on the southwest boundary of this convex hull. Now ℓ is on the boundary of $[L]$ if and only if there is a $p \in \mathbb{R}^n$ such that

$$p \cdot \ell = c, \text{ where } c := \inf_{\ell' \in [L]} p \cdot \ell'.$$

For ℓ to be on the southwest boundary, $p \cdot (\ell + \mathrm{d}\ell) > c$ for all $\mathrm{d}\ell > \mathbf{0}$, which in turn implies that $p \gg \mathbf{0}$.[1] Thus if $w \cdot a$ is an admissible action with loss vector ℓ, and $w' \cdot a$ is any action with loss vector ℓ', then there is a $p \gg \mathbf{0}$ for which

$$\sum_j p_j \, L(w \cdot a, x_j) \leq \sum_j p_j \, L(w' \cdot a, x_j).$$

And hence $w \cdot a$ is a Bayes action for $f_X(x_j) := p_j / \sum_{j'} p_{j'}$, where the support of f_X is $\mathcal{X}$. □

[1] Notation. If x and y are vectors, then (i) $x > y$ exactly when $x_i \geq y_i$ for all i and $x_i > y_i$ for some i; (ii) $x \gg y$ exactly when $x_i > y_i$ for all i.

If the client specifies an f_X with support on the whole of $\mathcal{X}$, then the first part of the theorem shows how the statistician can ensure that the recommended action is admissible by choosing a Bayes action. The second part of the theorem shows that all admissible actions can be represented this way. In other words, there is always an f_X implicit in the choice of an admissible action. So the auditor can say to the statistician,

> "Prove to me that the action you have recommended is admissible, by providing an f_X under which it is the Bayes action."

And even if the statistician can produce such an f_X his ordeal is still not over, because now the auditor can ask him to defend this choice of f_X as a reasonable representation of the client's judgements. Therefore if admissibility is taken to be a minimal requirement for any defensible action, there is really no evading eliciting an f_X from the client.

Finally, just to clear up the situation with mixed actions, the following theorem shows that it is never necessary to recommend a mixed action, on the basis of minimised expected loss.

Theorem 2.2. *Suppose that $w \cdot a$ is an admissible mixed action. Let f_X be the distribution under which this is a Bayes action; such a distribution must exist according to Theorem 2.1. For this f_X there is an admissible pure action with the same expected loss as $w \cdot a$.*

Proof. Admissible mixed actions lie in the interior of facets of the convex hull of the rows of the loss matrix. Every point on the facet is admissible and has the same expected loss; this includes the corner points, which are pure actions. □

2.2 *With-data decision analysis*

Suppose that the client will know the value of some *observations* Y at the time when the choice of action must be made. For this section I will assume that the client is able to specify the conditional

observations

distribution $f_{Y|X}$. For many types of observation this would be reasonable; for example, Y might be imperfect measurements on a subset of X, or a subset of the population represented by X, or causally dependent on X. The more general case is discussed in Section 2.5.

The dependence of the choice of action on the value of the observations is represented as a *decision rule* decision rule

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

Hence if $\delta(y) = a$ then δ is a rule that says "choose a when $Y = y$". Rather than looking for the optimal action, now we look for the optimal rule. The space of all rules is denoted $\mathcal{D}$. We can have pure rules and mixed rules.

Happily, it will turn out that this analysis is analogous to the no-data analysis. First, it is necessary to define some terms.

Definition 2.3 (Risk function). $R : \mathcal{D} \times \mathcal{X} \to \mathbb{R}$ *is a* risk function *exactly when* risk function

$$R(\delta, x) = \mathsf{E}\{L(\delta(Y), x) \mid X = x\}.$$

In a risk function, the expectation is taken over Y conditional on $X = x$. The risk function is like the loss function, except with a rule instead of an action. An *admissible rule* is one which is not dominated by any other rules. admissible rule

If the client is able to specify f_X, her distribution for X, then we can also define the integrated risk. This is like the expected loss, except with a rule instead of an action.

Definition 2.4 (Integrated risk). $R : \mathcal{D} \to \mathbb{R}$ *is the* integrated risk *exactly when* integrated risk

$$R(\delta) = \mathsf{E}\{L(\delta(Y), X)\}.$$

The expectation in the integrated risk is over the joint distribution of $\{X, Y\}$. The risk function and the integrated risk are related through the LIE (Theorem 1.11 and eq. (1.6)):

$$\begin{aligned} R(\delta) &= \mathsf{E}\left[L(\delta(Y), X)\right] \\ &= \mathsf{E}\left[\mathbb{E}\{L(\delta(Y), X) \mid X\}\right] \qquad \text{by the LIE} \\ &= \mathsf{E}\left[R(\delta, X)\right], \end{aligned}$$

where the expectation in the final line is over the client's distribution for X.

Finally, the Bayes rule is like the Bayes action, except that it minimises the integrated risk instead of the expected loss. Computing the Bayes rule is discussed in Section 2.2.1.

Definition 2.5 (Bayes rule). $\delta^* : \mathcal{Y} \to \mathcal{A}$ *is a* Bayes rule *exactly when* Bayes rule

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

Precisely the same reasoning as before, with δ instead of a and $R(\delta, x)$ instead of $L(a, x)$, gives the following theorem (the analogue of Theorem 2.1). Note that the space of all possible pure rules is finite, if $\mathcal{A}$ and $\mathcal{Y}$ are finite.

Theorem 2.3 (Bayes rules and admissible rules)**.** *Let $\mathcal{A}$, $\mathcal{X}$, and $\mathcal{Y}$ be finite.*

1. *If the support of f_X is $\mathcal{X}$, then a Bayes rule is admissible.*
2. *Allowing for mixed rules, every admissible rule is a Bayes rule for some f_X with support $\mathcal{X}$.*

The analogous result that every mixed rule can be replaced by a pure rule with the same integrated risk also holds (Theorem 2.2). And the same conclusions hold: if the statistician is going to recommend a rule to the client, it might as well be a Bayes rule for a defensible choice of f_X. More general situations, with the same conclusion, are outlined in Section 2.2.2.

* * *

A decision rule is like a *playbook*. Before knowing the value of the observations, the client is able to say how she would act for each possible outcome in $\mathcal{Y}$. Thus, a decision rule is about being prepared. If the client is responsible for real-time risk management, then she can respond rapidly to the observations as they come in, as the Bayes action has already been computed (see Section 2.2.1).

playbook

Of course it is rarely that simple in practice, as the actual observations will tend not to be precisely the ones that were anticipated, and not a superset of them either. In this situation the decision rule is more about guidance: the client might find a y in the playbook that is sufficiently like the actual observations that $\delta(y)$ is a reasonable candidate for a good action. And presumably the process of computing the decision rule, involving specifying an action set and a loss function and thinking about uncertainty, also equips the client to make better decisions under pressure.[2]

[2] "Plans are worthless, but planning is everything.", Dwight D. Eisenhower, 1957.

In other situations, where a rapid response is not required, decision rules are important in experimental design (Section 2.3), but less so in choosing between actions. If the observations are already known, say $Y = y^{\text{obs}}$, then there is little reason to compute the decision rule for any value of y other than y^{obs}. Generally, it is helpful to distinguish between a *pre-data analysis*, where $\mathcal{Y}$ is known but the value of Y is not, and a *post-data analysis*, where Y is known to take the value y^{obs}. Chapter 4 is all about post-data analysis.

pre-data analysis

post-data analysis

2.2.1 *Computing the Bayes rule*

Computing the Bayes rule looks like a difficult problem, because the minimisation is over the space of functions which map $\mathcal{Y}$ to $\mathcal{A}$, which might be a very inconvenient space to work with. But in fact there is a celebrated result which shows that this problem is much easier than it appears.

Theorem 2.4 (Bayes rule theorem)**.** *The Bayes rule is*

$$\delta^*(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathsf{E}\{L(a, X) \mid Y = y\} \quad \textit{for each } y \in \mathcal{Y}. \tag{†}$$

In other words, one simply minimises the expected loss conditional on $Y = y$ to find the optimal action $\delta^*(y)$.

Proof. Let δ be any rule. For any y,

$$\begin{aligned} \mathsf{E}[L(\delta(y), X) \mid Y = y] &\geq \min_a \mathsf{E}[L(a, X) \mid Y = y] \\ &= \mathsf{E}[L(\delta^*(y), X) \mid Y = y] \end{aligned}$$

using the definition in (†). Hence, using the LIE and the monotonicity property of expectations, if δ is any rule, then

$$\begin{aligned} R(\delta) &= \mathsf{E}\{L(\delta(Y), X)\} \\ &= \mathsf{E}\{\mathbb{E}[L(\delta(Y), X) \mid Y]\} \\ &\geq \mathsf{E}\{\mathbb{E}[L(\delta^*(Y), X) \mid Y]\} \\ &= \mathsf{E}\{L(\delta^*(Y), X)\} \\ &= R(\delta^*) \end{aligned}$$

showing that δ^* is an optimal rule, as its integrated risk is the lower bound of all possible rules. □

Thus the Bayes rule can be deduced from the client's $\mathcal{A}$, L, $f_{Y|X}$, and f_X. The statistician's role is:

1. Helping the client to quantify her judgements for $\mathcal{A}$, L, $f_{Y|X}$, and f_X;
2. Computing $f_{X|Y}$ from f_X and $f_{Y|X}$, by Bayes's Theorem (Theorem 1.19);
3. Minimising $\mathsf{E}\{L(a, X) \mid Y = y\}$ over a to find $\delta^*(y)$.

If there is a line to be drawn between statistical inference and decision theory it lies between inferring $f_{X|Y}$, which is statistical inference, and recommending an action, which additionally requires an action set and a loss function. Conceivably, the same statistical inference could serve many different clients, if all clients were to agree on $f_{Y|X}$ and f_X.

2.2.2* *Complete class theorems*

The results presented in Theorem 2.3 hold for finite $\mathcal{A}$, $\mathcal{X}$, and $\mathcal{Y}$. As discussed in Chapter 1, notably Section 1.3 and Section 1.8.1, there will be times when the statistician wants more general results, covering the cases where these sets are non-finite, and possibly non-bounded. The mathematics is complicated, but there are some general conclusions which emerge.[3]

It is helpful to classify sets of rules in the following way.

[3] This material borrows heavily from Berger (1985, ch. 8).

Definition 2.6 (Complete class)**.** *The set of rules $\mathcal{C} \subset \mathcal{D}$ is* complete *exactly when every $\delta \notin \mathcal{C}$ is dominated by a $\delta' \in \mathcal{C}$. The set $\mathcal{C}'$ is* minimal complete *if it is the smallest possible complete class.*

complete

minimal complete

It is easy to check that if $\mathcal{C}$ is complete and $\delta \notin \mathcal{C}$, then δ is inadmissible; that there can be $\delta' \in \mathcal{C}$ which are also inadmissible (dominated by other members of $\mathcal{C}$); and that if $\mathcal{C}'$ is minimal complete it is exactly the class of admissible rules. Therefore the simplest guideline for favouring admissible rules is only to select rules from a complete class. In this way the statistician will not always find an admissible rule (unless his class is minimal complete) but at least he has the *possibility* of finding an admissible rule. Once he strays outside a complete class, he has no possibility of finding an admissible rule. In other words, being in a complete class is a necessary condition for admissibility.

Now it is possible to give a more general result for the case where both $\mathcal{A}$ and $\mathcal{X}$ are finite. *If the loss function is bounded below, then the set of Bayes rules is a complete class, and the set of admissible Bayes rules is a minimal complete class* (Berger, 1985, sec. 8.2.3). This generalises Theorem 2.3 because it does not require f_X to have support on the whole of $\mathcal{X}$; see Figure 2.3. A very similar result is available for the case where $\mathcal{A}$ is non-finite.

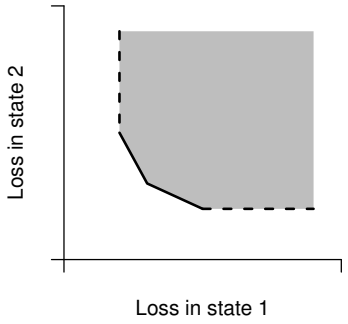


Figure 2.3: Admissible Bayes rules (solid line) form a minimal complete class. The dashed line shows Bayes rules which are not admissible.

However, the extension to non-finite $\mathcal{X}$ is more complicated. In this case it turns out that Bayes rules do not form a complete class, and it is necessary to introduce a more general type of rule. I will treat X as a continuous random quantity with an uncountable realm, as this is the common case.

A real-valued function g_X on $\mathbb{R}$ for which $g_X(x) \geq 0$ and $\int g_X(x)\,\mathrm{d}x = \infty$ is said to be an *improper PDF* if taken as a PDF for X. Nevertheless, the conditional distribution of X given Y,

improper PDF

$$f_{X|Y}(x \mid y) \propto f_{Y|X}(y \mid x)\, g_X(x)$$

may still be a proper PDF, if the first term on the righthand side has sufficiently tight tails as a function of x for given y. The use of improper PDFs for X allows for more general Bayes rules.

Definition 2.7 (Generalised Bayes rule)**.** *δ^* is a* generalised Bayes rule *exactly when it is a Bayes rule for an improper PDF for X.*

generalised Bayes rule

The general result is that, in many situations, *the union of Bayes rules and generalised Bayes rules forms a complete class* (Berger, 1985, sec. 5.8). Or, to put it another way,

> "… all admissible procedures are approximately Bayes. […] So not much is lost by confining attention to Bayes procedures." (Diaconis and Freedman, 1986, p. 10)

Improper PDFs are often limits of proper PDFs,[4] and therefore the set of generalised Bayes rules can be thought of as the boundary of the set of Bayes rules, and the union of Bayes rules and generalised Bayes rules can be thought of as the closure of the set of Bayes rules.

[4] Here is the most common instance. Let ϕ be the PDF of a standard Normal random quantity. Then $\phi(x/\sigma)/\sigma$ is the PDF of a Normal random quantity with expectation zero and variance σ^2, and $\lim_{\sigma\to\infty} \phi(x/\sigma)/\sigma$ is the improper uniform distibution on the whole of $\mathbb{R}$.

2.3* *More complicated decisions*

The two complications we need to analyse are where the client's f_X depends on her choice of action, and where multiple decisions must be made sequentially. Here I consider a classic situation where an experiment is chosen and performed, and then an action is chosen: termed *experimental design*. This was first analysed in detail by Raiffa and Schlaifer (1961, chapter 1); generalisations to finite sequences of decisions are straightforward, although the notation gets intense (see, e.g., Bernardo and Smith, 2000, sec. 2.6).

experimental design

A sequential decision analysis can be effectively represented in terms of a *rollback tree* (see Smith, 2010, ch. 2, which also discusses other types of decision tree). This is a type of graph termed a rooted tree[5] in which actions are represented with square vertices, and random quantities with round ones. The actions and random quantities are represented sequentially, so that any action or random quantity in the path from the root to a vertex has the capacity to affect that vertex, either by being known at the time where the decision is made, or by having an effect on the outcome.

rollback tree

[5] A connected graph with no cycles, and a designed root which in our case represents the first decision to be made.

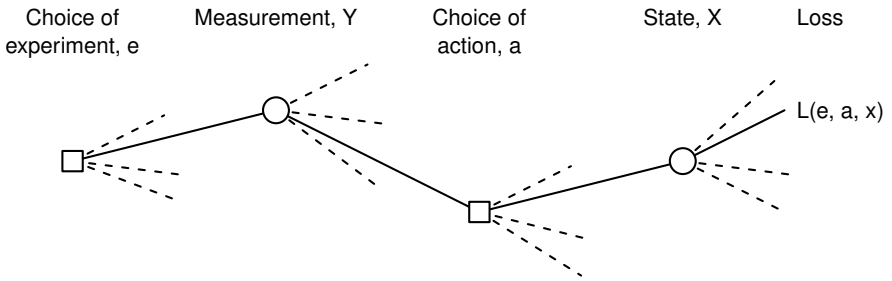


Figure 2.4: Rollback tree for experimental design.

One path through the rollback tree for experimental design is shown in Figure 2.4. Initially, an experiment is chosen from a finite set $\mathcal{E} := \{e_0, e_1, \ldots, e_k\}$. Then observations $Y \in \mathcal{Y}_i$ are made, where $\mathcal{Y}_i$ is the realm for experiment i. Then an action $a \in \mathcal{A}_i$ is chosen, where the action space might also depend on the experiment. The loss incurred depends on the experiment, the action, and the state of nature, X; hence the loss function has the form $L(e, a, x)$.[6]

[6] I am simplifying here by treating the loss as invariant to the value of Y, but allowing this would cause no additional difficulties.

The two random quantities in this analysis are X and Y, and the client's uncertainty is represented in a joint distribution $f_{X,Y}(x, y; e, a)$. The presence of e and a indicate that this joint distribution depends on the experiment that was done (because this determines the realm of Y), and also on the action chosen. Thus e and a are control parameters, not random quantities, and by convention they are included as arguments after a semicolon.

It is quite common that a distribution for the state of nature X

will depend on a. For example, climate change mitigation strategies such as geo-engineering are designed to alter future weather. More local interventions have the same intention. For example, building a levee does not change the weather, but it does change the impact of a flood following a storm. In both of these examples some uncertain thing in the future (X) will change as a result of the client's choice of action (a), and so the client's distribution for the thing should depend on the chosen action.

From the joint distribution for $\{X, Y\}$ we can derive the client's conditional and marginal distributions, notably $f_{X|Y}(x \mid y; e, a)$, $f_Y(y; e, a)$, and $f_X(x; e, a)$. The second of these has the property

$$f_Y(y; e, a) = f_Y(y; e),$$

that is, any distribution for the outcome of a experiment is invariant to the chosen action, which happens subsequently.

There is an attractive way to process a rollback tree, which is to start from the final decision and work backwards, termed *backward recursion*. This implements *Bellman's principle of optimality*.

backward recursion
Bellman's principle of optimality

Definition 2.8 (Bellman's principle of optimality)**.** *When making any choice, assume that all future choices will be made optimally.*

So, starting at the end of the rollback tree, the first step is to decide on the optimal choice of action, for a given experiment and observations. For this simple analysis, we can just use a Bayes rule, which is

$$\delta^*(e_i, y) := \operatorname*{argmin}_{a \in \mathcal{A}_i} \mathsf{E}\{L(e_i, a, X) \mid Y = y; e_i, a\}.$$

The minimised risk at this point is then

$$R^*(e_i, y) := \mathsf{E}\left\{L(e_i, \delta^*(e_i, y), X) \mid Y = y; e_i, \delta_i^*(e_i, y)\right\}.$$

We ought not to call it the Bayes risk (just yet), so instead call it the CJO risk, where CJO stands for *current judgement optimal* (Smith, 2010, sec. 2.5).

current judgement optimal

Now we have to choose the experiment. Bellman's principle is to select the experiment with the smallest CJO risk, now treating Y as unknown. The CJO risk for experiment i is

$$R^*(e_i) := \mathsf{E}\{R^*(e_i, Y); e_i\},$$

where there is no a in the arguments of the expectation because the marginal distribution of Y is invariant to a. Then the optimal experiment is $e^* := \operatorname{argmin}_e R^*(e)$. The CJO risk of the whole decision analysis is $R^* := R^*(e^*)$.

There is no doubt that backward recursion will result in a strategy $\{e^*, \delta^*\}$. What is not at all obvious is that $\{e^*, \delta^*\}$ is a Bayes rule; nevertheless, it is true under very natural conditions (Whittle, 1996, Appendix B). The proof is trickier than you might think, because of the mixture of random quantities and optimisation, and

simple proofs tend to ignore the important possibility that (e, a) can affect the client's distribution for X.

In the case where the experiment is fixed but the choice of action affects the client's uncertainty about the state of nature, the Bayes rule is easily seen to be

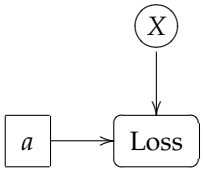
$$\delta^*(y) := \operatorname*{argmin}_{a \in \mathcal{A}} \mathsf{E}\{L(a, X) \mid Y = y; a\},$$

applying a simple generalisation of Theorem 2.4.

2.3.1* *Influence diagrams*

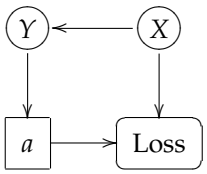
Rollback trees are thorough but they involve a lot of branching. *Influence diagrams* are more helpful in the early stages of structuring and visualising a decision analysis (Clemen, 1996, chapter 3). The influence diagram for the no-data analysis in Section 2.1 is

Influence diagrams



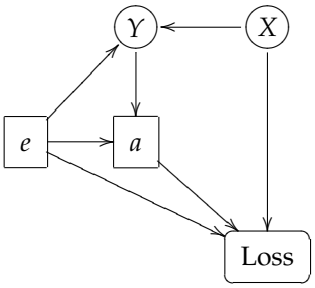
Rectangles represent choices, circles represent random quantities, and rounded rectangles represent consequences. Edges indicate either relevance or sequence; edges into random quantities or consequences represent relevance, while edges into actions represent sequence, showing what is known at the point where the choice is made.

The influence diagram for the with-data analysis (Section 2.2) is



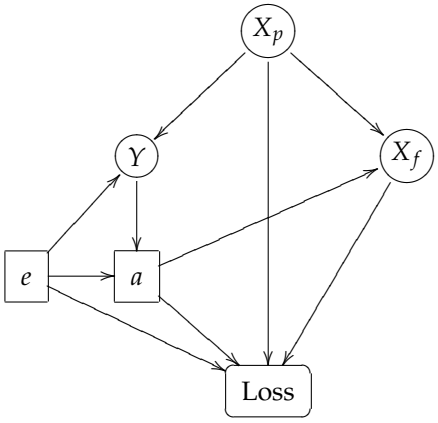
The edge from X to Y shows that the value of X is relevant for the value of Y, and the edge from Y to a shows that Y will be known when a is chosen.

The influence diagram for the experimental design analysis of this section is



The edge from e to 'Loss' represents the additional costs incurred in performing the experiment. It is tempting to add an edge from

a to X, to show that the action affects judgements about the state of nature. But this would create a closed loop, which is logically invalid. Instead, X can be split into 'past and current X', say X_p, which affects Y, and 'future X', say X_f, which can be affected by a:



2.4* *Valuing information*

Someone trustworthy[7] claims to know the actual value of Y—how much should the client pay for this information? The obvious point to make is that it depends on her action set and her loss function. For simplicity, I will assume in this section that the client's loss function is expressed in currency.

[7] Unstrustworthy is another interesting problem.

It helps to consider some extreme cases. First, suppose that the client does without the information. In that case, her expected loss would be

$$R^* := \underset{a \in \mathcal{A}}{\operatorname{argmin}} \, \mathsf{E}\{L(a, X)\}.$$

Second, suppose that the information is free. In the special case where the information completely determined X, her expected loss would be

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

because X would be known before the action was chosen. In the more general case, her expected loss would be $R(\delta^*)$, the minimised integrated risk. These three expected losses must satisfy the inequalities

$$R^{**} \leq R(\delta^*) \leq R^*,$$

the first inequality because Y cannot be more than perfectly informative, and the second because ignoring Y is a possible rule.

Now suppose that the client is contemplating paying p for the information. Clearly she would be happy with $p = 0$, but what is the most that she would pay? Adding p to the loss function, the largest acceptable value of p satisfies

$$R(\delta^*) + p \leq R^*.$$

Hence she would pay $p < p^{\max}$, where $p^{\max} := R^* - R(\delta^*)$. But finding $p^{\max}$ involves computing the Bayes rule. An upper bound on $p^{\max}$ can be found using

$$p^{\max} = R^* - R(\delta^*) \leq R^* - R^{**}.$$

This final term $R^* - R^{**}$ is the *Expected Value of Perfect Information (EVPI)*.

Expected Value of Perfect Information (EVPI)

The EVPI is quick to compute, because it is not necessary to introduce Y at all. Its value depends on the richness of the client's action set, and on the sensitivity of the loss function to different actions. It is a very good recommendation that a client contemplating a decision should first compute her EVPI. Then she will have a much clearer idea about how much she should pay for the observations Y, and whether or not she should hire a statistician to compute the Bayes rule and the minimised integrated risk.

2.5 *Statistical parameters*

This section is in some ways an iteration of Section 2.2, being a repackaging of the with-data analysis in the same way that the with-data analysis was a repackaging of the no-data analysis. The difference is that we are about to introduce new quantities, *statistical parameters*, which are extremely subtle and interpreted in different ways by different tribes of statisticians.

statistical parameters

The objective of introducing statistical parameters is to constrain the set of distributions in which $f_{X,Y}$ lies. The two interpretations are as follows.

1. The client judges that her distribution for $\{X, Y\}$ lies in some parametric class. This class is defined by a *parameter space* Ω and a family of *statistical models* $f_{X,Y;\theta}$. Her judgement is

   parameter space
   statistical models

   There exists a $\theta \in \Omega$ such that

   $$\Pr(X = x, Y = y) = f_{X,Y;\theta}(x, y; \theta)$$

   for every $\{x, y\}$ in $\mathcal{X} \times \mathcal{Y}$.

   It is traditional in this interpretation to separate the $\{x, y\}$ and θ arguments in the statistical model with a semicolon, to emphasise that θ is *not* a random quantity but an index.

2. The client finds it helpful to structure her distribution for $\{X, Y\}$ using a set of auxiliary random quantities $\theta \in \Omega$, a family of statistical models $f_{X,Y|\theta}$, and a prior distribution π_θ for which

   $$\Pr\{\theta \in [t, t + \mathrm{d}t)\} := \pi_\theta(t)\,\mathrm{d}t.$$

   Her judgement is

   $$f_{X,Y}(x, y) = \int f_{X,Y|\theta}(x, y \mid t)\,\pi_\theta(t)\,\mathrm{d}t.$$

   In this interpretation θ are random quantities, whose role is to help the client structure her judgements. The statistical model is a conditional probability distribution.

No matter what the interpretation, the parameter space Ω is usually finite-dimensional and convex, and I will take Ω to be a convex subset of $\mathbb{R}^d$. I will also settle on the notation

$$\text{statistical model} := f_{X,Y|\theta}(x, y \mid t)$$

because whether or not θ is a random quantity, it is definitely not a control variable, and I think it is a useful convention to use a semicolon to identify control variables (as in Section 2.3).

The formal effect of introducing parameters into the distribution of $\{X, Y\}$ is to replace X with θ. Hence the risk function becomes

$$R(\delta, t) := \mathsf{E}\left\{L(\delta(Y), X) \mid \theta = t\right\},$$

where the expectation is taken over $\{X, Y\} \mid (\theta = t)$. If a prior distribution π_θ is specified, the integrated risk becomes

$$\begin{aligned} R(\delta) &:= \mathsf{E}\left[L(\delta(Y), X)\right] \\ &= \mathsf{E}\left[\mathbb{E}\{L(\delta(Y), X) \mid \theta\}\right] \qquad \text{by the LIE} \\ &= \mathsf{E}\{R(\delta, \theta)\}, \end{aligned}$$

where the final expectation is taken over π_θ. All of the previous results regarding admissibility and Bayes rules still hold, except with θ rather than X and π_θ rather than f_X. So, for the record (and taking Ω to be finite—the generalisation was discussed in Section 2.2.2),

Theorem 2.5 (Bayes rules and admissible rules again)**.** *Let $\mathcal{A}$, $\mathcal{Y}$, and Ω be finite.*

1. *If the support of π_θ is Ω, then a Bayes rule is admissible.*
2. *Allowing for mixed rules, every admissible rule is a Bayes rule for some π_θ with support Ω.*

The analogous result that every mixed rule can be replaced by a pure rule with the same integrated risk also holds (Theorem 2.2).

* * *

There is, however, a difference in interpretation. In Section 2.1 and Section 2.2 the client may have been reluctant to specify a distribution for X, but at least she acknowledged that X was a random quantity. But now the situation is not so clear: *θ is not operationally defined*. In the first interpretation it is a manifestation of the client's difficulty in making a precise statement about $\{X, Y\}$; in the second, it is a device to structure $f_{X,Y}$. This is the point at which statisticians split into different tribes. *Bayesian statisticians* are prepared to provide a π_θ, but *Frequentist statisticians* are reluctant to provide a π_θ for something which they do not consider to be a random quantity.

Bayesian statisticians

Frequentist statisticians

Unfortunately, Frequentist statisticians are trapped by Theorem 2.5, if they want to avoid inadmissible rules. There are two possible

responses. The first is to agree that admissibility is a minimal requirement for any defensible rule, and so to use a prior distribution over Ω, even though θ is not thought of as a random quantity. According to the duck test[8] these statisticians are effectively Bayesian, although they can be identified by their strong preference for choosing the prior distribution according to a rule.[9] Recently this practice has acquired the label 'objective Bayes analysis'—in contrast to 'subjective Bayes analysis', where π_θ is an opportunity for the client to incorporate her judgements into the decision analysis. Of course 'objective' is hardly appropriate, given the amount of judgement that must go into the statistical model (and the loss function). Berger (2006) and Goldstein (2006) put the contrasting viewpoints.

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

[9] See Kass and Wasserman (1996) for a review of possible rules.

The second response is to abandon admissibility. This is not *terra incognito*, because a large amount of statistical theory was developed before admissibility and the complete class theorems were properly understood; this theory is mostly inadmissible, but its methods are often cheap, where they are applicable, and it has endured. The idea is to replace θ in the risk function with an *estimator* of θ based on y. Armed with an estimator, say $\tilde{\theta} : \mathcal{Y} \to \Omega$, the optimal rule becomes

estimator

$$\tilde{\delta}(y) := \operatorname*{argmin}_{a \in \mathcal{A}} R\big(a, \tilde{\theta}(y)\big) \quad \text{for each } y \in \mathcal{Y},$$

which I will term the *plug-in rule*.

plug-in rule

In general there is no reason at all for the plug-in rule $\tilde{\delta}$ and the Bayes rule δ^* to agree, since not only are the objective functions different in the two cases, but one depends on the choice of estimator, $\tilde{\theta}$, and the other depends on the choice of prior distribution, π_θ. The plug-in rule is unlikely to be admissible. But in some circumstances, some estimators give rise to plug-in rules which are very similar to Bayes rules, and are therefore effectively admissible. Estimators are covered in more detail in Chapter 5, and admissibility of the plug-in rule in Section 5.4.

2.6 *Choosing between two hypotheses*

This is one of the most studied situations, and it has a complete answer. The decision is to choose between two hypotheses, traditionally termed H_0 and H_1. We can treat $\theta \in \Omega := \{0, 1\}$ as the index of the two hypotheses, and the action set is also $\mathcal{A} = \Omega$. The client's loss function can be written

$$L(a,t) = \begin{array}{c|cc} & t=0 & t=1 \\ \hline a=0 & c_{00} & c_{01} \\ a=1 & c_{10} & c_{11} \end{array}$$

where making a mistake causes a larger loss, so that $c_{00} < c_{10}$ and $c_{11} < c_{01}$. Observations are available, with distributions

$$H_0 : Y \sim f_0 \quad \text{and} \quad H_1 : Y \sim f_1$$

under the two hypotheses. It is the fact that f_0 and f_1 are different under the two hypotheses that allows us to use Y to choose between them.

Now suppose that the client has prior probabilities

$$\mathsf{Pr}(\theta = 0) = \pi_0 \quad \text{and} \quad \mathsf{Pr}(\theta = 1) = \pi_1$$

where there is no necessity for $\pi_0 + \pi_1 = 1$, because H_0 and H_1 may be just two of a large collection of possible hypotheses. The Bayes rule solves

$$\delta^*(y) = \operatorname*{argmin}_{a \in \Omega} \mathsf{E}\{L(a, \theta) \mid Y = y\}$$

(Theorem 2.4). Thus the Bayes rule will select H_0 when the conditional expected loss for choosing H_0 is less than that for choosing H_1,

$$c_{00}\, f_{\theta|Y}(0 \mid y) + c_{01}\, f_{\theta|Y}(1 \mid y) < c_{10}\, f_{\theta|Y}(0 \mid y) + c_{11}\, f_{\theta|Y}(1 \mid y),$$

or, after rearranging,

$$\frac{f_{\theta|Y}(0 \mid y)}{f_{\theta|Y}(1 \mid y)} > c \quad \text{where } c := \frac{c_{01} - c_{11}}{c_{10} - c_{00}} > 0. \tag{†}$$

But Bayes's Theorem states

$$f_{\theta|Y}(t \mid y) = \frac{f_t(y)\, \pi_t}{f_Y(y)} \quad \text{for } t = 0, 1,$$

and so (†) can be written

$$\frac{f_0(y)}{f_1(y)} > c' \quad \text{where } c' := c\, \frac{\pi_1}{\pi_0} > 0.$$

Following exactly the same reasoning for when the Bayes rule will select H_1 gives the result:

$$\delta^*(y) = \begin{cases} 0 & \frac{f_0(y)}{f_1(y)} > c' \\ \text{toss a coin} & \frac{f_0(y)}{f_1(y)} = c' \\ 1 & \frac{f_0(y)}{f_1(y)} < c'. \end{cases} \qquad \text{for some } c' > 0. \tag{2.1}$$

According to Theorem 2.5, this is precisely the form of all possible admissible rules, and we summarise this in the following result.

Theorem 2.6 (Choosing between two hypotheses)**.** *A rule for choosing between two hypotheses H_0 and H_1 is admissible if and only if it has the form $f_0(y)/f_1(y) \gtrless c$ for some $c > 0$.*

* * *

The ratio $f_0(y)/f_1(y)$ has acquired several different names. If f_0 and f_1 are both directly specified distributions then it is termed the *likelihood ratio* for H_0 versus H_1. When parameters are involved, so that

likelihood ratio

$$f_0(y) = \int f_{Y|\psi}(y \mid v)\, \pi_\psi(v)\, \mathrm{d}v$$

and possibly similarly for f_1 (with a different statistical model and parameter space) then it is termed the *Bayes factor* for H_0 versus H_1.

Bayes factor

I will refer to $f_0(y)/f_1(y)$ as the *odds ratio* for H_0 versus H_1:

odds ratio

$$\text{odds ratio for } H_0 \text{ versus } H_1 := \frac{f_0(y)}{f_1(y)} = \frac{\mathsf{Pr}(Y = y \mid H_0)}{\mathsf{Pr}(Y = y \mid H_1)}.$$

The term 'odds' refers to a ratio of probabilities. Thus π_0/π_1 is the prior odds for H_0 versus H_1, and $f_{\theta|Y}(0 \mid y)/f_{\theta|Y}(1 \mid y)$ is the posterior odds. Applying Bayes's Theorem in odds form (see after Theorem 1.19),

$$\underbrace{\frac{f_{\theta|Y}(0 \mid y)}{f_{\theta|Y}(1 \mid y)}}_{\substack{\text{posterior}\\ \text{odds}}} = \underbrace{\frac{f_0(y)}{f_1(y)}}_{\substack{\text{odds}\\ \text{ratio}}} \times \underbrace{\frac{\pi_0}{\pi_1}}_{\substack{\text{prior}\\ \text{odds}}}$$

and hence $f_0(y)/f_1(y)$ is the ratio of the posterior odds to the prior odds, hence 'odds ratio'.

If the odds ratio is greater than one, then the observations y have changed the balance of probabilities between H_0 and H_1 in favour of H_0. An odds ratio greater than one does *not* mean that H_0 is more probable than H_1. For example, if the prior odds for H_0 versus H_1 is 0.1 and the odds ratio is 2 then the posterior odds is 0.2, and H_1 is still five times more probable than H_0. But if the odds ratio is, say, larger than 100, the posterior odds will favour H_0 for any reasonably balanced prior odds, and hence a large odds ratio is often taken as strongly supportive of H_0 over H_1. Of course this does not mean that the client should choose H_0. She should also consider the costs of making a mistake, captured in c. But if these costs are also reasonably balanced, then a large odds ratio for H_0 versus H_1 would suggest choosing H_0 over H_1.

Exactly the same reasoning applies in reverse if the odds ratio is small, say less than 0.01, which would suggest choosing H_1 over H_0.

Several authors, notably Jeffreys (1961, App. B), have proposed a scale for the odds ratio with conventional labels indicating the strength of evidence. Thus the statistician might report to the client that there is 'very strong evidence' for H_0 over H_1, rather than saying that the odds ratio for H_0 versus H_1 is, say, 55 (2 sf).[10] But the odds ratio is not a very complicated concept, and I think it would be better to present the value itself. In this way the client can assess her posterior odds for a range of prior odds, and her choice between H_0 and H_1 for a range of possible costs.

[10] 'Very strong evidence' indicating an odds ratio of between $10^{3/2}$ and 10^2.

Lindley (1991) presents an interesting application of this decision analysis, where H_0 and H_1 represent the innocence or guilt of a defendant in a court of law. Each piece of evidence is summarised in terms of its odds ratio, and the combined odds ratio for all of the evidence is the product of the odds ratios for each piece. Lindley also discusses prior probabilities, and society's loss function.

2.6.1* *The Neyman-Pearson approach*

What happens in the situation where the client feels unable to provide a c' in (2.1). There is a theory for this situation, associated with the statisticians Jerzy Neyman and Egon Pearson, and known as the *Neyman-Pearson (NP) approach*. For each possible value of c', compute

Neyman-Pearson (NP) approach

$$\alpha(c') := \mathsf{Pr}\left\{\frac{f_0(Y)}{f_1(Y)} < c' \,\middle|\, H_0\right\} \quad \text{and} \quad \beta(c') := \mathsf{Pr}\left\{\frac{f_0(Y)}{f_1(Y)} > c' \,\middle|\, H_1\right\}$$

where α and β are termed the *Type 1 and Type 2 error levels*, respectively. That is, $\alpha(c')$ is the probability of incorrectly choosing H_1, and $\beta(c')$ is the probability of incorrectly choosing H_0. Then, among all the possible values of $\big(\alpha(c'), \beta(c')\big)$, choose that c' which gives the best trade-off between the two errors. No further guidance is possible about what constitutes a good trade-off. The convention is to make the *status quo* hypothesis H_0, and set c' so that $\alpha(c') \approx 5\%$; although a very large $\beta(c')$ in this situation might prompt an increase in c' (giving a larger α and a smaller β).

Type 1 and Type 2 error levels

In the special case where it is also possible to control n, the number of observations, one can go further, and find the smallest value of n for which there exists a c' satisfying $\alpha(c') \leq \alpha_0$ and $\beta(c') \leq \beta_0$. Common values from medical science would be $\alpha_0 = 5\%$ and $\beta_0 = 20\%$.[11] These are purely conventional values.

[11] Still trying to track down a source for these values.

There is no doubt that one could adopt the NP approach to deciding between H_0 and H_1. But it seems self-evident that decisions *ought* to take account of costs of errors, and of existing information about which of the two hypotheses is more probable, *a priori*. Conventional thresholds for α and β completely ignore these factors. Moreover, were I the client, I expect I would find it easier to specify c and π_0/π_1 than to specify which point on $\big(\alpha(c'), \beta(c')\big)$ I prefer. Savage *et al.* (1962, pages 63–67) provides a more detailed discussion on this issue.

Rothman *et al.* (2008, ch. 10) provides a non-technical assessment of hypothesis testing (and related methods) in epidemiology.

3
Statistical modelling

This chapter defines the notion of conditional independence, and explains how it is the cornerstone of statistical modelling (Section 3.1 and Section 3.3). Two general classes of statistical models are developed, to serve as exemplars for the chapters than follow: Markov random fields (Section 3.2) and hierarchical models (Section 3.4). This chapter is *not* about the practice of statistical modelling and statistical computing; see, e.g., Davison (2003).

3.1 Conditional independence

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

conditional independence

Because independence is a special case of conditional independence, I will just explore conditional independence in this section. In the material below, independence results can be recovered simply by dropping the conditioning on $\boldsymbol{Z}$ (this follows from Theorem 1.12).

Now may be a good time to review Section 1.6. I will write $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ as boldface to emphasise that all three represent collections of random quantities. Here is a formal definition of conditional independence.[1]

[1] One possible definition, because there are several equivalent ones, as shown immediately below.

Definition 3.1 (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 $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$.*

Informally, this states that the optimal prediction of any function of $\boldsymbol{X}$ which is based on 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 any information about $\boldsymbol{X}$ which is not already present in $\boldsymbol{Z}$. Definition 3.1 appears to be asymmetric with respect to $\boldsymbol{X}$ and $\boldsymbol{Y}$, but the following equivalence theorem shows that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are symmetric in this definition, so that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \iff \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z}$.

Theorem 3.1 (Conditional independence, equivalencies)**.** *Let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ be collections of random quantities. The following are equivalent:*

1. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$.
2. $f_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{z})$ *whenever* $f_{\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{y}, \boldsymbol{z}) > 0$.
3. $f_{\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{z}) \cdot f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{y} \mid \boldsymbol{z})$ *whenever* $f_{\boldsymbol{Z}}(\boldsymbol{z}) > 0$.

Proof. I'll dispense with the bold symbols.

(1. $\Rightarrow$ 2.) If $f_{Y,Z}(y,z) > 0$ then $f_{X|Y,Z}(x \mid y,z)$ is well-defined. Setting $g(x) \leftarrow (x = x')$, (1.) then implies that $f_{X|Y,Z}(x' \mid y,z)$ is invariant to y, or $f_{X|Y,Z}(x' \mid y,z) = f_{X|Z}(x' \mid z)$ according to Theorem 1.12.

(2. $\Rightarrow$ 3.) It is always true that

$$f_{X,Y|Z}(x,y \mid z) = f_{X|Y,Z}(x \mid y,z) \cdot f_{Y|Z}(y \mid z) \qquad (\dagger)$$

whenever $f_Z(z) > 0$. This is because $f_{Y,Z}(y,z) = 0$ if and only if $f_{Y|Z}(y \mid z) = 0$, since $f_Z(z) > 0$. But if $f_{Y|Z}(y \mid z) = 0$ then $f_{X,Y|Z}(x,y \mid z) = 0$, and hence (†) has the form $0 = f_{X|Y,Z}(x \mid y,z) \cdot 0$, and the ambiguity of $f_{X|Y,Z}(x \mid y,z)$ if $f_{Y,Z}(y,z) = 0$ is immaterial. The result follows immediately.

(3. $\Rightarrow$ 1.) In general, if $\psi_g \in \mathcal{E}\{g(X) \mid Y, Z\}$ and $f_{Y,Z}(y,z) > 0$,

$$\psi_g(y,z) = \mathsf{E}\{g(X) \mid Y = y, Z = z\} = \sum_{x \in \mathcal{X}} g(x)\, f_{X|Y,Z}(x \mid y,z)$$

by the FTP (Theorem 1.3 and Section 1.8). By (†)

$$f_{X|Y,Z}(x \mid y,z) = \frac{f_{X,Y|Z}(x,y \mid z)}{f_{Y|Z}(y \mid z)}.$$

If (3.) holds,

$$f_{X|Y,Z}(x \mid y,z) = \frac{f_{X|Z}(x, \mid z) \cdot f_{Y|Z}(y \mid z)}{f_{Y|Z}(y \mid z)} = f_{X|Z}(x \mid z)$$

i.e. invariant to y, hence $\psi_g(y,z)$ is invariant to y. □

* * *

Finally, the following special case of conditional independence is frequently used, for reasons that will be explained in Section 3.4.

Definition 3.2 (Mutual conditional independence)**.** *Let $\boldsymbol{X} := (X_1, \ldots, X_m)$, and let $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ be disjoint subsets of $\boldsymbol{X}$. Then $\boldsymbol{X}$ is* mutually conditionally independent *given $\boldsymbol{Z}$ exactly when $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{Z}$ for all A and B.*

mutually conditionally independent

It follows immediately from Theorem 3.1 that $\boldsymbol{X}$ is mutually conditionally independent given $\boldsymbol{Z}$ if and only if

$$f_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{i=1}^{m} f_{X_i|\boldsymbol{Z}}(x_i \mid \boldsymbol{z}), \tag{3.1}$$

which may be written as

$$X_1, \ldots, X_m \mid \boldsymbol{Z} \overset{\text{ind}}{\sim} f_{X_i|\boldsymbol{Z}}.$$

A stronger version of this property holds if $f_{X_i|\boldsymbol{Z}} = f_{X|\boldsymbol{Z}}$; i.e. the same for all i. This venerable statistical model is also expressed as $\boldsymbol{X}$ is *independent and identically distributed (IID)* given $\boldsymbol{Z}$, which may be written as

independent and identically distributed (IID)

$$X_1, \ldots, X_m \mid \boldsymbol{Z} \overset{\text{iid}}{\sim} f_{X|\boldsymbol{Z}}. \tag{3.2}$$

This is a cornerstone of modelling using exchangeability (Section 3.3).

3.2 *Modelling using conditional independence*

Conditional independence is a representation of judgements about the *structure* of the relationship between random quantities (Cowell *et al.*, 1999; Smith, 2010). In this section I consider some of the more theoretical aspects of conditional independence judgements, and the practical issue of computation.

Consider a thought experiment in which all the of random quantities are taken two at a time, and in each case I ask whether, in my judgement, the pair X and Y are conditionally independent given all of the other quantities. One might not think that pairwise judgements of this nature could be sufficient to completely characterise all possible conditional independence relationships, which may often involve more than two random quantities at a time. But surprisingly and gratifyingly, this is exactly what happens.

Let $\boldsymbol{X} := \{X_1, \ldots, X_m\}$. For $C \subset \{1, \ldots, m\}$ let

$$\boldsymbol{X}_C := \bigcup_{i \in C} \{X_i\} \quad \text{and} \quad \boldsymbol{X}_{-C} := \bigcup_{i \notin C} \{X_i\}$$

be subsets of $\boldsymbol{X}$. The probability distributions

$$f_{X_i|\boldsymbol{X}_{-i}} \quad i = 1, \ldots, m$$

are known as the *full conditionals* of $f_{\boldsymbol{X}}$. The following result links $f_{\boldsymbol{X}}$ to its full conditionals.

full conditionals

Theorem 3.2 (Brook's lemma)**.** *Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be two points in the support of $f_{\boldsymbol{X}}$. Then*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x}') \prod_{i=1}^{m} \frac{f_{X_i|\boldsymbol{X}_{-i}}(x_i \mid x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_m)}{f_{X_i|\boldsymbol{X}_{-i}}(x'_i \mid x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_m)}.$$

Proof. It suffices to give this proof for $m = 3$.

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= f_{X_1,X_2,X_3}(x_1, x_2, x_3) \\ &= f_{X_3|X_1,X_2}(x_3 \mid x_1, x_2)\, f_{X_1,X_2}(x_1, x_2) \\ &= \frac{f_{X_3|X_1,X_2}(x_3 \mid x_1, x_2)}{f_{X_3|X_1,X_2}(x_3' \mid x_1, x_2)}\, f_{X_1,X_2}(x_1, x_2)\, f_{X_3|X_1,X_2}(x_3' \mid x_1, x_2) \\ &= \frac{f_{X_3|X_1,X_2}(x_3 \mid x_1, x_2)}{f_{X_3|X_1,X_2}(x_3' \mid x_1, x_2)}\, f_{X_1,X_2,X_3}(x_1, x_2, x_3'). \end{aligned}$$

Now iterate on the final term, starting with

$$\begin{aligned} & f_{X_1,X_2,X_3}(x_1, x_2, x_3') \\ &\quad = f_{X_2|X_1,X_3}(x_2 \mid x_1, x_3')\, f_{X_1,X_3}(x_1, x_3') \\ &\quad = \frac{f_{X_2|X_1,X_3}(x_2 \mid x_1, x_3')}{f_{X_2|X_1,X_3}(x_2' \mid x_1, x_3')}\, f_{X_1,X_3}(x_1, x_3')\, f_{X_2|X_1,X_3}(x_2' \mid x_1, x_3') \\ &\quad = \frac{f_{X_2|X_1,X_3}(x_2 \mid x_1, x_3')}{f_{X_2|X_1,X_3}(x_2' \mid x_1, x_3')}\, f_{X_1,X_2,X_3}(x_1, x_2', x_3') \end{aligned}$$

and so on. □

This result shows something quite remarkable—that the distribution $f_{\boldsymbol{X}}$ is completely determined by its full conditionals. This is because $\boldsymbol{x}'$ can be set to any value in the support of $f_{\boldsymbol{X}}$, and then the initial term $f_{\boldsymbol{X}}(\boldsymbol{x}')$ is just the normalising constant which ensures that $\sum_{\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}) = 1$.[2]

[2] I am not sure that the name 'Brook's lemma' is standard, but this is the name given in Rue and Held (2005, sec. 2.2).

Now consider any one of the full conditionals,

$$f_{X_i|\boldsymbol{X}_{-i}}(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m).$$

If this probability distribution is invariant to the value of x_j for some $j \neq i$, then $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-ij}$, according to Theorem 3.1. If this probability distribution is *not* invariant to x_j, then write $X_i \sim X_j$ to denote "X_i is a neighbour of X_j". Because conditional independence is symmetric, this relationship is reflexive. Were we to perform this operation for every pair of random quantities, making $m(m-1)/2$ judgements in all, we could construct a graph on the vertices $\boldsymbol{X}$, where there is an edge between X_i and X_j exactly when $X_i \sim X_j$. Denote this graph as $\mathcal{G}$; it encodes the following property.

Definition 3.3 (Pairwise Markov property (P))**.** $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-ij}$ *exactly when there is no edge between* X_i *and* X_j *in* $\mathcal{G}$.

Now it does not seem likely that arbitrary choices for the full conditionals which respect $\mathcal{G}$ will automatically give rise to a valid $f_{\boldsymbol{X}}$, as is apparent from Theorem 3.2—there is clearly a very complicated relationship between the full conditionals of $\boldsymbol{X}$ and $f_{\boldsymbol{X}}$ itself. The following very famous result completely characterises probability distributions that respect (P). In this result, a *clique* is either a single X_i or a subset of $\boldsymbol{X}$ with a full set of edges.[3]

clique

[3] Below I write '$f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$', but what I mean is the slightly more complicated statement that $f_{\boldsymbol{X}}(x_1, \ldots, x_m) > 0$ whenever $f_{X_i}(x_i) > 0$ for $i = 1, \ldots, m$.

Theorem 3.3 (Hammersley-Clifford theorem)**.** *If $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ then $f_{\boldsymbol{X}}$ satisfies the pairwise Markov property of $\mathcal{G}$ if and only if there is a set of positive functions $\{G_C\}$ for which*

$$z := \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{C \in \mathcal{C}} G_C(\boldsymbol{x}_C) < \infty$$

and

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{z} \prod_{C \in \mathcal{C}} G_C(\boldsymbol{x}_C) \tag{F}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$.

Proof.

($\Leftarrow$) Starting from eq. (F),

$$\begin{aligned} f_{X_i|\boldsymbol{X}_{-i}}(x_i \mid \boldsymbol{x}_{-i}) &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{\boldsymbol{X}_{-i}}(\boldsymbol{x}_{-i})} \\ &= \frac{\prod_C G_C(\boldsymbol{x}_C)}{\sum_{x_i'} \prod_C G_C(\boldsymbol{x}'_{C_i})} \end{aligned}$$

where

$$\boldsymbol{x}'_{C_i} := \begin{cases} \{x_i'\} \cup \boldsymbol{x}_{C \setminus i} & X_i \in C \\ \boldsymbol{x}_C & \text{otherwise.} \end{cases}$$

Now divide the cliques in $\mathcal{C}$ into two types: those that contain X_i, denoted $\mathcal{C}_i$, and those that do not. Then continue with

$$\begin{aligned} f_{X_i|\boldsymbol{X}_{-i}}(x_i \mid \boldsymbol{x}_{-i}) &= \frac{\prod_{C \in \mathcal{C}_i} G_C(\boldsymbol{x}_C) \cdot \prod_{C \notin \mathcal{C}_i} G_C(\boldsymbol{x}_C)}{\sum_{x_i'} \prod_{C \in \mathcal{C}_i} G_C(\boldsymbol{x}'_{C_i}) \cdot \prod_{C \notin \mathcal{C}_i} G_C(\boldsymbol{x}_C)} \\ &= \frac{\prod_{C \in \mathcal{C}_i} G_C(\boldsymbol{x}_C)}{\sum_{x_i'} \prod_{C \in \mathcal{C}_i} G_C(\boldsymbol{x}'_{C_i})}. \end{aligned}$$

If there is no edge from X_i to X_j in $\mathcal{G}$ then X_j is not in any of the sets in $\mathcal{C}_i$, and hence $f_{X_i|\boldsymbol{X}_{-i}}(x_i \mid \boldsymbol{x}_{-i})$ is invariant to x_j, as required.

($\Rightarrow$) See Besag (1974) for a beautiful and insightful proof. □

However, this is not the end of the story. Now consider a second possible property of $\mathcal{G}$.

Definition 3.4 (Global Markov property (G))**.** *Let A, B, and C be non-intersecting subsets of $\{1, \ldots, m\}$. Then $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{X}_C$ whenever every path from A to B in $\mathcal{G}$ passes through C.*

It is easy to see that

$$(G) \implies (P)$$

and, following the same pattern as the ($\Leftarrow$) branch of Theorem 3.3, to show that

$$(F) \implies (G).$$

But since Theorem 3.3 asserts that $(F) \iff (P)$, we have

$$(G) \iff (P).$$

In other words, the pairwise conditional independence graph $\mathcal{G}$ embodies the complete set of conditional independence judgements

about $\boldsymbol{X}$ (at least, when $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$). This is the surprising and gratifying result mentioned at the start of this section. Therefore we are fully justified in referring to $\mathcal{G}$ as the *conditional independence graph (CIG)* for $f_{\boldsymbol{X}}$.

conditional independence graph (CIG)

* * *

Many statistical applications deal with large collections of random quantities, which have a rather natural neighbourhood structure, represented by $\mathcal{G}$. For example: the pixels in an image, where each pixel's neighbours might be the eight pixels adjacent to it (i.e. including diagonals). Or the regions of a spatial map, where each region's neighbours might be the regions with which it shares a common boundary. Models that are built around conditional independence graphs are termed *Markov random fields (MRFs)*; see, for example, Cressie and Wikle (2011).

Markov random fields (MRFs)

MRFs have a crucial property, that will be relevant several times in the chapters that follow. Except in trivial cases and special cases, it is not possible to evaluate $f_{\boldsymbol{X}}(\boldsymbol{x})$ explicitly for specified $\boldsymbol{x}$, due to the intractable z term in Theorem 3.3. If $\boldsymbol{X}$ represented the 64 pixels in an 8×8 two-tone image, the sum over $\mathcal{X}$ for computing z would include $2^{64} \approx 2 \times 10^{19}$ terms—quite out of the question.

By extension, it is not possible to evaluate the marginal distribution $f_{\boldsymbol{X}_A}(\boldsymbol{x}_A)$ for $A \subset \{1, \ldots, m\}$, which also contains z, plus a sum over $\mathcal{X}_{-A}$. And nor is it possible to evaluate the conditional distribution $f_{\boldsymbol{X}_A|\boldsymbol{X}_B}(\boldsymbol{x}_A \mid \boldsymbol{x}_B)$, which does not contain z, but which contains sums over $\mathcal{X}_C$ and $\mathcal{X}_{A\cup C}$, where $A \cup B \cup C = \{1, \ldots, m\}$. The only probability distributions that can be easily evaluated are $f_{X_i|\boldsymbol{X}_{B_i}}$ where B_i is any superset of the neighbours of X_i. In this case

$$f_{X_i|\boldsymbol{X}_{B_i}}(x_i \mid \boldsymbol{x}_{B_i}) \propto \sum_{C \in \mathcal{C}_i} G_C(\boldsymbol{x}_C),$$

as shown in the proof of Theorem 3.3; the normalising constant involves a one-dimensional sum.

Naturally, there has been a lot of interest in non-trivial special cases, where a careful choice of $\mathcal{G}$ and the G_C functions makes it possible to evaluate marginal and conditional distributions. One very useful special case is *Gauss Markov random fields*, which is a rapidly-developing area (see Rue and Held, 2005; Lindgren *et al.*, 2011). More generally, there are advantages to choosing the G_C functions so that $f_{\boldsymbol{X}}$ is a member of the exponential family of distributions (see, e.g., Davison, 2003, sec. 5.2).

Gauss Markov random fields

3.2.1* *Introducing parameters*

As outlined in Section 2.5, there are two viewpoints regarding parameters. The Frequentist viewpoint is that they are an index within a family of distributions. Thus they may occur as arguments to the probability distribution of the X's. From the Hammersley-Clifford theorem (Theorem 3.3), they must therefore occur as argu-

ments to the G_C functions, to give

$$f_{\boldsymbol{X}|\theta}(\boldsymbol{x} \mid t) = \frac{1}{z(t)} \prod_{C \in \mathcal{C}} G_C(\boldsymbol{x}_C, t).$$

I note for later reference that, as explained immediately above, the function $f_{\boldsymbol{X}_A|\theta}(\boldsymbol{x}_A \mid t)$ cannot be evaluated as a function of t for given $\boldsymbol{x}_A$, except in trivial cases and special cases, because of the computational cost of $z(t)$ and of summing over $\mathcal{X}_{-A}$.

The Bayesian viewpoint is that parameters are random quantities in their own right. Therefore they can be represented as additional vertices in a larger CIG, for which the joint distribution now becomes

$$f_{\boldsymbol{X},\theta}(\boldsymbol{x}, t) = \frac{1}{z} \prod_{C \in \mathcal{C}} G_C([\boldsymbol{x}, t]_C).$$

Precisely how the parameters are included in the CIG is explained in Section 3.4.1.

3.2.2* *Gibbs sampling*

Statistical models based around conditional independence may seem rather intractable, according to the analysis at the end of Section 3.2. The reason that statistics has not ground to a halt in the face of increasing model size and complexity is that although it is hard to evaluate $f_{\boldsymbol{X}}$ and its marginal and conditional distributions, it is often very easy to sample from $f_{\boldsymbol{X}_A}$, and also from $f_{\boldsymbol{X}_A|\boldsymbol{X}_B}(\cdot \mid \boldsymbol{x}_B)$. This is due to the *Gibbs sampler*; see Cowell *et al.* (1999), Robert and Casella (2004), Gelman *et al.* (2003) or Lunn *et al.* (2013).

Gibbs sampler

In the simplest implementation, the Gibbs sampler is (setting $m = 4$ for convenience):

initialise
- $\boldsymbol{x}^0 \leftarrow (x_1^0, x_2^0, x_3^0, x_4^0)$, some point in $\mathcal{X}$ with $f_{\boldsymbol{X}}(\boldsymbol{x}^0) > 0$
- $j \leftarrow 0$

repeat
- Sample $x_1^{j+1} \sim f_{X_1|X_2,X_3,X_4}(\cdot \mid x_2^j, x_3^j, x_4^j)$
- Sample $x_2^{j+1} \sim f_{X_2|X_1,X_3,X_4}(\cdot \mid x_1^{j+1}, x_3^j, x_4^j)$
- Sample $x_3^{j+1} \sim f_{X_3|X_1,X_2,X_4}(\cdot \mid x_1^{j+1}, x_2^{j+1}, x_4^j)$
- Sample $x_4^{j+1} \sim f_{X_4|X_1,X_2,X_3}(\cdot \mid x_1^{j+1}, x_2^{j+1}, x_3^{j+1})$
- $j \leftarrow j + 1$

until *j is sufficiently large*

Thus, the Gibbs sampler cycles repeatedly through the full conditionals, simulating and updating one component of $\boldsymbol{X}$ at each step. Under very general conditions,[4] the random process constructed in this way converges in distribution to $f_{\boldsymbol{X}}$. Gibbs samplers have to be *spun up*, in order to forget the initial value $\boldsymbol{x}^0$, which might be in an improbable or inaccessible part of the realm of $\boldsymbol{X}$; details are given in, e.g., Gelman *et al.* (2003).

[4] But $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ is sufficient.

spun up

This process simulates the whole of $\boldsymbol{X}$; but if only a margin $\boldsymbol{X}_A$ is required, then the rest can be discarded. For simulating $\boldsymbol{X}_{-B}$

conditionally on $(\boldsymbol{X}_B = \boldsymbol{x}_B)$, simply set $\boldsymbol{X}_B \leftarrow \boldsymbol{x}_B$ throughout, and the Gibbs sampler cycles through the full conditionals of $\boldsymbol{X}_{-B}$. As explained in Section 3.2.1, the set of random quantities $\boldsymbol{X}$ may include parameters, if a Bayesian approach has been adopted. Otherwise (Frequentist approach), the Gibbs sampler is used to simulate from $f_{\boldsymbol{X}|\theta}(\cdot \mid t)$ for specified t.

A simple example of a Gibbs sampler including parameters and conditioning is given in Section 3.4.2.

* * *

The Gibbs sampler breaks the problem of simulating all of the uncertain elements of $\boldsymbol{X}$ down into a set of one-dimensional simulations from the full conditionals. Furthermore, when the CIG $\mathcal{G}$ is sparse (i.e. is missing lots of edges) each of these full conditionals can be expressed in terms of the small set of neighbours on $\mathcal{G}$. A careful choice of $\mathcal{G}$ and the G_C functions can result in many of the full conditionals being standard distributions, for which there exist fast algorithms for simulation. Otherwise, there are generic algorithms for simulating a one-dimensional random quantity from a distribution known up to proportionality (see Robert and Casella, 2004). Consequently Gibbs sampling has been packaged into software such as the powerful and ubiquitous BUGS (Bayesian inference Using Gibbs Sampling, see Lunn *et al.*, 2013).

3.3 *Exchangeability*

In many situations $X_1, \ldots, X_m$ represent the same operational definition, applied to m different units; for example, m different children in a classroom. One possibility in this case is to treat the unit labels, the i's on the X_i's, as uninformative. For children in a classroom, where the labels might be names, this would be inappropriate for the class teacher, but appropriate for the Schools Inspector. If the labels are treated as uninformative then the X's are *exchangeable random quantities*.

exchangeable random quantities

Definition 3.5 (Exchangeable random quantities)**.** *$X_1, \ldots, X_m$ is an exchangeable sequence exactly when*

$$f_{X_1,\ldots,X_m}(x_1, \ldots, x_m) = f_{X_1,\ldots,X_m}(x_{\pi_1}, \ldots, x_{\pi_m})$$

where $(\pi_1, \ldots, \pi_m)$ is any permutation of $(1, \ldots, m)$.

If $X_1, \ldots, X_m$ are treated as exchangeable, then there are two ways to represent $f_{X_1,\ldots,X_m}$. The first is to use a mixture of hypergeometric distributions, if m is finite and the realm of the X's is finite.[5] The other method is far more popular. This is to introduce random quantities $\theta \in \Omega$ for which the X's are IID given θ (see (3.2)). Then

[5] See Schervish (1995, ch. ch1) and Lad (1996, ch. 5).

$$f_{X_1,\ldots,X_m}(x_1, \ldots, x_m) = \int \prod_{i=1}^{m} f_{X|\theta}(x_i \mid t)\, \pi_\theta(t)\, \mathrm{d}t \qquad (3.3)$$

for some specified model $f_{X|\theta}$ and specified prior distribution π_θ. Clearly this $f_{X_1,\ldots,X_m}$ is invariant to permutations of the x's, but,

equally clearly, $X_1, \ldots, X_m$ are not mutually independent, because $f_{X_1,\ldots,X_m}$ does not factorise into an m-fold product—this would only occur if π_θ were a *Dirac delta function* $\pi_\theta(t) = \delta(t - \theta_0)$ for some specified θ_0.

Dirac delta function

In (3.3), each choice of statistical model $f_{X|\theta}$ and prior distribution π_θ specifies an exchangeable distribution for $X_1, \ldots, X_m$. *De Finetti's Representation Theorem* asserts that if $X_1, X_2, \ldots$ is an infinite sequence for which every finite sequence is exchangeable then (3.3) is the *only* way to represent the distribution of any finite subset. And furthermore, if $X_1, \ldots, X_m$ is a finite exchangeable sequence but m is large, then there exists a $f_{X|\theta}$ and π_θ for which (3.3) is close to $f_{X_1,\ldots,X_m}$. For this reason, (3.3) is the natural way to represent judgements of exchangeability, except in the case where m is small.

De Finetti's Representation Theorem

The mathematics of the Representation Theorem is hard. The original conception was due to de Finetti (1937). There is a simple half-proof in Heath and Sudderth (1976), a beautiful complete proof in Kingman (1978), and a useful summary in Schervish (1995, ch. 1).

The dominant position of exchangeable distributions in statistical inference is really down to two factors. First, exchangeability represents our qualitative judgement that the elements of $X_1, \ldots, X_m$ are like each other, without insisting that they are identical. More generally, it allows us to introduce element-specific information, but not to have to assert that this information exhausts the systematic differences between one element and another (see Section 3.4).

But this would be only theoretically attractive were it not for the second factor, which is that the representation in (3.3) is so tractable. All it requires is a model $f_{X|\theta}$ and a prior distribution π_θ—given these two, we immediately have a joint distribution over exchangeable $X_1, \ldots, X_m$ for any finite m. As Section 4.5 will discuss, the product form of the statistical model is particularly tractable for approximations.

3.4 *Hierarchical models*

As explained in Section 3.3, if $\boldsymbol{X} := (X_1, \ldots, X_m)$ is IID conditional on some parameter θ, then $\boldsymbol{X}$ is exchangeable. A distribution of this type has two 'levels', typically written as

$$\begin{aligned} X_1 \ldots, X_m \mid \theta &\overset{\text{iid}}{\sim} f_{X|\theta} \\ \theta &\sim \pi_\theta, \end{aligned}$$

see (3.2) and (3.3). The first level is the statistical model, and the second is the prior distribution. This is the simplest example of a *hierarchical model*.

hierarchical model

This simple template can be expanded either upwards or downwards. Upwards, to provide a richer description of the operationally-defined quantities, and downwards to provide a richer description of the parameters. It is easiest to explain this with an illustration.

Suppose that X_i was the number of paid hours worked per week by person i. If all we knew were the X's then we would be forced, through ignorance, to assert that $\boldsymbol{X}$ was exchangeable, if we wanted make inferences about unobserved X's on the basis of a sample of observed ones. In this case we would necessarily make the same inference about every unobserved X.

We would like to do much more than this, and make inferences that differ, one person to another, in the same way that people differ one to another. Typically we have additional information about each person, in the form of covariate information, $\boldsymbol{v}_i$ say, a p-vector of relevant information such as gender, age, family size, education, and so on, for each i. By incorporating this information we can make inferences about policy-relevant quantities, such as whether there is a difference between male and female working practices, after allowing for other factors.[6]

[6] Although it is obligatory to insert the warning that causal inference from observational data is hard; see Pearl (2000).

A simple hierarchical model for $\{X_1, \boldsymbol{v}_1\}, \ldots, \{X_m, \boldsymbol{v}_m\}$ might be expressed in terms of additional random quantities $\boldsymbol{B} := \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m\}$ and $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_p)$, in the general form

$$f_{\boldsymbol{X},\boldsymbol{B},\boldsymbol{\mu}}(\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{m}) = f_{\boldsymbol{X}|\boldsymbol{B},\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{b}, \boldsymbol{m})\, f_{\boldsymbol{B}|\boldsymbol{\mu}}(\boldsymbol{b} \mid \boldsymbol{m})\, f_{\boldsymbol{\mu}}(\boldsymbol{m}) \tag{3.4}$$

which is always valid, according to the factorisation theorem (Theorem 1.17). A natural implementation of the righthand side of (3.4) for the illustration would be

$$\begin{aligned} \boldsymbol{X} \mid \boldsymbol{B}, \boldsymbol{\mu} &\overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{v}_i^T \boldsymbol{\beta}_i, s_x^2) \\ \boldsymbol{B} \mid \boldsymbol{\mu} &\overset{\text{iid}}{\sim} \mathrm{N}_p(\boldsymbol{\mu}, S_\beta) \\ \boldsymbol{\mu} &\sim \mathrm{N}_p(\boldsymbol{0}, S_\mu), \end{aligned} \tag{3.5}$$

where N_p indicates a p-dimensional Normal distribution with specified expectation and variance. $\boldsymbol{X}$ are operationally-defined quantities, at the top of the hierarchy; $\boldsymbol{B}$ are the parameters; and $\boldsymbol{\mu}$ are termed *hyperparameters*, because they are parameters in the distribution of the parameters. For the moment, treat the three variance terms s_x, S_β, and S_μ as specified.

hyperparameters

The generality of this type of hierarchical model can be appreciated by considering the special case where $S_\beta = \boldsymbol{0}$, which implies that $\boldsymbol{\beta}_i = \boldsymbol{\mu}$ for all i. This is known as a *regression analysis*. In a regression analysis, the only difference between two people with the same covariates would be in the residual term, which has variance s_x^2. If we think there might be more systematic differences between these two people, then this is captured by $S_\beta > \boldsymbol{0}$, which allows two different people to have different $\boldsymbol{\beta}$'s. But the attraction of the regression analysis is that there are only p (hyper) parameters; while the hierarchical model has mp parameters and p hyperparameters, when there are only m observations.

regression analysis

3.4.1* *Conditional independence in hierarchical models*

Hierarchical models such as (3.5) are extremely common in modern statistical inference. They are typically treated using the Gibbs

sampler, and for this we need to know their CIG. This involves a two-stage procedure. First, the hierarchical model is represented in terms of its directed acyclic graph (DAG, see immediately below), and then this DAG is transformed into its CIG. The construction of the DAG is very similar to that of Section 3.2, except that the conditioning is different.

Let $\boldsymbol{X} := (X_1, \ldots, X_m)$ be a set of random quantities in a prescribed ordering. Then the factorisation theorem (Theorem 1.17) allows us to write

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) \prod_{i=2}^{m} f_{X_i|X_1,\ldots,X_{i-1}}(x_i \mid x_1, \ldots, x_{i-1}).$$

It will be convenient below to write $\boldsymbol{X}^{1:i} := (X_1, \ldots, X_i)$ and $\boldsymbol{X}^{1:i}_{-j} := (\boldsymbol{X}^{1:i})_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_i)$. In this notation the factorisation is written

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) \prod_{i=2}^{m} f_{X_i|\boldsymbol{X}^{1:(i-1)}}(x_i \mid x^{1:(i-1)}).$$

Now consider any one of these conditional distributions,

$$f_{X_i|\boldsymbol{X}^{1:(i-1)}}(x_i \mid x^{1:(i-1)})$$

If this probability distribution is invariant to x_j, where $j < i$, then $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}^{1:(i-1)}_{-j}$, according to Theorem 3.1. If this probability distribution is not invariant to x_j then write $X_j \to X_i$ to denote "X_i is a child of X_j". Were we to perform this operation for every pair of random quantities, $m(m-1)/2$ judgements in all, we could construct a directed graph, termed a *directed acyclic graph (DAG)*. The absence of an edge from X_i to X_j would either be because $j \geq i$, or because $j < i$ and $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}^{1:(i-1)}_{-j}$.

directed acyclic graph (DAG)

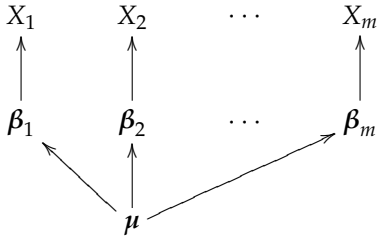
For the model in (3.5), the natural ordering (from the bottom) is $\boldsymbol{\mu}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, X_1, \ldots, X_m$. From the middle level,

$$\boldsymbol{\beta}_i \perp\!\!\!\perp \boldsymbol{\beta}^{1:(i-1)} \mid \boldsymbol{\mu} \qquad i = 2, \ldots, m.$$

From the top level,

$$\begin{aligned} X_1 &\perp\!\!\!\perp \boldsymbol{\beta}^{2:m}, \boldsymbol{\mu} \mid \boldsymbol{\beta}_1 \\ X_i &\perp\!\!\!\perp \boldsymbol{X}^{1:(i-1)}, \boldsymbol{\beta}^{1:m}_{-i}, \boldsymbol{\mu} \mid \boldsymbol{\beta}_i \quad i = 2, \ldots, m. \end{aligned}$$

Thus the DAG for (3.5) is simply



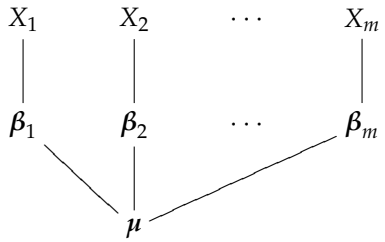
(†)

A DAG is not a CIG, because the conditional distributions in the factorisation are not full conditionals, but conditionals taken sequentially along an ordering of the random quantities. But there is

a simple algorithm to convert a DAG into a CIG. First, 'marry' any pairs of vertices that have a common child, by joining them with an undirected edge. Second, turn all directed edges into undirected edges. This procedure is quaintly known as *moralising* the DAG, and the resulting *moral graph* of the DAG is its CIG. Cowell *et al.* (1999, App. B) justifies this algorithm (it is quite intuitive).

moral graph

The moral graph for (3.5) is the same as its DAG except without the arrows, because there are no pairs of vertices with a common child:



(‡)

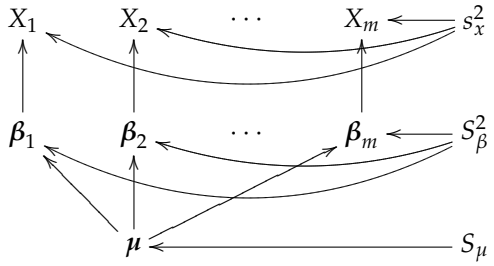
Using the global Markov property (Definition 3.4) we can immediately infer from (‡) that the X's are mutually conditionally independent given $\boldsymbol{\mu}$, because every path from X_i to X_j passes through $\boldsymbol{\mu}$. Thus

$$f_{\boldsymbol{X}|\boldsymbol{\mu}}(\boldsymbol{x} \mid \boldsymbol{m}) = \prod_{i=1}^{m} f_{X_i|\boldsymbol{\mu}}(x_i \mid \boldsymbol{m}). \tag{3.6}$$

This factorisaton property has two advantages. First, the distribution $f_{\boldsymbol{X}_A|\boldsymbol{\mu}}$ for any set A can be found very quickly, without having to jointly sum over $\mathcal{X}_{-A}$. Generally, it is the hyperparameters ($\boldsymbol{\mu}$ in this case) which describe the population, and so this type of model is extremely effective for making inferences about a population based on a sample, A. Second, the product structure of $f_{\boldsymbol{X}_A|\boldsymbol{\mu}}$ is very attractive computationally, as discussed further in Section 4.5. Note that it may not be possible to evaluate $f_{X_i|\boldsymbol{\mu}}(x_i \mid \boldsymbol{m})$, but this does not matter: it is the product structure itself that is important.

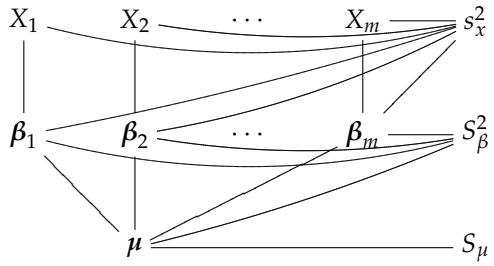
* * *

So far we have treated the three variances s_x^2, S_β and S_μ as specified. What happens if we treat them as random quantities (i.e. as additional hyperparameters) and prepend them to $\boldsymbol{\mu}$, treating all four as mutually independent? The DAG would become



Now there are many pairs of vertices that share a common child,

and so the CIG acquires some extra edges after moralisation:



The main reason to draw this CIG is to make the point that additional hyperparameters that are common across the components of the model that vary with i induce lots of extra edges, and extra complication in the sampling. But it is also interesting to note that the X's are still mutually conditionally independent given the (augmented) hyperparameters, and so the advantages of (3.6) still hold.

3.4.2 Gibbs sampling for DAGs*

Here is the Gibbs sampler for models with the same DAG as (†) on p. 59. Suppose that the first n components of $\boldsymbol{X}$ have been observed, and write, therefore, $\boldsymbol{X} := (\boldsymbol{y}^{\text{obs}}, \boldsymbol{Z})$ where $\boldsymbol{y}^{\text{obs}} := (y_1^{\text{obs}}, \ldots, y_n^{\text{obs}})$ and $Z_i := X_{n+i}$ for $i = 1, \ldots, m-n$. The intention is to sample from $f_{\boldsymbol{Z}|\boldsymbol{X}^{1:n}}(\cdot \mid \boldsymbol{y}^{\text{obs}})$, and to do this we sample from

$$f_{\boldsymbol{Z},\boldsymbol{B},\boldsymbol{\mu}|\boldsymbol{X}^{1:n}}(\cdot \mid \boldsymbol{y}^{\text{obs}})$$

and then ignore the $\boldsymbol{B}$ and $\boldsymbol{\mu}$ components.

initialise
  $\{\boldsymbol{z}^0, \boldsymbol{B}^0, \boldsymbol{m}^0\} \leftarrow \{\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}\}$
  $\boldsymbol{x}^0 \leftarrow (\boldsymbol{y}^{\text{obs}}, \boldsymbol{z}^0)$
  $j \leftarrow 0$
repeat
  Sample $\boldsymbol{\mu}^{j+1} \sim f_{\boldsymbol{\mu}|\boldsymbol{B}}(\cdot \mid \boldsymbol{B}^j)$
  for $i \leftarrow 1$ **to** m **do**
    Sample $\boldsymbol{\beta}_i^{j+1} \sim f_{\boldsymbol{\beta}_i|X_i,\boldsymbol{\mu}}(\cdot \mid x_i^j, \boldsymbol{\mu}^{j+1})$
  end
  for $i \leftarrow 1$ **to** $m-n$ **do**
    Sample $z_i^{j+1} \sim f_{X_{n+i}|\boldsymbol{\beta}_{n+i}}(\cdot \mid \boldsymbol{\beta}_{n+i}^{j+1})$
  end
  $\boldsymbol{x}^{j+1} \leftarrow (\boldsymbol{y}^{\text{obs}}, \boldsymbol{z}^{j+1})$
  $j \leftarrow j+1$
until *j is sufficiently large*

In the initialisation I have used $\boldsymbol{0}$ for all random quantities. The constraint is that the intial point must have positive probability conditional on $(\boldsymbol{X}^{1:n} = \boldsymbol{y}^{\text{obs}})$, but a sensible choice is helpful here, to reduce the time spent spinning up.

The sampling distributions can all be inferred from the model. In the case of (3.5), an experienced statistician would see immediately that $f_{\boldsymbol{\mu}|\boldsymbol{B}}$ is multivariate Normal, because $(\boldsymbol{\mu}, \boldsymbol{B})$ is multivariate Normal, and $f_{\boldsymbol{\beta}_i|X_i,\boldsymbol{\mu}}$ is multivariate Normal because $(\boldsymbol{\beta}_i, X_i) \mid \boldsymbol{\mu}$ is multivariate Normal. So all three of the sampling distributions in this case are multivariate Normal—this is an extremely tractable model! Things would get messier with the inclusion of s_x^2, S_β, and S_μ, but there are tractable choices for the marginal distributions of these too (inverse Gamma for s_x^2 and and inverse Wishart for S_β and S_μ).

4
Bayesian inference

The complete class theorems outlined in Section 2.2.2 state that being a Bayes rule or a generalised Bayes rule is a necessary condition for a decision rule to be admissible—precise results for the finite case were given in Theorem 2.3 and Theorem 2.5. And the Bayes Rule Theorem (Theorem 2.4) states that the Bayes rule minimises the expected loss with respect to the conditional probability distribution of the 'state of nature' given the values of the observations.

Bayesian inference is about this distribution: how it is formulated (Section 4.1), computed (Section 4.2), summarised and communicated (Section 4.3), checked (Section 4.4), and some of its interesting and useful properties (Section 4.5).

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

4.1 Prior, posterior, and predictive distributions

The object of Bayesian inference is to infer a *predictive distribution* predictive distribution

$$f_{X|Y}(x \mid y) := \mathsf{Pr}(X = x \mid Y = y)$$

where X are the random quantities which are influential on the outcome of a decision (the 'state of nature' in Chapter 2) , and Y are observations relevant to X. I will refer to X as the *predictands* and Y as the *observations*. predictands observations

From a formal Bayesian point of view, the client, assisted by the statistician and her experts, either specifies $f_{X|Y}$ directly, or, if it is easier, specifies $f_{X,Y}$, possibly in the form of $f_{Y|X}$ and f_X. In this second case, where $f_{X|Y}$ is not specified directly, the statistician can compute $f_{X|Y}$ from the given distribution(s), using Bayes's Theorem (Theorem 1.19).

In practice, however, none of these distributions is specified directly, but indirectly through the introduction of statistical parameters $\theta \in \Omega$, where I will take Ω to be a convex subset of $\mathbb{R}^d$. The purpose of introducing statistical parameters is to simplify the specification of the joint distribution $f_{X,Y}$; see Section 2.5 and Chapter 3. Statistical parameters are not required, but experience suggests that they are invaluable.

Having introduced statistical parameters, the client's joint distribution over $\{X, Y, \theta\}$ is expressed as

$$f_{X,Y,\theta}(x,y,t)\,\mathrm{d}t = f_{X,Y|\theta}(x,y \mid t)\,\pi_\theta(t)\,\mathrm{d}t \tag{4.1}$$

where both distributions on the righthand side must be specified. $f_{X,Y|\theta}$ is termed the *statistical model* and π_θ the *prior distribution*. Eq. (4.1) is an example of the factorisation theorem (Theorem 1.17): the statistician and his client find it easier to specify a joint distribution for X and Y by specifying a conditional distribution $f_{X,Y|\theta}$ and a marginal distribution π_θ; if θ is not needed in the decision analysis it can be integrated out.

statistical model
prior distribution

From the statistical model and the prior distribution we can immediately deduce the *marginal distribution* of the observations

marginal distribution

$$f_Y(y) = \sum_x \int f_{X,Y,\theta}(x,y,t)\,\mathrm{d}t, \tag{4.2a}$$

the *predictive distribution* (of the predictands)

predictive distribution

$$f_{X|Y}(x \mid y) = \int \frac{f_{X,Y,\theta}(x,y,t)}{f_Y(y)}\,\mathrm{d}t, \tag{4.2b}$$

and, if we need it, the *posterior distribution* of the parameters

posterior distribution

$$f_{\theta|Y}(t \mid y)\,\mathrm{d}t = \sum_x \frac{f_{X,Y,\theta}(x,y,t)}{f_Y(y)}\,\mathrm{d}t. \tag{4.2c}$$

These distributions are expressed for arbitrary y, and as such they represent a pre-data analysis. In a post-data analysis the statistician will have access to actual observations, denoted y^{obs}. For these actual observations I write the posterior distribution as

$$\pi^*_\theta(t)\,\mathrm{d}t := f_{\theta|Y}(t \mid y^{\mathrm{obs}})\,\mathrm{d}t.$$

The value of the marginal likelihood is termed the *evidence*,

evidence

$$\text{evidence} := f_Y(y^{\mathrm{obs}}),$$

and $-2\log(\text{evidence})$ is termed the *deviance*.

deviance

One very common statistical modelling approach is to specify a θ for which the client judges that X and Y are *conditionally independent* given θ. In this case the statistical model factorises as

conditionally independent

$$f_{X,Y|\theta}(x,y \mid t) = f_{Y|\theta}(y \mid t)\,f_{X|\theta}(x \mid t);$$

see Section 3.4. This has the following simplifying effects:

$$f_Y(y) = \int f_{Y|\theta}(y \mid t)\,\pi_\theta(t)\,\mathrm{d}t \tag{4.3a}$$

$$f_{\theta|Y}(t \mid y)\,\mathrm{d}t = \frac{f_{Y|\theta}(y \mid t)\,\pi_\theta(t)}{f_Y(y)}\,\mathrm{d}t \tag{4.3b}$$

$$f_{X|Y}(x \mid y) = \int f_{X|\theta}(x \mid t)\,f_{\theta|Y}(t \mid y)\,\mathrm{d}t. \tag{4.3c}$$

4.2 *Bayesian computation*

Ultimately, the objective of a Bayesian analysis is to compute the expected loss conditional on the hypothetical (pre-data) or actual (post-data) observations. Formally, this requires a lot of summing and integrating:

$$\begin{aligned}\mathsf{E}^*\left\{L(a,X)\right\} &= \sum_x L(a,x)\, f_X^*(x) \\ &= \sum_x L(a,x)\, f_{X|Y}(x \mid y^{\text{obs}}) \\ &= \sum_x L(a,x) \int \frac{f_{X,Y,\theta}(x, y^{\text{obs}}, t)}{f_Y(y^{\text{obs}})}\, \mathrm{d}t \\ &= \sum_x L(a,x) \frac{\int f_{X,Y,\theta}(x, y^{\text{obs}}, t)}{\sum_{x'} \int f_{X,Y,\theta}(x', y^{\text{obs}}, t')\, \mathrm{d}t'}\, \mathrm{d}t.\end{aligned}$$

But, as noted in Section 3.2, in many cases it is not possible to evaluate $f_{X,Y,\theta}(x, y^{\text{obs}}, t)$ explicitly, and even if it is possible, then the sums and integrals may be over huge spaces, and completely infeasible. As a result, only the smallest and most tractable inferences are now treated formally; everything else is done using simulation, and *sequential simulation* in particular.

sequential simulation

At the heart of a sequential simulation is the notion of what I will call a *sufficiently random sequence (SRS)*. In the definition below I write the sequence with its index in the superscript, because X will often be a vector, and I might want to write X_j^i to be the jth component of the ith element of the sequence.[1]

sufficiently random sequence (SRS)

[1] For the following definition, recollect from Section 1.4 that $(X^i = x)$ is a random proposition. If necessary, consult Section A.4 for the definition of $\xrightarrow{\text{P}}$, 'convergence in probability'.

Definition 4.1 (Sufficiently random sequence, SRS)**.** *$X^1, X^2, \ldots$ is a sufficiently random sequence for f_X exactly when*

$$m^{-1} \sum_{i=1}^{m} (X^i = x) \xrightarrow{P} f_X(x) \quad \textit{for all } x \in \mathcal{X}. \tag{4.4}$$

The value of an SRS is that it can be used to compute the expectation of any real-valued function of X to arbitrary accuracy, subject only to having enough CPU cycles.

Theorem 4.1. *If $X^1, X^2, \ldots$ is an SRS for f_X then*

$$m^{-1} \sum_{i=1}^{m} g(X^i) \xrightarrow{P} \mathsf{E}\{g(X)\} \quad \textit{for all real-valued } g.$$

Proof. For fixed m,

$$m^{-1} \sum_{i=1}^{m} g(X^i) = m^{-1} \sum_{i=1}^{m} \sum_x g(x)(X^i = x) = \sum_x g(x) \Big\{ m^{-1} \sum_{i=1}^{m} (X^i = x) \Big\}.$$

And then, because $X^1, X^2, \ldots$ is an SRS,

$$\sum_x g(x) \Big\{ m^{-1} \sum_{i=1}^{m} (X^i = x) \Big\} \xrightarrow{\text{P}} \sum_x g(x) f_X(x) = \mathsf{E}\{g(X)\},$$

using Theorem A.10. □

To understand how an SRS is used, consider the post-data decision analysis in which $Y = y^{\text{obs}}$, and now a Bayes action from $\mathcal{A}$ must be chosen. Suppose the statistician can generate a long SRS for $f^*_X(\cdot) := f_{X|Y}(\cdot \mid y^{\text{obs}})$. From this one sequence he can compute the expected posterior loss $\mathsf{E}^*\{L(a, X)\}$ for each action $a \in \mathcal{A}$, by taking the sample mean of $L(a, X)$ over the SRS for f^*_X. Hence he can identify a Bayes action for the client. So the main computing preoccupation of the statistician is to generate a long enough SRS for f^*_X: after that, everything is straightforward. In practice, the statistician will typically generate an SRS for $f^*_{X,\theta}$, and then simply ignore the θ component, there being no θ in the loss function.

The simplest type of SRS is an independent random sample. This is an SRS according to the Weak Law of Large Numbers (see Theorem A.9), because the elements of an independent random sequence are mutually uncorrelated.

The most popular approach to generating an SRS is *Markov chain Monte Carlo (MCMC)*; see for example, Besag *et al.* (1995), Robert and Casella (2004), Gelman *et al.* (2003), or Lunn *et al.* (2013). Gibbs sampling, mentioned in Chapter 3, is an example of an MCMC algorithm. Gibbs sampling takes advantage of the conditional independence structure of the collection $\{X, Y, \theta\}$ to break the problem of generating an SRS for $f^*_{X,\theta}$ down into a sequence of much simpler sampling tasks.

Markov chain Monte Carlo (MCMC)

Sometimes it is helpful to have individual candidates from an SRS, which behave like an independent random sample. In this case these candidates need to be well-spaced in the sequence to minimise the effect of autocorrelation. 'Thinning' the SRS is a possibility: this involves keeping only, say, every hundredth element (or some other number sufficiently large that the autocorrelation coefficient after thinning is small, say less than 0.05). Thinning is never recommended for computing expectations (since the discarded elements still contain relevant information), but might be necessary if computer memory is limited, or if a short SRS is more helpful to the client.

Finally, an obvious point, but an important one: the proportionate histogram of an SRS for $f^*_{X,\theta}$ approximates the posterior distribution of $\{X, \theta\}$. This is because, for any $A \subset \mathcal{X} \times \Omega$,

$$m^{-1} \sum_{i=1}^{m} (\{X^i, \theta^i\} \in A) \xrightarrow{\text{P}} \mathsf{E}^* \left\{ (\{X, \theta\} \in A) \right\} = \mathsf{Pr}^*(\{X, \theta\} \in A)$$

if $\{X^i, \theta^i\}$ is an SRS for $f^*_{X,\theta}$. Again, it would be better not to thin the SRS before computing the histogram.

4.2.1 Computing the evidence from an SRS*

As discussed in Section 2.6.1, admissible methods for choosing between hypotheses require the value of the 'evidence', $f_Y(y^{\text{obs}})$, for each of the competing hypotheses. An approximation to this value is given in Section 4.5.3. Here, I outline how an arbitrarily

exact value can be computed from an SRS for π_θ^*, the posterior distribution, using an approach suggested by Gelfand and Dey (1994).

Let L be the likelihood function, $L(t) := f_{Y|\theta}(y^{\text{obs}} \mid t)$. Note that, according to Bayes's Theorem

$$\{f_Y(y^{\text{obs}})\}^{-1} = \frac{\pi_\theta^*(t)}{L(t)\,\pi_\theta(t)} \quad \text{for any } t \in \Omega.$$

Therefore, if h is any proper distribution on Ω,

$$\{f_Y(y^{\text{obs}})\}^{-1} = \int \frac{h(t)}{L(t)\pi_\theta(t)}\pi^*(t)\,\mathrm{d}t = \mathsf{E}^*\left\{\frac{h(\theta)}{L(\theta)\pi_\theta(\theta)}\right\},$$

where E^* indicates the expectation with respect to the posterior distribution. Consequently Gelfand and Dey suggest the estimator

$$\widehat{f_Y}(y^{\text{obs}}) := \left\{ m^{-1} \sum_{j=1}^{m} \frac{h(\theta^j)}{L(\theta^j)\pi_\theta(\theta^j)} \right\}^{-1}$$

where $\theta^1, \theta^2, \ldots$ is an SRS for π_θ^*.

The distribution h can be chosen to approximate the posterior distribution π_θ^*. Such a choice is helpful for stabilising the terms in the sum, and reducing the variability of the estimator, which means that a smaller m is required for a given accuracy. The obvious choice in this case is the Normal approximation given in (4.5) below.

4.3 Summarising distributions*

In a modern post-data Bayesian analysis, the physical outcome of the inference is an SRS for $f_{X,\theta|Y}(\cdot \mid y^{\text{obs}})$. Think of this as a spreadsheet of numbers, with one column for each component of X and θ, and one row for each element of the SRS. We might denote this table as the matrix T. It would not be unusual for T to have a thousand or a million rows. Naturally, the statistician will think about summarising T in some way, if possible. He may indeed be under pressure to produce very short summaries, for example when the Minister says "Don't bamboozle me with a spreadsheet, just give me a number!"

However, the statistician must be careful. If there is a chance that the summary will subsequently be used in place of T, he must take care not to suppress any information in the columns of T which could affect the choice of actions. And, in the absence of an explicit action set $\mathcal{A}$ and loss function L, he would be wise not to suppress any information at all. The only compression of T that is completely acceptable is to ignore the ordering of the rows in T, because any permutation of an SRS is also an SRS.

In the absence of any specific action set and loss function, the statistician ought to be sensitivity to two properties of serious decision analysis (i.e. the kind of analysis where the client pays

serious money to hire an applied statistician). First, losses tend to depend jointly on several predictands. For example, an industrial accident tends to occur when several things go wrong at the same time. Weather hazards tend to depend jointly on sequences of temperature and precipitation (see, e.g., Edwards and Challenor, 2013). Our future climate depends on economics, demographics, and technology. Therefore low-dimensional summaries, focusing on one or two predictands, may misrepresent the distribution of the loss.

Second, the loss function is often convex at the top end, meaning that extreme outcomes of some of the predictands result in disproportionate losses. So the loss from, say, an outcome for X at the 95th percentile can be far less severe than one at the 99th percentile. In addition, in many natural hazards like earthquakes, floods, and volcanoes, the observed frequency/magnitude relationship is well fitted by a power law (see, e.g. Woo, 2011; Rougier *et al.*, 2013). The 95th percentile for the magnitude of an earthquake is *much* smaller than the 99th percentile. Summarising X by its 95th percentile seems very misleading, if the actions are about managing losses.

Therefore, the statistician should resist simple summaries of his SRS unless it is clear from the client's action set and loss function that (i) only one or two of the predictands are involved in the loss function, and (ii) the loss function is approximately linear.

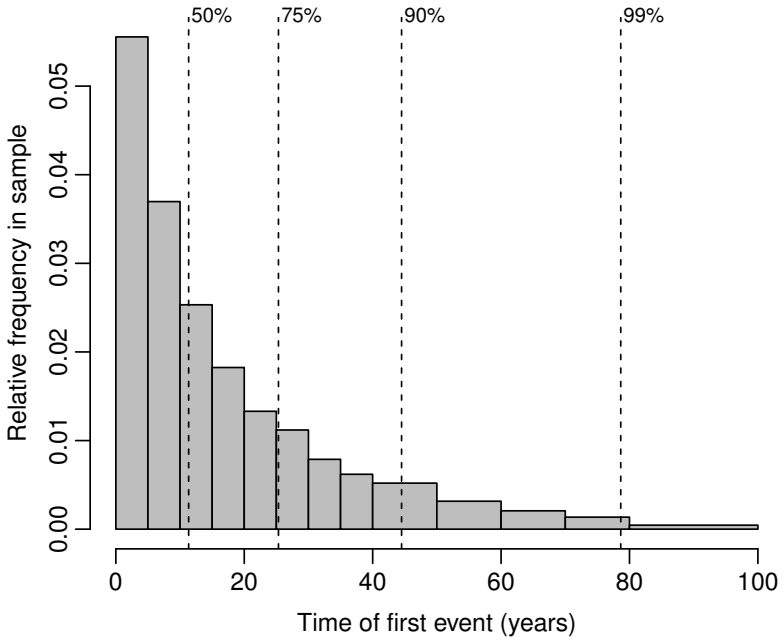


Figure 4.1: Histogram of the arrival times of the first event in a generalised Poisson process, with some quantiles indicated.

In the case where only one predictand is important the statistician might provide a histogram of the SRS for that predictand, with helpful quantiles indicated (these can be hard to assess from a histogram).[2] See Figure 4.1. If the loss function is approximately linear in the predictand he might provide a simple interval summary, such as the 5th and 95th percentiles—this is termed a 90%

[2] R users should use the `truehist` function from the `MASS` package, see Venables and Ripley (2002, sec. 5.3).

equi-tailed *credible interval*.

credible interval

In the case where only two predictands are important he might provide a *scatter plot* of the SRS, with one predictand on each axis. Rather than showing the individual points, this could be coloured by the density of points, but my preference would be to use *convex peeling* to colour the plot using helpful percentiles, with extreme outliers shown individually: see Figure 4.2 for an example. This approach works best when the cloud of points is roughly elliptical, and this might be improved by a transformation.

convex peeling

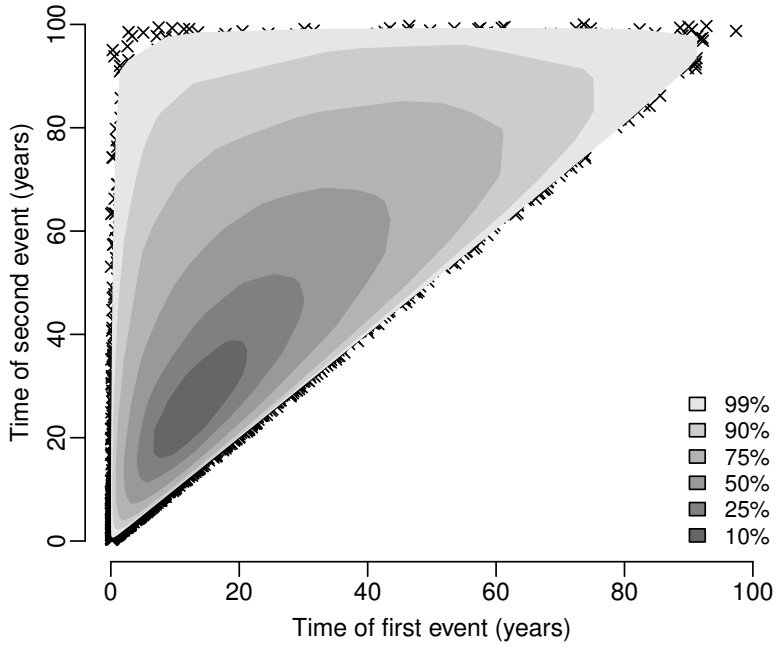


Figure 4.2: Convex peeling to show quantiles of two random quantities. In this case, these are the arrival times of the first two events in a generalised Poisson process.

* * *

Clearly, this section is not the conventional "here's the definition of a credible set, here's the definition of a high posterior density region" section that might be anticipated in lecture notes—for this you should consult, e.g., Tanner (1996). It is a plea that we recognise the complexity of modern decision analysis, and don't over-simplify. Obviously the client is busy, and I expect she would prefer the statistical inference to come in a small and tidy package; perhaps with a colourful image for the front of the report. But if she is serious about making the right decision then she will understand the need to think hard about her action set and her loss function, and what they imply for effective summaries of the predictands.

4.4 *Visualisation and diagnostics*

A model is a device for organising our knowledge and judgements. This is true of all models, including statistical models, which are a device for organising the knowledge and judgement of the client and her experts, including her statistician. Thus the client and her

experts should concur that the distribution for $\{X, Y\}$

$$f_{X,Y}(x, y) = \int f_{X,Y|\theta}(x, y \mid t)\, \pi_\theta(t)\, \mathrm{d}t$$

is a reasonable reflection of their knowledge and judgements.[3] As shown, this distribution usually arises as an implication of the choice of a statistical model for $\{X, Y\}$ and a prior distribution for the parameter θ. Often θ can be rather abstract, which makes specifying π_θ challenging; but X and Y are operationally defined and Y is observed.

[3] There is always a difficulty in representing the collective judgement of a group of people. According to Prof. Willy Aspinall, who is very experienced in group elicitation (Aspinall, 2010; Aspinall and Cooke, 2012), a common response from an individual expert is that while the collective assessment $f_{X,Y}(x, y)$ would not represent his personal judgement, he would not disagree with it as a representation of the judgement of a well-informed person.

The purpose of visualisation and diagnostics is for the client and her experts to satisfy themselves (and the auditor, in the narrative of Chapter 2) that their distribution for $\{X, Y\}$ is reasonable, by comparing the marginal distribution of Y with the value of the observations, y^{obs}. There are several approaches, some of which are outlined below.

In each approach the objective is to compare the actual observations with a distribution implied by the statistical model and the prior distribution. In most cases the distribution will be most easily represented as an SRS (Section 4.2). Candidates drawn from the SRS can be used in a *Turing test*. A number (say 15) of candidates are randomly drawn from the SRS. These candidates plus y^{obs} are shuffled, and all 16 are displayed to the client and her experts, in whatever format seems most appropriate (e.g. as a map, as a time-series, as a table). If the statistical model and the prior distribution are an adequate representation of Y then it will be hard to spot y^{obs} among the candidates. Statistical summaries can also be computed (Chapter 6).

Turing test

1. *Prior predictive* diagnostics, in which the value y^{obs} is compared with the marginal distribution f_Y (Box, 1980). However, this approach tends to be uninformative when π_θ is diffuse, and is not applicable at all if π_θ is improper.

   Prior predictive

2. *cross-validation* diagnostics (Stone, 1974). *Leave-one-out (LOO)* is the simplest and most popular approach: each y_i is compared with its predictive distribution based on the other $n - 1$ observations. This is the gold-standard but it is expensive because the inference needs to be repeated n times. Generalisations tend to go by the name *k-fold cross validation*, where the observations are divided into k subsets, and one is used to update the predictive distribution for the other $k - 1$ (so LOO is n-fold cross validation). If the y_i's are ordered, then *prequential diagnostics* are another variant (Dawid, 1984; Cowell *et al.*, 1999).

   cross-validation
   Leave-one-out (LOO)

   k-fold cross validation

   prequential diagnostics

3. *posterior predictive* diagnostics Rubin (1984).[4] In this approach the observations are 'cloned' into an identical set Y^{pre}, defined by the joint distribution

   posterior predictive

   [4] See also Meng (1994) and Gelman *et al.* (2003, ch. 6).

   $$f_{Y^{\text{pre}},Y}(y^{\text{pre}}, y) = \int f_{Y|\theta}(y^{\text{pre}} \mid t)\, f_{Y|\theta}(y \mid t)\, \pi_\theta(t)\, \mathrm{d}t.$$

   In other words, Y^{pre} is a set of observations that we might have got had we been able to duplicate the experiment represented by

the statistical model. An application of Bayes's Theorem gives

$$f^*_{Y^{\text{pre}}}(\cdot) := f_{Y^{\text{pre}}|Y}(\cdot \mid y^{\text{obs}}) = \int f_{Y|\theta}(\cdot \mid t)\, \pi^*_\theta(t)\, \mathrm{d}t$$

where π^*_θ is the posterior distribution, as usual. This distribution describes the set of candidates for Y that the statistical model and the prior distribution imply as reasonable alternatives to the actual value y^{obs}.

There is no equivalent check for $f_{X,Y}$, because X is not observed. But candidates can be randomly sampled from the posterior predictive distribution $f^*_X(\cdot) := f_{X|Y}(\cdot \mid y^{\text{obs}})$, and the client and her experts can at least satisfy themselves that the individual candidates look plausible, and that summary measures of the joint distribution are consistent with their judgements. Ideally, X and Y would be judged sufficiently alike that successfully representing Y will be taken as confirmatory of the joint distribution over $\{X, Y\}$.

4.5 *Bayesian asymptotics*

Consider the task of assessing the posterior distribution of the parameters,

$$\pi^*_\theta(t)\, \mathrm{d}t := \mathsf{Pr}(\theta \in [t, t + \mathrm{d}t) \mid \boldsymbol{Y} = \boldsymbol{y}^{\text{obs}})$$

where $\theta \in \Omega$ and Ω is a convex subset of $\mathbb{R}^d$, and $\boldsymbol{Y} := (Y_1, \ldots, Y_n)$. For a decision analysis, interest resides in the predictands rather than the parameters. But a direct interest in the posterior distribution arises in three situations:

1. When θ is decision-relevant in its own right (in which case we can take θ and X to be synonymous).
2. When X and $\boldsymbol{Y}$ are treated as conditionally independent given θ; see (4.3) and Section 4.5.2.
3. For posterior predictive diagnostics; see Section 4.4.

This section considers the case where n, the number of observations, is large, the so-called *asymptotic* case. The Bayesian approach has no particular need of asymptotic arguments. But it can be useful to have a tractable approximation to the posterior distribution (e.g. to approximate the evidence, see Section 4.2.1), and asymptotic arguments can provide this. Asymptotics also address an interesting regularity, which is that very often the margins of a posterior distribution appear to be approximately Normal, and not to depend on the prior distribution.

asymptotic

First, a quick review of the *Normal distribution*. Let $\boldsymbol{X} \in \mathbb{R}^m$ have a multivariate Normal distribution with expectation μ and non-singular variance matrix Σ, denoted $\boldsymbol{X} \sim \mathrm{N}_m(\mu, \Sigma)$. The PDF of $\boldsymbol{X}$ is

Normal distribution

$$\phi_m(\boldsymbol{x}; \mu, \Sigma) := |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x} - \mu)^T \Sigma^{-1} (\boldsymbol{x} - \mu)\right\}.$$

Let $\boldsymbol{Y} := a + B\boldsymbol{X}$, where a and B are specified; a is an n-vector and B is an $n \times m$ matrix. Then $\boldsymbol{Y} \sim \mathrm{N}_n(a + B\mu, B\Sigma B^T)$ provided that $B\Sigma B^T$ is non-singular. Full details are given in Mardia *et al.* (1979, ch. 3).

The formal result for the asymptotic behaviour of the posterior distribution is as follows (the terms will be defined below).

Theorem 4.2 (Asymptotic Normality of the posterior distribution)**.** *Suppose that* $(Y_1, \ldots, Y_n) := \boldsymbol{Y}_n \sim f_{\boldsymbol{Y}_n|\theta}(\cdot \mid \theta_0)$ *for some* θ_0 *in the interior of* Ω*. Then, under suitable regularity conditions on the behaviour of* $f_{\boldsymbol{Y}_n|\theta}$ *with respect to* n *and the choice of prior distribution,* $\hat{\theta}(\boldsymbol{Y}_n) \xrightarrow{P} \theta_0$*,* $\hat{J}(\boldsymbol{Y}_n)$ *increases without limit, and*

$$\hat{J}(\boldsymbol{Y}_n)^{\frac{1}{2}}\{\theta - \hat{\theta}(\boldsymbol{Y}_n)\} \xrightarrow{D} N_d(0, I)$$

as n *increases. Here* $\hat{\theta}$ *is the maximum likelihood estimator and* $\hat{J}$ *is the observed Fisher Information matrix.*

This is a statement about the random quantities on the lefthand side, which is a d-vector involving both observations and parameters. But by conditioning on the observations and rearranging for θ we get, under the same regularity conditions,

$$\pi^*_\theta(t) \approx \phi_d\big(t; \hat{\theta}(\boldsymbol{y}_n^{\mathrm{obs}}), \hat{J}(\boldsymbol{y}_n^{\mathrm{obs}})^{-1}\big) \tag{4.5}$$

for large n; i.e. the posterior distribution is approximately multivariate normal, and concentrates around the 'true' value θ_0, no matter where in the interior of Ω this value happens to be.

In the asymptotic limit, the prior distribution drops out of the posterior distribution. Thus two statisticians who agree on $f_{\boldsymbol{Y}_n|\theta}$ will also agree, *a priori*, that their posterior distributions will be more-or-less the same regardless of their choices of prior distributions, for a sufficiently large number of observations.

4.5.1 More details*

The proof of Theorem 4.2 is very technical, and there are several versions of the regularity conditions. Here I will reproduce the heuristic argument of Bernardo and Smith (2000, sec. 5.3.2), which is insightful about the interplay of the statistical model and the prior distribution. The crucial feature (assumed implicitly above) is that it must be possible to increase the number of observations without increasing the number of parameters.

In general, Bayes's Theorem states that the conditional distribution of the parameters given the observations is

$$\log f_{\theta|\boldsymbol{Y}}(t \mid \boldsymbol{y}) = c + \log f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid t) + \log \pi_\theta(t) \tag{†}$$

where $c := -\log f_{\boldsymbol{Y}}(\boldsymbol{y})$. Let

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

$$m_0 := \operatorname*{argmax}_{t \in \Omega} \log \pi_\theta(t)$$

(formally these both solve the first-order conditions). $\hat{\theta}$ is termed the *maximum likelihood estimator (MLE)*, and m_0 is the mode of the prior distribution. Then expand each of the terms in (†) as a Taylor series around its maximum to give

maximum likelihood estimator (MLE)

$$\log f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid t) = \log f_{\boldsymbol{Y}|\theta}\big(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y})\big) - \frac{1}{2}\{t - \hat{\theta}(\boldsymbol{y})\}^T \hat{J}(\boldsymbol{y})\{t - \hat{\theta}(\boldsymbol{y})\} + R_n$$
$$\log \pi_\theta(t) = \pi_\theta(m_0) - \frac{1}{2}\{t - m_0\}^T H_0 \{t - m_0\} + R_0$$

where

$$\hat{J}(\boldsymbol{y}) := -\nabla^2 \log f_{\boldsymbol{Y}|\theta}\big(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y})\big)$$
$$H_0 := -\nabla^2 \pi_\theta(m_0)$$

and ∇^2 denotes the matrix of second partial derivatives with respect to the components of t. $\hat{J}$ is termed the *observed Fisher Information*. The regularity conditions ensure that these derivatives exist, that $\hat{J}(\boldsymbol{y})$ is positive definite, and that the remainder terms R_n and R_0 can be ignored for sufficiently large n. Then, collecting all additive constants together in c and dropping the remainder terms,

observed Fisher Information

$$\begin{aligned}\log f_{\theta|\boldsymbol{Y}}(t \mid \boldsymbol{y}) &\approx c - \tfrac{1}{2}\Big[\{t - \hat{\theta}(\boldsymbol{y})\}^T \hat{J}(\boldsymbol{y})\{t - \hat{\theta}(\boldsymbol{y})\} + \{t - m_0\}^T H_0 \{t - m_0\}\Big] \\ &= c - \tfrac{1}{2}\Big[\{t - m_n(\boldsymbol{y})\}^T H_n(\boldsymbol{y})\{t - m_n(\boldsymbol{y})\}\Big]\end{aligned}$$

after completing the square in t, where

$$H_n(\boldsymbol{y}) := \hat{J}(\boldsymbol{y}) + H_0 \tag{4.6a}$$
$$m_n(\boldsymbol{y}) := H_n(\boldsymbol{y})^{-1}\{\hat{J}(\boldsymbol{y})\hat{\theta}(\boldsymbol{y}) + H_0\, m_0\}. \tag{4.6b}$$

Thus the conditional distribution is approximately Normal,

$$f_{\theta|\boldsymbol{Y}}(t \mid \boldsymbol{y}) \approx \phi_d\big(t; m_n(\boldsymbol{y}), H_n(\boldsymbol{y})^{-1}\big)$$

where the expectation and the variance have contributions from both the statistical model (the MLE and the observed Fisher Information) and from the prior distribution (m_0 and H_0).

The regularity conditions ensure that $\hat{J}(\boldsymbol{y})$ increases without limit in n, the number of observations in $\boldsymbol{y}$. And hence $\hat{J}(\boldsymbol{y})^{-1} H_0 \to \mathbf{0}$, implying that

$$H_n(\boldsymbol{y}) \to \hat{J}(\boldsymbol{y}) \text{ and } m_n(\boldsymbol{y}) \to \hat{\theta}(\boldsymbol{y})$$

as n increases, which is the asymptotic result given in (4.5). It will usually speed up convergence to a Normal distribution if the parameters are transformed so that $\Omega = \mathbb{R}^d$.

4.5.2 Hierarchical models*

Section 3.4 introduced exchangeable hierarchical distributions for a vector of random quantities. These have exactly the right property, that the number of hyperparameters is constant, once the parameters themselves have been integrated out, and that $\hat{J}$

is linear in n. Typically, the hyperparameters are given very flat prior distributions, so that H_0, which measures the curvature of the prior around its mode, is small or zero. In this situation the posterior distribution of the hyperparameters will often converge quite rapidly to a Normal distribution.

However, this does not mean that the posterior distribution of the parameters will converge to a Normal distribution. For example, if

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p \mid \boldsymbol{\mu} \overset{\text{iid}}{\sim} f_{\boldsymbol{\beta}|\boldsymbol{\mu}}(\boldsymbol{b} \mid \boldsymbol{m})$$

in the second level of (3.5), then, for sufficiently large n, the hyperparameter $\boldsymbol{\mu}$ would be effectively known, say $\hat{\boldsymbol{\mu}}$. In this case, the posterior distribution of $\boldsymbol{\beta}_i$ would be proportional to

$$f^*_{\boldsymbol{\beta}_i}(\boldsymbol{b}_i) \propto \phi(y_i^{\text{obs}}; \boldsymbol{v}_i^T \boldsymbol{b}_i, s_x^2) \times f_{\boldsymbol{\beta}|\boldsymbol{\mu}}(\boldsymbol{b}_i \mid \hat{\boldsymbol{\mu}})$$

and there is no reason at all for $f^*_{\boldsymbol{\beta}_i}$ to be Normally distributed if, say, $f_{\boldsymbol{\beta}|\boldsymbol{\mu}}$ is non-normal.

As implemented in this example, when n is sufficiently large, little is lost by 'plugging in' the posterior expectation (or the maximum likelihood estimate) for the hyperparameters and effectively removing the bottom level from the hierarchical model. This approximation is known as *parametric empirical Bayes* (Morris, 1983). It is not necessary, but it can lead to helpful simplifications, as well as avoiding the difficulty of specifying a prior distribution for the hyperparameters. It is 'almost admissible' according to the analysis of Section 5.4.

parametric empirical Bayes

4.5.3 Evidence and model choice*

As one example of the use of Theorem 4.2, here is an approximation to the 'evidence', $f_{\boldsymbol{Y}}(\boldsymbol{y}^{\text{obs}})$. Recall from Section 2.6 that this is the crucial quantity in admissible rules for choosing between competing hypotheses. An approach to computing the evidence from an SRS for π^*_θ was outlined in Section 4.2.1. The approximation given here is quicker, and also illuminating.

Theorem 4.3 (Evidence approximation)**.** *Under approximately asymptotic conditions,*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}^{obs}) \approx L(\hat{\theta})\, \pi_\theta(\hat{\theta})\, (2\pi)^{\frac{d}{2}}\, |\hat{J}^{-1}|^{\frac{1}{2}},$$

where L is the likelihood function, $L(t) := f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}^{obs} \mid t)$, $\hat{\theta}$ is the ML estimate of θ, and $\hat{J}$ is the observed Fisher information.[5]

[5] I am dropping the $\boldsymbol{y}^{\text{obs}}$ argument on these two functions to avoid clutter.

Proof. A rearrangement of Bayes's Theorem (also used in Section 4.2.1) gives

$$f_{\boldsymbol{Y}}(\boldsymbol{y}^{\text{obs}}) = \frac{L(t)\, \pi_\theta(t)}{\pi^*_\theta(t)} \quad \text{for any } t \in \Omega, \tag{†}$$

where $L(t)$ is the likelihood function, and π^*_θ is the posterior distribution. Set $t \leftarrow \hat{\theta}$ and note that

$$\pi^*_\theta(\hat{\theta}) \approx \phi_d(\hat{\theta}; \hat{\theta}, \hat{J}^{-1}) = |2\pi \hat{J}^{-1}|^{-\frac{1}{2}}, \tag{‡}$$

using the asymptotic approximation to the posterior distribution given in (4.5). Then, combining (†) and (‡)

$$f_{\boldsymbol{Y}}(\boldsymbol{y}^{\text{obs}}) = \frac{L(\hat{\theta})\,\pi_\theta(\hat{\theta})}{\pi^*_\theta(\hat{\theta})} \approx \frac{L(\hat{\theta})\,\pi_\theta(\hat{\theta})}{|2\pi\hat{J}^{-1}|^{-\frac{1}{2}}},$$

and the result follows from a simple re-arrangement. □

The approximation in Theorem 4.3 is very useful in clarify two of the properties of $f_{\boldsymbol{Y}}$, and in turn clarifying when one hypothesis is favoured over another according to $\boldsymbol{y}^{\text{obs}}$. First, the logarithm of the evidence decomposes into a goodness of fit term and a model complexity penalty:

$$\log f_{\boldsymbol{Y}}(\boldsymbol{y}^{\text{obs}}) \approx \underbrace{\log L(\hat{\theta}) + \log \pi_\theta(\hat{\theta})}_{\text{Goodness of fit}} - \underbrace{\tfrac{1}{2}\big(\log|\hat{J}| - d\log 2\pi\big)}_{\text{complexity penalty}}.$$

Here model complexity is indexed by the number of parameters, d. For fixed n, $\log|\hat{J}|$ is increasing in d, because $\hat{J}$ is a $d \times d$ positive-definite matrix.[6] This decomposition of the evidence into 'goodness of fit less complexity penalty' is the basis for the claim that the Bayesian approach automatically implements the principle of *Occam's Razor*—which asserts that unnecessary complexity should be avoided. Thus a more complex model might improve the goodness of fit, but not by enough to increase the evidence, given the increase in penalty caused by the additional parameters.

[6] Hence $\log|\hat{J}| = \sum_{i=1}^{d} \log \lambda_j$, where $\lambda_j > 0$ is the jth eigenvalue of $\hat{J}$.

Occam's Razor

Second, when choosing between hypotheses the prior distribution π_θ cannot be ignored: in particular, it cannot be replaced by an improper prior. To illustrate, suppose that $\pi_\theta(t;\sigma^2) = \phi_d(t;\mathbf{0},\sigma^2 I)$, for which

$$\lim_{\sigma^2\to\infty} \pi_\theta(\hat{\theta};\sigma^2) = (2\pi\sigma^2)^{-\frac{d}{2}},$$

the (improper) uniform distribution on $\mathbb{R}^d$. Then the log evidence becomes, for large σ^2,

$$\log f_{\boldsymbol{Y}}(\boldsymbol{y}^{\text{obs}}) \approx \ell(\hat{\theta}) - \tfrac{d}{2}\log\sigma^2 - \tfrac{1}{2}\log|\hat{J}|,$$

which is decreasing without limit in σ^2. A hypothesis with an improper prior distribution for its parameters has a log evidence of $-\infty$, and so its evidence is smaller than any hypothesis with a proper prior distribution for its parameters. And, by extension, a hypothesis with a diffuse prior distribution will often have smaller evidence than one with a concentrated prior distribution. So when comparing hypotheses using the evidence (which is the only admissible approach), the choice of prior distribution for the parameters within each hypothesis matters.

This is in contrast to the posterior distribution within a given hypothesis, where the choice of prior distribution is often washed out by large numbers of observations, according to (4.5).

5
Estimators

As described in Section 2.5, there is an approach to decision analysis which does not require a prior distribution on the parameters of the statistical model, but which replaces the parameters in the risk function by an 'estimator' based on the observations. I termed the resulting rule the 'plug-in' rule. This chapter is about estimators, and the final section (Section 5.4) is about the admissibility of the plug-in rule.

The theory of estimators has been a core part of statistics for nearly a century. More details are available in textbooks such as Cox and Hinkley (1974), Schervish (1995), and Casella and Berger (2002).

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

5.1 Estimators and decision analysis

An estimator is a function of the observations which is designed to be like the 'true' parameter values. If the statistical model is $f_{Y|\theta}$ and the parameter space is Ω, then an estimator for θ is a function

$$\tilde{\theta} : \mathcal{Y} \to \Omega$$

where, one hopes,

$$\tilde{\theta}(Y) \approx \theta_0 \text{ when } Y \sim f_{Y|\theta}(\cdot \mid \theta_0), \text{ for all } \theta_0 \in \Omega.$$

But we need a more formal criterion than this, if we are to distinguish good estimators from bad ones. Hence, estimation is treated as a decision analysis (Chapter 2).

In this treatment the action set is $\mathcal{A} := \Omega$, and the loss function $L(a, \theta_0)$ quantifies the consequence of choosing $\theta = a$ when in fact $\theta = \theta_0$. An estimator $\tilde{\theta}$ is simply a rule, with risk function

$$R(\tilde{\theta}, \theta_0) := \mathsf{E}\left\{L\big(\tilde{\theta}(Y), \theta_0\big) \mid \theta_0\right\} = \sum_y L\big(\tilde{\theta}(y), \theta_0\big)\, f_{Y|\theta}(y \mid \theta_0).$$

inadmissible estimators

At the very least, *inadmissible estimators* ought to be avoided—these are clearly bad estimators, because they are dominated by other estimators. The complete class theorems of Section 2.2.2 suggest

that we focus on estimators which are Bayes rules, which will be discussed in Section 5.2.

However, before getting ahead of ourselves, we should contemplate the loss function. A estimator is admissible or inadmissible with respect to a specific loss function, and if the loss function is inappropriate then there is no reason to think that an inadmissible estimator is poor. The difficulty is that the parameters are typically not operationally defined, and the client's actual decision analysis depends not on the parameters themselves, but on the 'state of nature' (or 'predictands'), denoted X in previous chapters. So a loss function for estimating the parameters may not be something that the client has any particular judgement about.

Now this could be taken as an argument for abandoning estimators and plug-in decision rules, and doing the decision analysis 'properly' according to Section 2.5. But an alternative is to propose a generic loss function for parameter estimation. The most popular choice is the *quadratic loss function*.

quadratic loss function

Definition 5.1 (Quadratic loss function)**.** *L is a quadratic loss function for parameter estimation exactly when*

$$L(a, \theta_0) = (a - \theta_0)^T A (a - \theta_0)$$

where A is any $d \times d$ symmetric non-negative definite matrix.[1]

[1] That is, a symmetric matrix for which all the eigenvalues are non-negative, so that $L(a, \theta_0) \geq 0$ for all $a, \theta_0 \in \Omega$.

It is necessary to allow for a matrix A to account for the fact that the parameters may have different units and ranges. A can rescale the loss function to put all of the parameters on the same footing, or it can be used to prioritise particular linear combinations of parameters, including single parameters.

The simplest justification for a quadratic loss function arises from the defensible assertion that the loss function is a differentiable convex function of $a - \theta_0$, with a minimum $L(\mathbf{0}) = 0$ (see, e.g. Savage, 1954, ch. 15). In this case a Taylor series expansion gives

$$\begin{aligned} L(a, \theta_0) &= L(a - \theta_0) \\ &= L(\mathbf{0}) + (a - \theta_0)^T \nabla L(\mathbf{0}) + \tfrac{1}{2}(a - \theta_0)^T \nabla^2 L(\mathbf{0})(a - \theta_0) + \text{h.o.t.} \\ &= \tfrac{1}{2}(a - \theta_0)^T \nabla^2 L(\mathbf{0})(a - \theta_0) + \text{h.o.t.} \end{aligned}$$

where h.o.t. are higher-order terms. Hence $A = \frac{1}{2}\nabla^2 L(\mathbf{0})$ in this justification.[2] Dropping the higher-order terms imposes symmetry with respect to negative and positive values of $a - \theta_0$. This symmetry is one of the weak features of the quadratic loss function. Another is that it is unbounded above for unbounded parameter spaces, which is hardly realistic because if losses really were unbounded then we would all be paralysed into inaction.

[2] A is always symmetric, and A is non-negative definite because L is convex.

There is another justification for a quadratic loss function, based on our aspiration for an estimator. If θ were a scalar, then the best possible outcome for an estimator $\tilde{\theta}$ would be that $\tilde{\theta}(Y)$ and θ_0 were *not materially different* when $Y \sim f_{Y|\theta}(\cdot \mid \theta_0)$; see Section 1.6.o. This would be true if and only if

not materially different

$$\mathsf{E}\{(\tilde{\theta}(Y) - \theta_0)^2 \mid \theta_0\} = 0.$$

If the loss function were the quadratic loss $L(a, \theta_0) = (a - \theta_0)^2$, the lefthand side would be the risk function, $R(\tilde{\theta}, \theta_0)$. Therefore, the quadratic loss function represents our aspiration that the estimator be not materially different from the parameters, in the sense that smaller risks are closer to the ideal situation, where the risk is zero.[3] The quadratic loss function in Definition 5.1 is a generalisation of this to a d-dimensional parameter vector.

[3] A similar argument was used to justify the squared loss function in the characterisation of conditional expectation in Section 1.6.1.

5.2 *The Bayes estimator*

If we want to avoid inadmissible estimators, then at the very least we should satisfy the necessary condition that our estimator is a Bayes rule or a generalised Bayes rule. This involves specifying a prior distribution π_θ. The Bayes rule for the quadratic loss function is termed the *Bayes estimator*; it has a very simple form.

Bayes estimator

Theorem 5.1 (Bayes estimator)**.** *The Bayes estimator is*

$$\theta^*(y) := E\{\theta \mid Y = y\}.$$

In other words, the Bayes estimator is the conditional expectation of the distribution of θ given $(Y = y)$. For a post-data analysis, the Bayes estimator is the posterior expectation. Theorem 5.1 is an immediate consequence of a more general result, that expected quadratic loss is minimised at the expected value.

Theorem 5.2. *If L is quadratic loss, then*

$$\operatorname*{argmin}_m E\{L(m, X)\} = E(X).$$

Proof. let X be any vector of square integrable random quantities with $\bar{m} := \mathsf{E}(X)$. Then

$$\begin{aligned}(m - X)^T A(m - X) \\ = [(m - \bar{m}) + (\bar{m} - X)]^T A[(m - \bar{m}) + (\bar{m} - X)].\end{aligned}$$

Multiplying out and taking expectations gives

$$\begin{aligned}\mathsf{E}\{(m - X)^T A(m - X)\} \\ = (m - \bar{m})^T A(m - \bar{m}) + \mathsf{E}\{(\bar{m} - X)^T A(\bar{m} - X)\},\end{aligned}$$

as the two cross-product terms have zero expectation. Minimising with respect to m gives $m = \bar{m}$ if A is non-negative definite. □

There is a different Bayes estimator for each choice of prior distribution π_θ, although, as noted in Section 4.5, the prior distribution will often play only a small role in the posterior distribution when the number of observations is large. Following Section 2.2, the Bayes estimator is admissible if π_θ is proper and its support is the whole of the parameter space Ω.

One feature of the Bayes estimator which will be relevant below is that it is not transformation-invariant, because the expectation is not transformation-invariant. For example , if $\mathsf{E}(X) = \bar{m}$ and $Y = g(X)$, then

$$\mathsf{E}(Y) = \sum_x g(x) f_X(x) \neq g\left(\sum_x x f_X(x)\right) = g(\bar{m}).$$

The expectation only transforms correctly if g is linear, or if the support of f_X is a single value (in which case $f_X(\bar{m}) = 1$). So if $g : \theta \mapsto \psi$ then the Bayes estimator for ψ cannot be found simply by applying g to the Bayes estimator for θ. Transformation-invariance is discussed in Section 5.3.1.

5.3 *The Maximum Likelihood estimator*

In this chapter I will focus on one particular 'non Bayes estimator', the *maximum likelihood estimator (MLE)*. The MLE is by far the most popular estimator in practice. To complement the MLE, there is a large theory of likelihood-based inference, which I will not address; see, e.g., Pawitan (2001). The main alternative to the MLE is based on estimating functions (for which the MLE is a special case); see Jesus and Chandler (2011) for a recent review.

maximum likelihood estimator (MLE)

Definition 5.2 (MLE). *$\hat{\theta} : \mathcal{Y} \to \Omega$ is a maximum likelihood estimator (MLE) exactly when*

$$\hat{\theta}(y) = \operatorname*{argmax}_{t \in \Omega} f_{Y|\theta}(y \mid t) \quad \textit{for every } y \in \mathcal{Y}.$$

With this definition there are issues of both existence and uniqueness. Existence, because Ω may not be a closed set, and uniqueness because for a given y there may be more than one maximising value of t. However, these are rarely a problem in practice (due in part to judicious choices for statistical models).

Sometimes it is necessary to refer to the MLE of a component of θ, say θ_1. In that case, the MLE of θ_1 is denoted as $\hat{\theta}_1$, and defined as the first component of $\hat{\theta}$.

When the value of the observation y^{obs} is the argument, the value $\hat{\theta}(y^{\text{obs}})$ is termed the *maximum likelihood estimate*. It should always be clear whether one is talking of the MLE (a function) or the ML estimate (a value in Ω).

maximum likelihood estimate

* * *

One very important *caveat*. In order to compute the ML estimate for observations y^{obs}, we must be able to evaluate

$$\text{Lik}(t) := f_{Y|\theta}(y^{\text{obs}} \mid t)$$

at any specified $t \in \Omega$, up to a multiplicative constant that does not depend on t. This is a strong restriction, and rules out many of the statistical models that are regularly used in modern inference,

such as Markov random fields—see Section 3.2 for more details. In this situation a somewhat *ad hoc* alternative is to replace the likelihood function with a tractable approximation, termed a *composite likelihood*; see Varin *et al.* (2011).

composite likelihood

Alternatively, finding the Bayes estimate in this situation is conceptually straightforward. The parameters are treated as random quantities along with X and Y and added to the joint distribution, and the conditional independence graph (CIG). Then Gibbs sampling is used to generate a sufficiently random sequence (SRS, see Section 4.2) for $f^*_{X,\theta}$. The sample mean of the θ components of the SRS gives the Bayes estimate $\theta^*(y^{\text{obs}})$. See Chapter 3 for more details on Gibbs sampling.

5.3.1 *Transformation invariance*

In a parametric model the parameter simply indexes a family of distributions. Any injective function of an index is also an index.[4] So, formally, there is no special index, even though some choices may seem more natural than others. Sometimes the choice of index will be made by the statistician, because some parameterisations are more convenient for one thing, and some for another. For example, for the multivariate Normal distribution, a variance parameter is more convenient for marginalisation, but an inverse variance parameter is more convenient for conditioning (see, e.g., Rue and Held, 2005).

[4] The function g is injective exactly when $g(x) = g(x')$ implies that $x = x'$. So there exists a g^{-1} such that $g^{-1}g(x) = x$ for any x.

However, even though the index may change, the family stays the same. Let $g : \theta \mapsto \psi$ be an injective function. Then the statistical model can be written in terms of θ, or it can be written in terms of ψ, and the two must be related as

$$f_{Y|\psi}(y \mid v) = f_{Y|\theta}\big(y \mid g^{-1}(v)\big). \tag{5.1}$$

An estimator $\tilde{\theta}$ is *transformation invariant* exactly when $\tilde{\psi}(y) = g(\tilde{\theta}(y))$ for all $y \in \mathcal{Y}$. This is an attractive property precisely because the choice of index is arbitrary. However, it should not be over-sold. In particular, it should not be prized above admissibility, particularly if there is a parameterisation which seems more natural.

transformation invariant

Theorem 5.3. *The MLE is transformation invariant.*

Proof. Follows from (5.1) because, for each y,

$$\max_v f_{Y|\psi}(y \mid v) = \max_v f_{Y|\theta}\big(y \mid g^{-1}(v)\big) = \max_t f_{Y|\theta}\big(y \mid t\big).$$

Hence, for any y, $g^{-1}(\hat{\psi}) = \hat{\theta}$, or $\hat{\psi} = g(\hat{\theta})$, as was to be shown. Here I am assuming that the maximum is unique for each y. Statistical models with this property are termed *identifiable*. □

identifiable

5.3.2* *Computing the ML estimate*

As explained at the start of Section 5.3, the MLE is only applicable in cases when the statistical model $f_{Y|\theta}(y^{\text{obs}} \mid t)$ can be evaluated as a function of t; so we must assume this from here on.

There are several points to be made. In principle the ML estimate can be found in any post-data analysis by maximising the *log-likelihood function*,

log-likelihood function

$$\ell(t) := \log f_{Y|\theta}(y^{\text{obs}} \mid t).$$

Taking logarithms is precautionary for numerical reasons, because $f_{Y|\theta}(y^{\text{obs}} \mid t)$ can vary over several orders of magnitude as t varies; but it is also theoretically attractive, as explained below.

However, there is no reason for an arbitrary log-likelihood to be a well-behaved function, with a single global maximum. Possibly there is a dominant maximum and then some foothills: these foothills are annoying computationally, but they are not inferentially worrying. But possibly it's all foothills. In this case there is very little reason to favour the MLE, because it is likely to be highly sensitive to the addition or deletion of a single observation.

So evidence that your numerical optimiser is struggling to find a global maximum of the log-likelihood is not simply a reason to buy a larger computer. It is telling you that the observations alone are not very helpful for making inferences about the parameters.

Happily, the log-likelihood tends to be well-behaved when the number of observations is much larger than the number of parameters. This follows directly from Theorem 4.2. If the posterior distribution is approximately Normal, then the log-likelihood function must be approximately quadratic. In this case, an optimiser which can exploit this knowledge will move rapidly to the global maximum, once it has got clear of the foothills. As noted in Section 4.5, a bijective transformation of the parameter space from Ω to $\mathbb{R}^d$ speeds up the convergence, and because the MLE is transformation invariant, this is perfectly acceptable.

Numerical optimisation of smooth functions that are fairly quadratic is a very well-studied problem; see, e.g., Nocedal and Wright (2006). Practically speaking, you can almost certainly do better than just plugging ℓ into the R function `optim`. In particular, if it is possible to evaluate $\ell(t)$ then it is often possible to compute the gradient vector $\nabla\ell(t)$. Access to this function can dramatically shorten the time needed to find a global maximum. A good optimiser will report an approximation to the observed Fisher information $\hat{J}(y^{\text{obs}})$, which is sequentially approximated during the optimisation. Multiple starting points are always a good idea, as there will usually be foothills.

5.3.3 *The MLE and admissibility under quadratic loss*

In general the MLE is inadmissible with respect to a quadratic loss function. This is because it does not represent a conditional expectation with respect to some prior distribution π_θ; therefore it fails the necessary condition of being a Bayes rule or generalised Bayes rule. However, under the same conditions that hold for Theorem 4.2, for large numbers of observations the MLE can be nearly a generalised Bayes rule and even, in the limit, nearly a

Bayes rule.

I assume that Ω has been bijectively transformed to $\mathbb{R}^d$, to make best possible use of the results in Section 4.5. Under the conditions of that section, the conditional expectation of θ given $(Y = y)$ is approximately given by $m_n(y)$ in (4.6), which I repeat here:

$$H_n(y) := \hat{J}(y) + H_0$$
$$m_n(y) := H_n(y)^{-1}\{\hat{J}(y)\hat{\theta}(y) + H_0\, m_0\}.$$

$\hat{J}(y)$, the observed Fisher Information, increases without limit in n, the number of observations in Y. H_0 is invariant to n, describing the curvature of the prior distribution around its mode m_0. Therefore, under the appropriate conditions,

$$\hat{\theta}(y) \longrightarrow m_n(y),$$

and convergence is accelerated when π_θ is an improper prior distribution (for which H_0 is small or even zero).

This relationship $\hat{\theta} \approx m_n$ is only a valid approximation under certain conditions on the statistical model and the prior distribution. But where these conditions hold, and where the number of observations is sufficiently large, the MLE is effectively the conditional expectation, and in this case it satisfies the necessary condition for admissibility under quadratic loss.

* * *

Here is a cautionary tale about the MLE and admissibility, the *Stein paradox*. Consider the statistical model

Stein paradox

$$\boldsymbol{Y} := (Y_1, \ldots, Y_n) \sim \mathrm{N}_n(\boldsymbol{\theta}, I)$$

where $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ and I is the $n \times n$ identity matrix; in other words, there is a parameter for each observation. A sensible estimator of $\boldsymbol{\theta}$ would seem to be

$$\tilde{\boldsymbol{\theta}}(\boldsymbol{y}) := \boldsymbol{y} \tag{5.3}$$

which is also the MLE. It is also a generalised Bayes rule under quadratic loss, with an improper prior distribution $\pi_{\boldsymbol{\theta}}(\boldsymbol{t}) \propto 1$. But, to the surprise and consternation of statisticians, Stein (1956) showed that (5.3) was inadmissible for $n \geq 3$. Stein's paradox is carefully analysed in Cox and Hinkley (1974, sec. 11.8), and there is an insightful and non-technical summary in Efron and Morris (1977).

This example illustrates the point made in Section 2.2.2, that the union of Bayes and generalised Bayes rules forms a complete class but not a minimal complete class. That is to say, there are some generalised Bayes rules that are admissible (here, in the case $n \leq 2$), so that Bayes rules on their own are not a complete class. But there are some generalised Bayes rules that are inadmissible, so that the union of Bayes and generalised Bayes rules contains some inadmissible rules (here $n \geq 3$).

More pragmatically, though, it warns us against the MLE when the number of observations is not substantially larger than the number of parameters: in Stein's paradox, $n = d$. In this situation the Bayes estimator offers the opportunity to augment the observations with additional judgements about θ, as represented by a proper prior distribution π_θ.

5.4* *Plug-in estimators and admissibility*

Now return to the primary decision analysis, described in Chapter 2. The risk function for rule δ is

$$R(\delta, t) := \mathsf{E}\left\{L\big(\delta(Y), X\big) \mid \theta = t\right\}.$$

The Bayes rule δ^* minimises the integrated risk $R(\delta) := \mathsf{E}\{R(\delta, \theta)\}$ with respect to some prior distribution π_θ. The necessary condition for any rule δ to be admissible is that it is a Bayes rule or a generalised Bayes rule (which allows for π_θ to be improper).[5]

[5] In this section I will write 'Bayes rule' for 'Bayes rule or generalised Bayes rule'.

Section 2.5 introduced the 'plug-in rule', which was not a Bayes rule, but which was simple and tractable. Bayes rules have the form

$$\delta^*(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathsf{E}\left\{L(a, X) \mid Y = y\right\},$$

for some prior distribution π_θ. The plug-in rule has the form

$$\tilde{\delta}(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathsf{E}\left\{L(a, X) \mid \theta = \tilde{\theta}(y)\right\},$$

for some estimator $\tilde{\theta}$. Not being a Bayes rule, the plug-in rule is inadmissible. Perhaps, however, there are conditions under which the plug-in rule is almost a Bayes rule? This would require that

$$\mathsf{E}\left\{L(a, X) \mid Y = y\right\} \approx \mathsf{E}\left\{L(a, X) \mid \theta = \tilde{\theta}(y)\right\}.$$

The following result gives the conditions. It might be a good idea to revisit Section 1.6 first.

Theorem 5.4. *If X and Y are conditionally independent given θ, and the regularity conditions of Theorem 4.2 hold, and n is sufficiently large, then the plug-in rule with the MLE for θ approximates the Bayes rule.*

Proof. Let $\psi_a \in \mathcal{E}\{L(a, X) \mid \theta, Y\}$. Definition 3.1 states that if X and Y are conditionally independent given θ then $\psi_a(t, y) = \phi_a(t)$ for some ϕ_a; that is, ψ_a is invariant to y. Recollect from Theorem 1.12 that $\phi_a \in \mathcal{E}\{L(a, X) \mid \theta\}$. Then

$$\begin{aligned}
\mathsf{E}\{L(a, X) \mid Y = y\} &= \mathsf{E}\left[\mathbb{E}\{L(a, X) \mid \theta, Y = y\} \mid Y = y\right] \quad \text{by the LIE} \\
&= \mathsf{E}\left[\psi_a(\theta, y) \mid Y = y\right] \\
&= \mathsf{E}\left[\phi_a(\theta) \mid Y = y\right] \quad \text{by conditional independence} \\
&\approx \phi_a\big(\mathsf{E}[\theta \mid Y = y]\big) \qquad (\dagger) \\
&= \mathsf{E}\left\{L(a, X) \mid \theta = \theta^*(y)\right\} \\
&\approx \mathsf{E}\left\{L(a, X) \mid \theta = \hat{\theta}(y)\right\} \qquad (\dagger)
\end{aligned}$$

where θ^* is the Bayes estimator, and $\hat{\theta}$ is the MLE. The approximations at (†) require that Y approximately determines θ and that $\theta^*(y) \approx \hat{\theta}(y)$. In other words, that the regularity conditions of Theorem 4.2 hold, with sufficiently large n.[6] □

[6] A more elegant proof would make a second-order approximation at the first (†), rather than a first-order approximation.

* * *

Section 3.4 discussed hierarchical models, which are often used for drawing inferences about populations from samples; e.g. in epidemiology (Rothman *et al.*, 2008). In hierarchical models, the predictands are often mutually conditionally independent given the hyperparameters θ, and the observations are a subset of the predictands. Thus if we write $\boldsymbol{X} := \{\boldsymbol{X}_A, \boldsymbol{X}_B\}$ and $\boldsymbol{Y} := \boldsymbol{X}_A$ then $\boldsymbol{X}_B \perp\!\!\!\perp \boldsymbol{Y} \mid \theta$. Hierarchical models tend to satisfy the regularity conditions of Theorem 4.2, see Section 4.5.2. And it is often possible to collect quite large samples (large n). Thus the conditions in Theorem 5.4 are satisfied. And so it is not surprising that the plug-in rule has endured: it is a good approximation to the Bayes rule in this situation, and so satisfies the necessary condition for admissibility.

But for a more complex decision analysis, conditional independence may not be appropriate, the conditions of Theorem 4.2 may not hold, or n may not be sufficiently large. In this case the plug-in rule is inadmissible. So to conclude, *the plug-in decision rule of Section 2.5 is only (approximately) admissible under restrictive conditions on the joint distribution of the predictands, the observations, and the parameters*. In particular, its common usage in some areas of statistical inference should not be taken to imply that it is generally defensible.

6
Signficance levels and confidence sets

A significance level measures the concordance between a set of observations y and judgements about Y, the latter represented in terms of a probability distribution, or a family of distributions. It takes the form of a p-value—one of the two most misunderstood concepts in statistical inference. This chapter covers the definition, computation, and choice of a p-value (Section 6.1, Section 6.2, and Section 6.3). P-values can be used to construct confidence sets for the parameters (Section 6.4 *et seq.*)—the other most misunderstood concept in statistical inference.

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

6.0 Preamble

The length of this chapter is a testament not so much to the importance of significance levels and confidence sets (although they are ubiquitous), but to the errors that are made in interpreting them. In order that there can be no confusion, let me stress at the very outset that significance levels and confidence sets should not be used for making decisions that involve choosing between options. The correct tool for this is Decision Analysis (Chapter 2). Any choices that are not derived within the framework of Decision Analysis are likely to be inadmissible, and indefensible. I deliberately avoid writing 'signficance *test*', as a p-value should not be used, formally, to test a theory, as will be explained in Section 6.1.

But it is a moot question whether the many intermediate choices that are made by the client and her experts (including the statistician) during the course of a statistical inference must also be derived within the framework of Decision Analysis. Ideally, yes—but in practice there are so many choices, that one looks for informal guidance in order to make some progress. In particular, one looks for 'sanity checks' to ensure that the analysis is proceeding in the right direction; most notably, not drifting too far from a path which is concordant with the observations y^{obs}. Thus a significance level is a type of alarm bell, which rings when the drift has become worryingly large. Likewise, it is a crude screening device, allowing a first-stage selection of promising paths. The crucial feature of

such informal guidance is that it must require much less effort to implement than the formal guidance—otherwise what is the point?!

The crucial feature of a significance level is that it involves only a single hypothesis, conventionally denoted H_0, the *null hypothesis*, without the explicit requirement for any alternative hypothesis. Thus the statistician asks: "I am currently assuming H_0 about $\{X, Y\}$—am I on the right path?", and it seems intuitive that this question will sometimes have a clear answer based on H_0 and y^{obs} alone, without any reference to other paths. Likewise the domain expert might ask "I would like to use process model H_0 to represent $\{X, Y\}$—did it do a good job of representing y^{obs}?"

null hypothesis

In statistical inference H_0 takes the form of a family of distributions for Y, written generically as

$$H_0 : Y \sim f_Y \in \mathcal{F},$$

for some specified family $\mathcal{F}$. It is helpful to distinguish two situations. In a *simple hypothesis* $\mathcal{F}$ has only one element, while in a *composite hypothesis* $\mathcal{F}$ has more than one element. If we take the statistical model $f_{Y|\theta}$ as given, then simple hypotheses either takes the form

simple hypothesis

composite hypothesis

$$f_Y(\cdot) = \int f_{Y|\theta}(\cdot \mid t)\, \pi_\theta(t)\, \mathrm{d}t, \tag{6.1a}$$

where one specifies a prior distribution π_θ,[1] or the form

[1] Giving rise to prior predictive diagnostics, see Section 4.4.

$$f_Y(\cdot) = f_{Y|\theta}(\cdot \mid \theta_0) \tag{6.1b}$$

where one specifies a point value for θ; i.e. $H_0 : \theta = \theta_0$. Composite hypotheses take the general form

$$H_0 : \theta \in \Omega_0$$

for some $\Omega_0 \subset \Omega$, where Ω_0 has more than one element. In this chapter I will focus on simple hypotheses until Section 6.6, which considers nuisance parameters.

6.1 P-values for simple hypotheses

A simple null hypothesis H_0 is represented as

$$H_0 : Y \sim f_Y$$

for some specified distribution f_Y. The idea of a significance level is to construct a *statistic* (a scalar real-valued function of the observations) which can be interpreted on a standard scale, no matter what H_0 might be. Formally, this idea is implemented as a *p-value*.

statistic

p-value

Definition 6.1 (*P*-value)**.** *A statistic $p : \mathcal{Y} \to \mathbb{R}$ is a p-value for a hypothesis H_0 exactly when the realm of $p(Y)$ is $[0, 1]$,*

1. *$p(Y)$ has a subuniform distribution under H_0, and*
2. *$p(Y)$ under H_0 stochastically dominates $p(Y)$ under some decision-relevant departure from H_0.*

If $p(Y)$ has a uniform *distribution under H_0 then I term it an* exact *p-value.*

In this definition, X has a *subuniform distribution* exactly when $\Pr(X \leq u) \leq u$ for all $u \in [0, 1]$. And X *stochastically dominates* Y exactly when

subuniform distribution

stochastically dominates

$$\Pr(X > a) \geq \Pr(Y > a) \quad \text{for all } a \in \mathbb{R}.$$

If X stochastically dominates Y then, visually, the distribution function of X lies to the right (or, rather, never to the left) of Y. The importance of the stochastic dominance property was emphasised by DeGroot (1973). To avoid writing "decision-relevant departure from H_0" I will just write H', where H' is a direction in 'distribution space' away from H_0.

Both properties in Definition 6.1 are important, but only the second is necessary to prove the following result, which relates the p-value to the odds ratio. Recollect from Theorem 2.6 that the odds ratio is the only admissible way to choose between two competing hypotheses; i.e. to choose between H_0 and the (notional) H'.

Theorem 6.1. *Small p-values indicate that the odds ratio for H_0 versus H' is likely to be less than one.*

Proof. I will assume that the p-value has a continuous distribution.[2]

Denote the distribution function of $p(Y)$ under H_0 as F_0, with PDF f_0. Let u denote the p-value, and let u be small, close to zero. In this case

$$f_0(u) \approx \frac{F_0(u) - F_0(0)}{u} = \frac{u_0}{u}$$

say, where $u_0 := F_0(u)$ and $F_0(0) = 0$. Let F' be the distribution function of $p(Y)$ under H', and f' the PDF. By the same reasoning,

$$f'(u) \approx \frac{u'}{u}$$

where $u' := F'(u)$ and $F'(0) = 0$, so that $u_0 \leq u'$ according to the stochastic dominance property in Definition 6.1. Hence the odds ratio for H_0 versus H' is

$$\frac{f_0(u)}{f'(u)} \approx \frac{u_0}{u'} \leq 1.$$

□

[2] This proof is really just illustrating the 'obvious' point that if two random quantities X and Y have the same realm $\mathcal{U}$ and X stochastically dominates Y then $f_Y(u) \geq f_X(u)$ when u is a sufficiently small value in $\mathcal{U}$.

There are three points to notice about this result. First, it only holds for small p-values. For large p-values, the odds ratio is undetermined. Second, a small p-value does not fix the value of the

odds ratio, beyond suggesting that it will be less than one. The value could be almost one, or tiny—it would be necessary to make an explicit choice for H' to resolve this. Third, recollect from the discussion in Section 2.6.1 that the odds ratio is not the same as the posterior odds, so that a small odds ratio does not imply that H_0 is improbable relative to H'. All three of these points correspond to common errors.

Errors in interpreting *p*-values:

1. The *large p-value fallacy*: a large p-value implies that the odds ratio favours H_0 over its alternatives. — large p-value fallacy
2. The *odds ratio fallacy*: a small p-value implies a small odds ratio for H_0 versus its alternatives. — odds ratio fallacy
3. The *base rate fallacy*: a small p-value makes H_0 less probable than its alternatives. — base rate fallacy

I have coined the name 'odds ratio fallacy', but its origins go back at least to Edwards *et al.* (1963), who noted that the p-value of 0.05 was much smaller than a lower bound for the odds ratio in the case where Y is Normal. The 'base rate fallacy' gets its name from forensics and medical science, where the prior odds $\mathsf{Pr}(H_0) / \mathsf{Pr}(H')$ is taken to be the ratio of the relative frequencies (base rates) of H_0 and H' in the population (see, e.g., Gigerenzer, 2003).

To return to the point made at the start of this chapter, it is a horrible error to think that a p-value can be used to 'reject H_0' or 'not reject H_0' on the basis of y^{obs}. To 'reject H_0' on the basis of a small p-value commits both the odds ratio fallacy and the base rate fallacy. To 'not reject H_0' on the basis of a large p-value commits the large p-value fallacy. P-values are never decision-relevant—they are purely indicative.

P-values have always been controversial, because misinterpretation is so easy; see, for example, Greenland and Poole (2013), which is just the latest exchange in fifty years of discussion. If there was an acid test for statistical competence, correctly defining and interpreting a p-value would do.

6.2 *Constructing and computing p-values*

I continue to assume that H_0 is a simple hypothesis of the form $H_0 : Y \sim f_Y$ for some specified f_Y.

6.2.1 *Using a test statistic*

The ubiquitous method for constructing p-values is to propose a test statistic $s : \mathcal{Y} \to \mathbb{R}$ with the property that large values of $s(y)$ are suggestive of a decision-relevant departure from H_0. Then we have the following result, which is precise in the case where $s(Y)$ is continuous, but imprecise when it is discrete.

Theorem 6.2. *Let H_0 be a simple hypothesis for Y, and let $S := s(Y)$ be a test statistic for which S under H_0 is stochastically dominated by S under some decision-relevant departure from H_0, labelled H'. Then*

$$p_s(y) := \mathsf{Pr}\{S \geq s(y) \mid H_0\}$$

has a subuniform distribution under H_0, and $p_s(Y)$ under H_0 nearly stochastically dominates $p_s(Y)$ under H'. If S is continuous then p_s is an exact p-value.

Proof. To prove the first property of Definition 6.1, I use a nifty trick from Casella and Berger (2002, section 8.3.4). Let G_0 be the distribution function of $-S$ under H_0. Then

$$p_s(y) = \mathsf{Pr}\{S \geq s(y) \mid H_0\} = \mathsf{Pr}\{-S \leq -s(y) \mid H_0\} = G_0(-s(y)).$$

Then since $p_s(Y) = G_0(-S)$, subuniformity of $p(Y)$ under H_0 follows from the Probability Integral Transform (Section A.3). If S is continuous, then $p_s(Y)$ is uniform under H_0.

For property 2, let H' be a decision-relevant departure from H_0, for which

$$\mathsf{Pr}(S > a \mid H_0) \leq \mathsf{Pr}(S > a \mid H') \qquad (\dagger)$$

for each $a \in (0, 1)$, according to the conditions of the theorem. If S is continuous, G_0^{-1} exists, and then

$$\begin{aligned}
\mathsf{Pr}\{p_s(Y) > u \mid H_0\} &= \mathsf{Pr}\{G_0(-S) > u \mid H_0\} \\
&= \mathsf{Pr}\{S < -G_0^{-1}(u) \mid H_0\} \qquad (\ddagger) \\
&\geq \mathsf{Pr}\{S < -G_0^{-1}(u) \mid H'\} \quad \text{by } (\dagger) \\
&= \mathsf{Pr}\{G_0(-S) > u \mid H'\} \qquad (\ddagger) \\
&= \mathsf{Pr}\{p_s(Y) > u \mid H'\},
\end{aligned}$$

and hence $p_s(Y)$ under H_0 stochastically dominates $p_s(Y)$ under H', as required.

In the more general case where S is discrete, a non-decreasing generalised inverse can be defined,

$$G_0^-(u) := \inf_s \{s \in \mathbb{R} : G_0(s) \geq u\}$$

but this has the property that $G_0^- G_0(s) \leq s$ and $G_0 G_0^-(u) \geq u$. Hence the lines between the (‡)'s cannot be true, but will be good approximations when the distances between points in $G_0(\mathcal{S})$ are small, where $\mathcal{S}$ is the realm of S. □

Theorem 6.2 is sometimes taken to be the definition of a p-value, with small p-values being interpreted as 'surprising' under H_0. This interpretation makes little sense, because such a p-value depends on the choice of test statistic, of which there is an infinite variety. Perhaps it is possible to choose a test statistic with the element of 'surprise'. But it would be much more sensible to choose a test statistic that indicated a decision-relevant departure from H_0.

6.2.2 *Computing p-values by simulation*

Occasionally it will be possible to choose a test statistic s with a known distribution under H_0, from which an explicit expression for $p_s(y)$ can be derived (see Section 6.3.1). But this puts the cart before the horse—what we want is to be able to choose our test statistic according to our judgement about decision-relevant departures from H_0. Happily, the p-value for any s can be computed by simulation using following result.

Theorem 6.3. *For any finite sequence of scalar random quantities $X^0, X^1, \ldots, X^m$, define the rank of X^0 in the sequence as*

$$R := \sum_{i=1}^{m} (X^i \leq X^0).$$

If $X^0, X^1, \ldots, X^m$ are exchangeable then R has a uniform distribution on the integers $0, 1, \ldots, m$, 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

$$\mathsf{Pr}(R = r) = \frac{1}{m+1} \quad \text{for } r = 0, 1, \ldots, m \tag{†}$$

and zero otherwise, proving the first claim.

To prove the second claim,[3]

$$\begin{aligned}
\mathsf{Pr}\left\{\frac{R+1}{m+1} \leq u\right\} &= \mathsf{Pr}\left\{R+1 \leq u(m+1)\right\} \\
&= \mathsf{Pr}\left\{R+1 \leq \lfloor u(m+1) \rfloor\right\} \quad \text{as } R \text{ is an integer} \\
&= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \mathsf{Pr}(R = r) \\
&= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \frac{1}{m+1} \quad \text{from (†)} \\
&= \frac{\lfloor u(m+1) \rfloor}{m+1} \leq u.
\end{aligned}$$

□

[3] Notation: $\lfloor x \rfloor$ is the largest integer no larger than x, termed the 'floor' of x.

Now take a statistic $s : \mathcal{Y} \to \mathbb{R}$ which has the property that larger values of $s(y)$ are suggestive of a decision-relevant departure from H_0. Define $S := s(Y)$ and $S^j := s(Y^j)$ where $Y^1, \ldots, Y^m \overset{\text{iid}}{\sim} f_Y$. Then $S, S^1, \ldots, S^m$ form a scalar exchangeable sequence under H_0. Hence if

$$R_s(y) := \sum_{j=1}^{m} (-S^j \leq -s(y)) = \sum_{j=1}^{m} (S^j \geq s(y))$$

then Theorem 6.3 implies that

$$P_s(y) := \frac{R_s(y) + 1}{m+1}$$

has a subuniform distribution under H_0.[4] Furthermore, the Weak

[4] Here I write both R_s and P_s as capitals, because they are functions of the random quantities $Y^1, \ldots, Y^m$.

Law of Large Numbers (Section A.4) shows that

$$\begin{aligned}\lim_{m\to\infty} P_s(y) &= \lim_m \frac{\lim_m R_s(y)+1}{m+1} \\ &= \lim_m \frac{m^{-1}\{R_s(y)+1\}}{m^{-1}\{m+1\}} \\ &= \mathsf{Pr}\{S \geq s(y) \mid H_0\}\end{aligned}$$

and so the asymptotic limit of P_s is the p-value defined in Theorem 6.2. Hence, asymptotically, P_s satisfies the stochastic dominance property in the same way as p_s. Thus a large m is preferable, even though P_s is subuniform for all m.

Therefore P_s can be computed from an SRS for f_Y, which is used to provide $Y^1, \ldots, Y^m$. In order for these to be exchangeable it is sufficient that they are independent, and hence these m values must be extracted from well-separated locations in the SRS (see Section 4.2).

Besag and Clifford (1989) made an elegant suggestion for null hypotheses of the form $H_0 : \theta = \theta_0$, based on an MCMC simulation on $\mathcal{Y}$ with stationary distribution $f_{Y|\theta}(\cdot \mid \theta_0)$. First, initialise the MCMC at y, and run the MCMC *backwards* for k steps. Then initialise m independent sequences from the same value and run each one forwards for k steps. Then y and the MCMC values will be exchangeable under H_0 (see Figure 6.1). If k is small these values will *not* be independent under H_0, but exchangeable is all that is required by Theorem 6.3. However, a large k and a large m are better, again because of the stochastic dominance property.

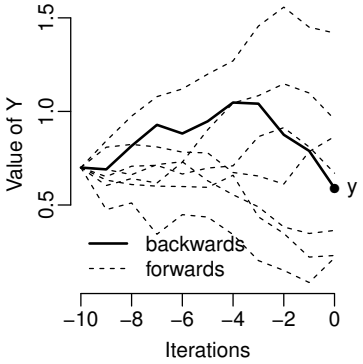


Figure 6.1: The MCMC approach of Besag and Clifford (1989), with $m = 7$ and $k = 10$.

6.3 *Choice of test statistic*

Ideally, the client would be able to specify a test statistic for which large values indicate a decision-relevant departure from H_0. But in many cases it is useful have a default test statistic. Such a choice can at least avoid any suspicion that the test statistic has been selected to reach a foregone conclusion.[5]

[5] Although recollect my warning about basing actions on significance levels, in Section 6.1.

However, there is one important *caveat* which applies in this section. It is the same *caveat* given at the start of Section 5.3. It has to be possible to evaluate and manipulate the statistical model $f_{Y|\theta}(y \mid t)$ as a function of t. This is almost taken for granted in textbooks, but it rules out many of the statistical models that are regularly used in modern statistical inference.

Here I consider testing hypotheses about the parameters. Suppose that

$$H_0 : \theta = \theta_0$$

for some specified $\theta_0 \in \Omega \subset \mathbb{R}^d$, but no specified alternative value $H' : \theta = \theta'$ is available. Imagine if such a θ' were specified. In this case, all admissible tests for H_0 versus H' would have the form

$$\text{choose } H' \text{ over } H_0 \text{ when } \frac{f_{Y|\theta}(y \mid \theta')}{f_{Y|\theta}(y \mid \theta_0)} > c$$

for some $c > 0$ (Theorem 2.6). Now suppose that θ' is close to θ_0. Taking the logarithm of the odds ratio and writing $\theta' = \theta_0 + \delta$ for small δ gives

$$\begin{aligned}\log \frac{f_{Y|\theta}(y \mid \theta_0 + \delta)}{f_{Y|\theta}(y \mid \theta_0)} &= \log f_{Y|\theta}(y \mid \theta_0 + \delta) - \log f_{Y|\theta}(y \mid \theta_0) \\ &\approx \delta^T u(y, \theta_0),\end{aligned}$$

using a Taylor series expansion to first order, where u is the *score function*,

score function

$$u(y, t) := \nabla \log f_{Y|\theta}(y \mid t)$$

and ∇ indicates differentiation with respect to the components of t.[6] Thus for testing for small departures from θ_0 in the direction δ, admissible tests would have the form $\delta^T u(y, \theta_0) > c$ for some $c \in \mathbb{R}$. Overall, this suggests that test statistics based on the score function will be powerful for examining small departures from H_0. In the very simple case where θ is a scalar parameter, $u(\cdot, \theta_0)$ is termed the *the locally most powerful test statistic*.

[6] Hence $u(y, t)$ is a d-vector.

the locally most powerful test statistic

The variance of the score function will be used below. This is termed the *expected Fisher information*,

expected Fisher information

$$I(t) := \mathrm{Var}\{u(Y, t) \mid \theta = t\}, \tag{6.2}$$

and it is, by construction, a $d \times d$ symmetric non-negative definite matrix.

6.3.1* *Special case of regular statistical models*

As well as being theoretically attractive, test statistics based on the score function have a practical advantage as well, for statistical models of a particular form: their distribution under $H_0 : \theta = \theta_0$ is approximately known. In this section I write $\boldsymbol{Y} := (Y_1, \ldots, Y_n)$.

The defining feature of *regular statistical models* is that the operations of taking expectations over $\boldsymbol{Y}$ and taking derivatives with respect to t can be interchanged. The dominant necessary condition for this is that the support of the statistical model does not depend on the values of the parameters. If this condition holds, a sufficient condition is that the realm of $\boldsymbol{Y}$ is bounded. See Casella and Berger (2002, ch. 2) for more details of regular statistical models.

regular statistical models

Theorem 6.4. *If $f_{\boldsymbol{Y}|\theta}$ is a regular statistical model then $\mathsf{E}\{u(\boldsymbol{Y}, t) \mid \theta = t\} = \mathbf{0}$.*

Proof. Start from the identity $\sum_{\boldsymbol{y}} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid t) = 1$. Then differentiate both sides with respect to t_j. The righthand side is zero. For a regular model, the lefthand side is

$$\begin{aligned}\frac{\partial}{\partial t_j} \sum_{\boldsymbol{y}} f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid t) &= \sum_{\boldsymbol{y}} \nabla_j f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid t) \\ &= \sum_{\boldsymbol{y}} \nabla_j \log f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid t) \times f_{Y|\theta}(\boldsymbol{y} \mid t) \\ &= \mathsf{E}\{u_j(\boldsymbol{y}, t) \mid \theta = t\},\end{aligned}$$

which must therefore equal zero. □

Now consider the special case where

$$f_{\boldsymbol{Y}|\theta}(\boldsymbol{y} \mid t) = \prod_{i=1}^{n} f_{Y|\theta}(y_i \mid t).$$

In this case, the $\boldsymbol{Y}$ is said to be IID given θ (Section 3.1). For the score function,

$$\begin{aligned} u(\boldsymbol{Y}, t) &= \nabla \log \prod_{i=1}^{n} f_{Y|\theta}(Y_i \mid t) \\ &= \nabla \sum_{i=1}^{n} \log f_{Y|\theta}(Y_i \mid t) \\ &= \sum_{i=1}^{n} \nabla \log f_{Y|\theta}(Y_i \mid t) = \sum_{i=1}^{n} u_1(Y_i, t), \end{aligned} \tag{6.3}$$

where u_1 is the score function for a single observation. And then the expected Fisher information (see (6.2)) equals

$$I(t) = n \operatorname{Var}\left\{u_1(Y_i, t) \mid \theta = t\right\} = n I_1(t),$$

where I_1 is the expected Fisher Information for a single observation. The following result then follows directly from (6.3) and the Central Limit Theorem (Section A.4).

Theorem 6.5. *If $\boldsymbol{Y}$ is IID given $\theta \in \Omega \subset \mathbb{R}^d$ and $f_{Y|\theta}$ is a regular statistical model, then*

$$u(\boldsymbol{Y}, \theta_0) \xrightarrow{D} N_d\big(\mathbf{0}, n I_1(\theta_0)\big)$$

where convergence is under $H_0 : \theta = \theta_0$.

Therefore, under the conditions of Theorem 6.5 the score function has a known distribution under H_0 in the limit as $n \to \infty$, and an approximately known distribution for large n. This means that approximate p-values for H_0 which are based on the score function can be computed directly, without simulation. In particular, there is the general purpose test statistic for H_0,

$$s(\boldsymbol{y}, \theta_0) = u(\boldsymbol{y}, \theta_0)^T \left\{n I_1(\theta_0)\right\}^{-1} u(\boldsymbol{y}, \theta_0) \tag{6.4a}$$

for which Normal distribution theory implies that $s(\boldsymbol{Y}, \theta_0) \xrightarrow{\mathrm{D}} \chi^2_d$ under H_0 (see, e.g., Mardia *et al.*, 1979, ch. 3, Cor. 3.2.1.1). Thus, for large n,

$$\begin{aligned} p_s(\boldsymbol{y}, \theta_0) &= \mathsf{Pr}\left\{s(\boldsymbol{Y}, \theta_0) \geq s(\boldsymbol{y}, \theta_0) \mid \theta = \theta_0\right\} \\ &\approx 1 - F_{\chi^2_d}\big(s(\boldsymbol{y}, \theta_0)\big) \end{aligned} \tag{6.4b}$$

where $F_{\chi^2_d}$ is the distribution function of a χ^2_d random quantity.

6.3.2* *Expected and observed Fisher Information*

For regular models, there is an additional result which is useful when computing the expected Fisher Information—it is often easier to compute derivatives than to compute variances.

Theorem 6.6 (Fisher's identity)**.** *If $f_{\boldsymbol{Y}|\theta}$ is a regular statistical model, then*

$$I(t) = -\,\mathsf{E}\big\{\nabla^2 \log f_{\boldsymbol{Y}|\theta}(\boldsymbol{Y} \mid t) \mid \theta = t\big\}. \tag{6.5}$$

Proof. The following cryptic outline should suffice:

$$\begin{aligned}
\nabla_i \nabla_j \log f &= \nabla_i \frac{\nabla_j f}{f} \\
&= \frac{\nabla_i \nabla_j f \cdot f - \nabla_j f \cdot \nabla_i f}{f^2} \quad \text{the quotient rule} \\
&= \frac{\nabla_j \nabla_j f}{f} - \nabla_i \log f \cdot \nabla_j \log f.
\end{aligned}$$

Now take expectations conditional on θ, to give

$$\mathsf{E}\left\{\nabla_i \nabla_j \log f \mid \theta = t\right\} = \mathsf{E}\left\{ \frac{\nabla_j \nabla_j f}{f} \,\middle|\, \theta = t \right\} - I_{ij}(t), \tag{†}$$

because, for the second term,

$$\begin{aligned}
&\mathsf{E}\{\nabla_i \log f \cdot \nabla_j \log f \mid \theta = t\} \\
&\quad = \mathsf{E}\left\{u_i \cdot u_j \mid \theta = t\right\} = \mathrm{Cov}\left\{u_i, u_j \mid \theta = t\right\}
\end{aligned}$$

as $\mathsf{E}\{u_i \mid \theta = t\} = \mathsf{E}\{u_j \mid \theta = t\} = 0$ by regularity and Theorem 6.4.

Finally, the first term in (†) is zero, because, by regularity, and suppressing the conditioning on $\theta = t$,

$$\begin{aligned}
\mathsf{E}\,\frac{\nabla_i \nabla_j f}{f} &= \sum_{\boldsymbol{y}} \frac{\nabla_i \nabla_j f}{f} f \\
&= \sum_{\boldsymbol{y}} \nabla_i \nabla_j f \\
&= \nabla_i \sum_{\boldsymbol{y}} \nabla_j \log f \cdot f \\
&= \nabla_i \,\mathsf{E}\left\{\nabla_j \log f\right\} = \nabla_i \,\mathsf{E}\left\{u_j\right\} = 0
\end{aligned}$$

again by Theorem 6.4. □

Now recollect the definition of the *observed Fisher Information*, given in Section 4.5,

$$\hat{\jmath}(\boldsymbol{y}) := -\nabla^2 \log f_{\boldsymbol{Y}|\theta}\big(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y})\big), \tag{6.6}$$

where $\hat{\theta}$ is the Maximum Likelihood Estimator (MLE). It is surely not a coincidence that $I(t)$ in (6.5) and $\hat{\jmath}(\boldsymbol{y})$ in (6.6) look so similar, even if they do have different arguments. And indeed it is not.

Theorem 6.7. *If $\boldsymbol{Y}$ is IID given θ and $f_{Y|\theta}$ is a regular statistical model then*

$$\hat{\jmath}(\boldsymbol{Y}) \xrightarrow{P} n I_1(\theta_0)$$

under $H_0 : \theta = \theta_0$.

Proof (informal). The conditions of the theorem satisfy the regularity conditions of Theorem 4.2, and hence $\hat{\theta}(\boldsymbol{Y}) \xrightarrow{\mathrm{P}} \theta_0$ under H_0. And then, because $\boldsymbol{Y}$ is IID given θ,

$$\hat{J}(\boldsymbol{y}) = \sum_{i=1}^{n} -\nabla^2 \log f_{Y|\theta}\big(y_i \,|\, \hat{\theta}(\boldsymbol{Y})\big).$$

Hence, under H_0,

$$n^{-1}\hat{J}(\boldsymbol{Y}) \xrightarrow{\mathrm{P}} \mathsf{E}\left\{-\nabla^2 \log f_{Y|\theta}(Y_i \,|\, \theta_0) \,|\, \theta = \theta_0\right\} = I_1(\theta_0)$$

according to the Weak Law of Large Numbers (WLLN, Section A.4) and Theorem 6.6. □

This result supports the practice of replacing $I(\theta_0)$ with $\hat{J}(\boldsymbol{y}^{\mathrm{obs}})$ in asymptotic approximations derived under H_0. Sometimes this will be simply for convenience. If the value of the MLE has been found, then $\hat{J}(\boldsymbol{y}^{\mathrm{obs}})$ ought to be readily available (see Section 5.3), whereas $I(\theta_0)$ requires an integration over $\mathcal{Y}$. But there are also theoretical reasons for preferring to use the observed Fisher Information rather than the expected Fisher information, explored in Efron and Hinkley (1978).

6.4 *Confidence sets*

> The only use I know for a confidence interval is to have confidence in it. —L.J. Savage (Savage *et al.*, 1962, p. 98)

Confidence sets are a Frequentist approach to quantifying parameter uncertainty, in terms of a set in parameter space. A Bayesian approach to the same problem would simply assert a loss function

$$L(A, \theta_0) \quad \text{for } A \subset \Omega \text{ and } \theta_0 \in \Omega,$$

the loss experienced from choosing set A when the 'true' θ equals θ_0, and then use the Bayes rule (see Chapter 2) for a specified prior distribution. Hence confidence sets are adopted by statisticians who would rather not provide either a loss function or a prior distribution.

Without a prior distribution, confidence sets cannot make probabilistic statements about θ directly. Instead, they make them indirectly, with reference to the behaviour of random sets in parameter space.

Definition 6.2 (Confidence set and coverage). *$\mathcal{C}_\beta$ is a level β confidence set for θ exactly when $\mathcal{C}_\beta(y) \subset \Omega$ and*

confidence set

$$Pr\left\{t \in \mathcal{C}_\beta(Y) \,|\, \theta = t\right\} \geq \beta \quad \textit{for all } t \in \Omega.$$

The probability on the lefthand side is defined as the coverage *of $\mathcal{C}$ at t. If the coverage is exactly β for all t, then the confidence set is 'exact'.*

coverage

There is a close relationship between confidence sets and p-values; for every p-value, there is a confidence set (and *vice versa*).[7]

[7] 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 $H_0 : \theta = \theta_0$. See Section 6.6.

Theorem 6.8. *Let $p(y, \theta_0)$ be a p-value for the hypothesis $H_0 : \theta = \theta_0$. Then*

$$\mathcal{C}_\beta(y) := \{t \in \Omega : p(y,t) > 1 - \beta\}$$

is a level β confidence set for θ. If the p-value is exact, then the confidence set is exact as well.

Proof. This proof uses the subuniformity property of p-values.

$$\begin{aligned}\mathsf{Pr}\{t \in \mathcal{C}_\beta(Y) \mid \theta = t\} &= \mathsf{Pr}\{p(Y,t) > 1 - \beta \mid \theta = t\} \\ &= 1 - \mathsf{Pr}\{p(Y,t) \leq 1 - \beta \mid \theta = t\} \\ &\geq 1 - (1 - \beta) = \beta,\end{aligned}$$

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. □

Note that although the stochastic dominance property of p-values is not used explicitly, it is used implicitly to define a sensible confidence set. Values for t with small p-values are excluded from the confidence set because their odds ratios for H_0 versus decision-relevant departures from H_0 are likely to be smaller than one (Theorem 6.1).

Section 6.2.1 showed how p-values could be constructed from test statistics: the p-value for H_0 using test statistic s was denoted $p_s(\cdot, \theta_0)$. For a given test statistic, confidence sets constructed with p-value have two important properties.

First, from the construction it is immediate that $\beta \leq \beta'$ implies that $\mathcal{C}_\beta(y) \subset \mathcal{C}_{\beta'}(y)$, 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.

always nested

Second, such confidence sets are *transformation-invariant* (see Section 5.3.1). That is to say, if $g : \theta \mapsto \phi$ is an injective mapping and $\mathcal{C}^\theta_\beta$ is the confidence set in the parameterisation θ, and $\mathcal{C}^\phi_\beta$ in the parameterisation ϕ, then $\mathcal{C}^\phi_\beta(\boldsymbol{y}) = g\mathcal{C}^\theta_\beta(\boldsymbol{y})$.

transformation-invariant

Proof. If $H_0 : \theta = \theta_0$, then, in terms of ϕ, $H_0 : \phi = \phi_0$ where $\theta_0 = g^{-1}(\phi_0)$. Then the result follows immediately from (5.1) and

$$\begin{aligned}p^\phi_s(y, \phi_0) &= \mathsf{Pr}\left\{s(Y) \geq s(y) \mid \phi = \phi_0\right\} \\ &= \textstyle\sum_{y'} \left(s(y') \geq s(y)\right) f_{Y|\phi}(y' \mid \phi_0) \\ &= \textstyle\sum_{y'} \left(s(y') \geq s(y)\right) f_{Y|\theta}(y' \mid \theta_0) \\ &= p^\theta_s(y, \theta_0).\end{aligned}$$

□

6.4.1* *Interpretation of confidence sets*

The first thing about confidence sets is that there are as many ways to construct them as there are ways to construct p-values. This

means that one can construct confidence sets which are completely meaningless, by choosing a p-value which is meaningless. For example, use $\{y, \theta_0\}$ to seed a deterministic uniform random number generator, and then select, say, the thousandth iterate as $p(y, \theta_0)$. In this case $p(Y, \theta_0)$ is uniform for all $f_{Y|\theta}$ and all $\theta_0 \in \Omega$, and

$$\mathcal{C}_\beta(y) := \{t \in \Omega : p(y,t) > 1 - \beta\}$$

is an exact level β confidence set under all statistical models; see Figure 6.2. This confidence set is exact, nested, and transformation-invariant—what's not to like?! It is also meaningless.

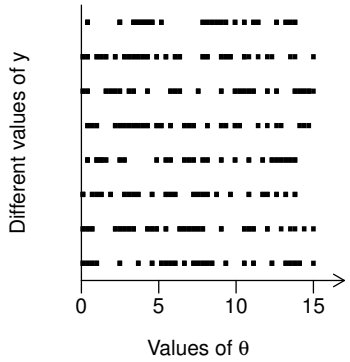


Figure 6.2: Meaningless exact 50% confidence sets for the parameter of a Poisson statistical model, for different values of y.

One way to protect against meaningless confidence sets is to enforce the stochastic dominance property of p-values. If strict stochastic dominance is required, then the distribution of the p-value must vary between the null model and decision-relevant alternatives. Better p-values have stronger stochastic dominance, and give rise to better confidence sets. Thus a statistician using a confidence set needs to be able to defend his particular choice of confidence set (e.g. his choice of test statistic) as a good choice among the infinity of possible choices, bearing in mind that this infinity includes many choices that are poor.

* * *

Second, a confidence set only promises a lower bound on coverage when averaging over the whole of $\mathcal{Y}$, but for particular y it may be unattractive: unbounded, say, or unconnected. However, in order for the coverage to be correct, the statistician must commit to a particular confidence set $\mathcal{C}_\beta$ *before* seeing $Y = y^{\text{obs}}$, and then report $\mathcal{C}_\beta(y^{\text{obs}})$ even though it may turn out that this particular set is unattractive. This has proved to be unpalatable for statisticians; see ch. 2 of Berger and Wolpert (1984) for more details.

One consequence has been the development of *conditional inference*, in which one conditions on *ancillary statistics*, representing a function of the observations that is independent of the parameters. But it has been difficult to formalise conditional inference into an acceptable principle, due to ambiguities about what constitutes an ancillary statistic (if such a thing exists); see Berger and Wolpert, *op. cit.*. Davison (2003, ch. 12) provides an update-to-date summary of conditional inference.

conditional inference
ancillary statistics

A Bayesian approach suffers none of these difficulties, because a Bayesian credible set is conditioned on y^{obs}; but then again, Bayesian credible sets make no guarantees about coverage.

* * *

Third, there is the issue of what happens when the statistician reports the confidence set to the client. Suppose that you tell the Minister that you have computed a 95% confidence interval for sea-level rise in 2100 to be $(0.25\,\text{m}, 0.85\,\text{m})$, and she says

> "OK, just to be clear, there is a 95% probability that sea-level rise will be greater than 0.25 m and less than 0.85 m."

To which you reply

> "Not quite Minister: $(0.25\,\mathrm{m}, 0.85\,\mathrm{m})$ is one realisation of a random interval that has at least a 95% probability of containing the true value of sea-level rise, no matter what that value happens to be"

She looks at you as though you have two heads, but you have no choice—this is the definition!

Non-statisticians expect Bayesian credible intervals (see, e.g., Rubin, 1984), and you—the statistician—should recognise this fact, and be extremely cautious about presenting confidence sets to non-statisticians. I doubt whether the client is interested in being 95% correct on average, over thousands of imaginary replications of the observations. She is much more concerned with *her* observations y^{obs}, and what they tell her about the parameters and the predictands. This requires a Bayesian analysis, and a prior distribution for the parameters.

6.5* *Quick confidence sets*

There is a practical issue when computing $\mathcal{C}_\beta(y^{\text{obs}})$. To enumerate the elements in this confidence set requires the calculation of a p-value for $H_0 : \theta = \theta_0$ for many candidate values of $\theta_0 \in \Omega$. In general each candidate requires a simulation, as described in Section 6.2.2. So computational cost is an issue.

This issue can be addressed by making a tractable choice of test statistic; in particular using the test statistic and p-value from (6.4) which does not require any simulation.[8] Unfortunately, however, the conditions of Theorem 6.5 will typically not hold. This means that the actual coverage of the resulting confidence set will not be the same as its nominal coverage β—the difference between these two is termed *level error*. So there needs to be a correction for level error.

[8] This is only possible when the score function is computable. The more general situation is discussed in Section 6.6, but this section should be read first.

level error

Let s be defined as in (6.4a), but with $I(\theta_0)$ in place of $nI_1(\theta_0)$, and p_s be defined as in (6.4b). Define the set

$$\mathcal{C}(y, \beta^*) := \left\{ t \in \Omega : p_s(y, t) > 1 - \beta^* \right\}.$$

For any y and t it is cheap to compute whether $t \in \mathcal{C}_\beta(y, \beta^*)$, because it is cheap to compute whether $p_s(y, t) > 1 - \beta^*$. But in general $\mathcal{C}_\beta(\cdot, \beta^*)$ will not have coverage of at least β^* for all $t \in \Omega$, because the conditions of Theorem 6.5 will not hold. So the simple idea is that β^* is adjusted until $\mathcal{C}(\cdot, \beta^*)$ has the desired coverage β in the region of parameter space favoured by the observations y^{obs}.

The tool for this adjustment is the *bootstrap*. Here is the algorithm.

bootstrap

1. Compute the ML estimate $\hat{\theta}(y^{\text{obs}})$, denote this as $\hat{\theta}^{\text{obs}}$.
2. Sample $y^1, \ldots, y^m \overset{\text{iid}}{\sim} f_{Y|\theta}(\cdot \mid \hat{\theta}^{\text{obs}})$, for some large m (say 1000).
3. Set $\beta^* \leftarrow \beta$.

4. Compute the empirical coverage

$$\begin{aligned}\text{coverage}(\beta^*) &:= m^{-1} \sum_{j=1}^{m} \left(\hat{\theta}^{\text{obs}} \in \mathcal{C}(y^j, \beta^*)\right) \\ &= m^{-1} \sum_{j=1}^{m} \left(p_s(y^j, \hat{\theta}^{\text{obs}}) > 1 - \beta^*\right).\end{aligned}$$

5. If $\text{coverage}(\beta^*) \approx \beta$ then go to step 6; otherwise, adjust β^* and go back to step 4.

6. Set

$$\mathcal{C}_\beta(\cdot) = \mathcal{C}(\cdot, \beta^*). \qquad (\dagger)$$

Eq. (†) will have coverage of almost exactly β at $\theta = \hat{\theta}^{\text{obs}}$, and—one hopes—coverage of close to β in a reasonably large region of Ω around $\hat{\theta}^{\text{obs}}$. The observed (approximately exact) level β confidence set $\mathcal{C}_\beta(y^{\text{obs}})$ can now be enumerated quickly.

The empirical coverage in step 4 approximates the true coverage at $\hat{\theta}$, according to the Weak Law of Large Numbers (WLLN, see Section A.4). As discussed in Section 4.2, the y's can be sampled from an SRS for $f_{X,Y|\theta}(\cdot \mid \hat{\theta})$, or for $f_{Y|\theta}(\cdot \mid \hat{\theta})$ if the joint distribution marginalises conveniently.

There are many approaches to using the bootstrap to construct confidence sets, reviewed in DiCiccio and Efron (1996). The approach outlined here is an example of what they term *bootstrap calibration*. Because there is an adjustment for level error, additional approximations are possible. In particular, because $\hat{\theta}^{\text{obs}}$ has been computed, the observed Fisher Information $\hat{J}(y^{\text{obs}})$ ought to be available (see Section 5.3.2 and Section 6.3.2). Replacing the expected Fisher Information with the observed Fisher Information simplifies the evaluation of $\mathcal{C}_\beta$ and $\mathcal{C}(\cdot, \beta^*)$ because then the denominator of (6.4a) does not depend on θ_0.

bootstrap calibration

6.6* *Nuisance parameters*

The calibration correction for level error in Section 6.5 is a general tool that can be used whenever it is possible to simulate from the statistical model. So it can be used to correct Bayesian set estimators for level error, and this is perhaps the most general way to construct confidence sets.

Suppose that the client is interested in some scalar function of the parameters, $\phi := g(\theta)$. This is a very common situation. The full parameter set θ is necessary to construct a statistical model which adequately represents the judgements of the client, but only a subset of the parameters is decision relevant, or some function of the parameters—I have just used a scalar function here but the results generalise immediately.[9] Convex confidence sets for scalar parameters are termed *confidence intervals*. The other functions of the parameters that are required to augment g into an injective function are termed *nuisance parameters*. In the simplest case the

[9] For a vector function, the credible interval below would need to be replaced by a high posterior density set (see, e.g. Tanner, 1996).

confidence intervals
nuisance parameters

client might be interested in, say, θ_1, and then $\theta_2, \ldots, \theta_d$ would be nuisance parameters.

The simple idea is to construct a Bayesian interval estimator of ϕ, and then correct for level error using bootstrap calibration. It is necessary to specify a prior distribution π_θ. As this is effectively a device, a rule-based specification may suffice (see, e.g., Kass and Wasserman, 1996). But it seems sensible and defensible to incorporate judgements about θ where they are widely shared. Then generate an SRS for π^*_θ, the posterior distribution of the parameters. This can immediately be transformed into an SRS for π^*_ϕ, according to Theorem 4.1. Using this SRS, compute the level β equi-tailed credible interval, which is the interval defined by the $(1-\beta)/2$ and $(1+\beta)/2$ quantiles of the SRS (see Section 4.4); denote these quantiles ℓ and u.

Now you have a level β credible interval for ϕ, based on y^{obs}. You can tell the Minister that there is a $100\beta\%$ probability that ϕ lies in the range $[\ell, u]$, based on the statistical model, the prior distribution, and the observations. It is just possible that the Minister will ask "and what is the coverage of this interval?" At this point you can do the calibration calculation of Section 6.5. Starting from the Bayes estimate (the posterior expectation of θ), generate a large set of candidate observations. Compute a level β equi-tailed credible interval for each one, and count the proportion that contain the Bayes estimate. This proportion is an estimate of the coverage at the Bayes estimate and—one hopes—in a reasonably large region around this estimate as well.[10]

[10] This is a computationally intensive calculation, because each candidate requires an SRS, but it can be performed in parallel.

Fingers crossed, there will only be a small discrepancy between β and the estimated coverage. If there was a large discrepancy, I would initially suspect a computing error. Having ruled this out, I would subject the statistical model and the prior distribution to more stringent checking—see Section 4.4. A more pragmatic response would be to try a different (flatter?) prior distribution.[11]

[11] There is a class of prior distributions called *matching priors* that are designed to give level β credible intervals with coverage of approximately β, but these prior distributions only exist for a small subset of statistical models; see Datta and Sweeting (2005).

6.6.1 P-values for composite hypotheses*

The hypothesis

$$H_0 : \phi = \phi_0$$

is an example of a *composite hypothesis*, because it does not completely determine the distribution of Y. This section has shown how to derive confidence intervals for ϕ, and these can be turned into a p-value for H_0, using the duality of p-values and confidence sets described in Theorem 6.8. Simply adjust the confidence level β until ϕ_0 lies on the boundary of its confidence interval. Then $1-\beta$ is a p-value for H_0. Similarly, one could compute a *Bayesian p-value* by using a Bayesian credible interval in place of a confidence interval.

Bayesian p-value

A
More on expectation and probability

Here is a *smörgåsbord* of additional results about expectation and probability.

A.1 Variances and covariances

Specifying my expectation for X describes only a single aspect of my judgements about X. I can increase the depth of my judgements about X by specifying expectations for non-linear functions of X. There are several approaches, but one in particular has proved mathematically tractable, which is to specify judgements about powers of X. The value $\mathsf{E}(X^r)$ is termed the *rth moment*, for $r = 1, 2, \ldots$; thus the first moment is the expectation. The second moment in particular plays an important role, because of its attractive theoretical properties. This is the *variance*. For vector random quantities, this generalises to the *covariance*.

*r*th moment

A.1.1 Variance and standard deviation

The *variance* of X is defined as

variance

$$\mathrm{Var}(X) := \mathsf{E}\left[\{X - \mathsf{E}(X)\}^2\right],$$

and the *standard deviation* of X,

standard deviation

$$\mathrm{Sd}(X) := \sqrt{\mathrm{Var}(X)}.$$

Both the variance and the standard deviation are well-defined provided that X is square integrable, as can be seen by multiplying out:

$$\{X - \mathsf{E}(X)\}^2 = X^2 - 2X\,\mathsf{E}(X) + \{\mathsf{E}(X)\}^2,$$

and hence $\mathrm{Var}(X) = \mathsf{E}(X^2) - \{\mathsf{E}(X)\}^2$. The standard deviation is the more intuitive of the two, having the same units as X itself. It has acquired much familiarity from its role as a standard parameter of the Normal distribution.

A.1.2 Covariance and correlation

I may judge two random quantities X and Y to be related, in the sense that my expectation for the random quantity

$$\{X - \mathsf{E}(X)\}\{Y - \mathsf{E}(Y)\}$$

is positive, or negative. Thus define the *covariance* of X and Y, covariance

$$\mathrm{Cov}(X, Y) := \mathsf{E}\left[\{X - \mathsf{E}(X)\}\{Y - \mathsf{E}(Y)\}\right],$$

implying that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$, and $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. Multiplying out,

$$\mathrm{Cov}(X, Y) = \mathsf{E}(XY) - \mathsf{E}(X)\,\mathsf{E}(Y),$$

so that the Schwarz inequality (Theorem 1.2) implies that $\mathrm{Cov}(X, Y)$ is finite if X and Y are both square integrable. Also,

$$\mathrm{Cov}(X, Y) = 0 \iff \mathsf{E}(XY) = \mathsf{E}(X)\,\mathsf{E}(Y).$$

The covariance is particularly tractable for specifying judgements about linear functions of random quantities. It is easy to check that if a, b, c, d are constants, then

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

so covariances (and variances as a special case) are invariant to shifts (a and c) but not to scalings (b and d). One special case is $\mathrm{Cov}(a, Y) = 0$, setting $b = 0$. More generally,

$$\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z).$$

This result can be iterated for any finite sum. In particular,

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i) + \sum_{ij} \mathrm{Cov}(X_i, X_j). \tag{A.1}$$

The unit of $\mathrm{Cov}(X, Y)$ is the product of the units of X and the units of Y, which is not very intuitive. The *correlation coefficient* is a unitless transformation of the covariance, defined as correlation coefficient

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\mathrm{Sd}(X)\,\mathrm{Sd}(Y)} \quad \text{if } \mathrm{Sd}(X)\,\mathrm{Sd}(Y) > 0,$$

and undefined otherwise. This is invariant to shifts and scalings. The correlation is bounded between -1 and 1, according to the Schwarz inequality.

Proof. Since covariances are invariant to shifts, we may take X and Y to have expectation zero, without loss of generality. Then $\mathrm{Cov}(X, Y) = \mathsf{E}(XY)$ and $\mathrm{Var}(X) = \mathsf{E}(X^2)$. By the Schwarz inequality, then,

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

Taking square roots implies

$$|\mathrm{Cov}(X, Y)| \leq \mathrm{Sd}(X)\,\mathrm{Sd}(Y),$$

from which $-1 \leq \mathrm{Corr}(X, Y) \leq 1$ follows directly, provided that the righthand side is positive. □

The next result sheds light on the interpretation of the correlation.

Theorem A.1. *If $Y \overset{\text{ms}}{=} a + bX$ for some real a, b then $|Corr(X, Y)| = 1$.*

Proof. Without loss of generality, let Y and $a + bX$ both have expectation zero. Then

$$\begin{aligned}\mathsf{E}\left[\{Y - (a+bX)\}^2\right] &= \mathsf{E}(Y^2) - 2\,\mathsf{E}\{Y(a+bX)\} + \mathsf{E}\{(a+bX)^2\} \\ &= \mathrm{Var}(Y) - 2\,\mathrm{Cov}(Y, a+bX) + \mathrm{Var}(a+bX) \\ &= \mathrm{Var}(Y) - 2b\,\mathrm{Cov}(Y, X) + b^2\,\mathrm{Var}(X).\end{aligned}$$

This must be equal to zero for all b, and hence

$$\{2\,\mathrm{Cov}(Y, X)\}^2 - 4\,\mathrm{Var}(Y)\,\mathrm{Var}(X) = 0$$

from which the result follows directly. □

Thus correlation can be interpreted as a measure of the strength of the linear association between two random quantities, with ± 1 indicating an effectively perfect linear association, and 0 indicating no linear association.

We say that X and Y are *uncorrelated* exactly when $\mathrm{Cov}(X, Y) = 0$. If X and Y are uncorrelated then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, from (A.1). This result can be iterated to any finite sum of *mutually uncorrelated* random quantities, for which $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$. If X and Y are probabilistically independent (see Section 3.1) then they are uncorrelated, but the converse is not true.

uncorrelated

mutually uncorrelated

A.2 Inequalities

The axioms of expectation give rise to some very powerful inequalities regarding expectations of functions of random quantities. Schwarz's inequality has already been given (Theorem 1.2). Here are some others.

Theorem A.2 (Jensen's inequality)**.** *If $\boldsymbol{x} := (x_1, \ldots, x_n)$ and g is any convex function of $\boldsymbol{x}$, then $E\{g(\boldsymbol{X})\} \geq g(E\{\boldsymbol{X}\})$.*

Proof. If g is convex then there is a supporting hyperplane through every point $g(\boldsymbol{x})$. Let $\bar{\boldsymbol{x}} := \mathsf{E}(\boldsymbol{X})$, and denote the supporting hyperplane through $g(\bar{\boldsymbol{x}})$ as $h(\boldsymbol{x}) := g(\bar{\boldsymbol{x}}) + \boldsymbol{a}^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$, for some $\boldsymbol{a} \in \mathbb{R}^n$; see Figure A.1. Then since $g(\boldsymbol{x}) \geq h(\boldsymbol{x})$ for all $\boldsymbol{x}$, so $\mathsf{E}\{g(\boldsymbol{X})\} \geq \mathsf{E}\{h(\boldsymbol{X})\}$ by monotonicity, and the result follows by linearity:

$$\mathsf{E}\{g(\boldsymbol{X})\} \geq \mathsf{E}\{g(\bar{\boldsymbol{x}}) + \boldsymbol{a}^T(\boldsymbol{X} - \bar{\boldsymbol{x}})\} = g(\bar{\boldsymbol{x}}) + \boldsymbol{a}^T(\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}) = g\{\mathsf{E}(\boldsymbol{X})\}.$$

□

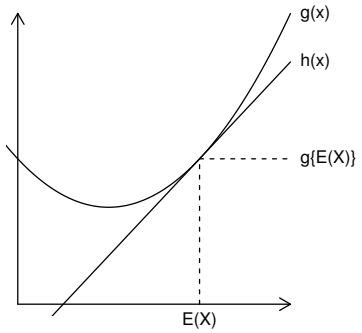


Figure A.1: Supporting hyperplane for a convex function

Theorem A.3 (Monotonicity of norms)**.** *Let $1 \leq p \leq r$. If $E(|X|^r)$ is finite then*

$$E(|X|^p)^{1/p} \leq E(|X|^r)^{1/r}.$$

Proof. As $\mathsf{E}(|X|^r)$ is finite by hypothesis, it is only necessary to show that

$$\mathsf{E}(|X|^p)^{r/p} \leq \mathsf{E}(|X|^r).$$

But $g(x) = |x|^{r/p}$ is a convex function and so

$$\begin{aligned}
\mathsf{E}(|X|^p)^{r/p} &= |\mathsf{E}(|X|^p)|^{r/p} && \text{by non-negativity} \\
&\leq \mathsf{E}(||X|^p|^{r/p}) && \text{by Jensen's inequality} \\
&= \mathsf{E}(|X|^{pr/p}) = \mathsf{E}(|X|^r)
\end{aligned}$$

as was to be shown. □

Here are some inequalities which link expectation and probability.

Theorem A.4 (Generalised Markov's inequality)**.** *If g is an increasing non-negative function of x, then*

$$\mathsf{Pr}(X \geq a) \leq \frac{E\{g(X)\}}{g(a)}.$$

Proof. If $g(a) = 0$ then this is certainly true, so let $g(a) > 0$. Because g is increasing, $X \geq a$ if and only if $g(X) \geq g(a)$. Let $Y := g(X)$ and $b := g(a)$. Then

$$\begin{aligned}
\mathsf{Pr}(X \geq a) &= \mathsf{Pr}(Y \geq b) \\
&= \mathsf{E}(Y \geq b) && \text{by definition} \\
&= \mathsf{E}(Y/b \geq 1) && \text{as } b > 0 \\
&\leq \mathsf{E}(Y/b) && \text{as } Y \geq 0 \text{ implies } (Y/b \geq 1) \leq Y/b \\
&= \mathsf{E}(Y)/b
\end{aligned}$$

which proves the result, on substituting for Y and b. □

The original Markov's inequality states that if Y is non-negative then $\mathsf{Pr}(Y \geq b) \leq \mathsf{E}(Y)/b$. Other famous inequalities follow from this one, such as Chebyshev's Inequality.

Theorem A.5 (Chebyshev's inequality)**.** *If X is a random quantity with expectation μ and variance σ^2 then*

$$\mathsf{Pr}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof. Since $|X - \mu| \geq a \iff (X - \mu)^2 \geq a^2$, so

$$\mathsf{Pr}(|X - \mu| \geq a) = \mathsf{Pr}\{(X - \mu)^2 \geq a^2\} \leq \frac{\mathsf{E}\{(X - \mu)^2\}}{a^2} = \frac{\sigma^2}{a^2}$$

where the inequality is the original Markov's inequality. □

Note that Markov's and Chebyshev's inequalities are tight (Whittle, 2000, ch. 15), so that one cannot get a better upper bound on the probability without constraining the distribution of X further. Here is one approach to getting a better bound with additional information about X.

Theorem A.6 (Tail probabilities)**.** *If $a > 0$,*

$$Pr\{|X| \geq a\} \leq \min_{r>0} \frac{E\{|X|^r\}}{a^r}.$$

Proof. Follows immediately from the generalised Markov's inequality (Theorem A.4) with $g(x) = x^r$, which must hold for all $r > 0$. □

A.3 The Probability Integral Transform (PIT)

Let $X \in \mathcal{X} \subset \mathbb{R}$ be a scalar random quantity with distribution function $F_X : \mathcal{X} \to [0,1]$. Define a new random quantity $Y := F_X(X)$; i.e. Y is the random quantity one gets by putting X into its own distribution function. It is a very useful fact that Y has a *subuniform distribution*, i.e.

subuniform distribution

$$F_Y(u) \leq u \quad \text{for all } u \in [0,1],$$

and that $F_Y(u) = u$ 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 $x \in \mathcal{X}$:

$$F_Y(u) = \mathsf{Pr}\{F_X(X) \leq F_X(x)\} = \mathsf{Pr}\{X \leq x\} = F_X(x) = u.$$

The 'cancellation' of F at the second equality occurs because of the bijective relationship between x and $F(x)$ for $x \in \mathcal{X}$.[1] This proves the second part of the claim: in the case where X is a continuous random quantity, the points u in $(0,1)$ are in a bijective relationship with the points x in $\mathcal{X}$, and Y is uniformly distributed.

[1] Technical note: here we can ignore points in $\mathcal{X}$ that have zero probability.

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 A.2. For a continuous random quantity the steps are infinitesimally small, and the distribution function and the 45° line coincide.

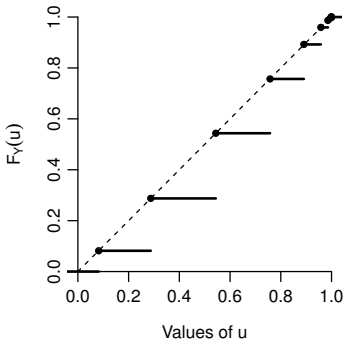


Figure A.2: Distribution function of $Y = F_X(X)$, where $X \sim \text{Poisson}(\lambda = 2.5)$.

A.4 Convergence of random quantities

This is a very brief outline of some useful results about probabilistic convergence. Details and proofs can be found in Grimmett and Stirzaker (2001, chs 5 and 7).

There are several different types of convergence for sequences of random quantities. Two if these are:

1. *Convergence in probability*: $X_n \xrightarrow{\mathrm{P}} X$, exactly when

Convergence in probability

$$\lim_{n\to\infty} \mathsf{Pr}\{|X_n - X| \geq \varepsilon\} = 0$$

for all $\varepsilon > 0$.

2. *Convergence in distribution*: $X_n \xrightarrow{\mathrm{D}} X$, exactly when

Convergence in distribution

$$\lim_{n\to\infty} F_{X_n}(x) \longrightarrow F_X(x)$$

for all x for which F_X is continuous. An equivalent condition is that $\lim_{n\to\infty} \mathsf{E}\{g(X_n)\} = \mathsf{E}\{g(X)\}$ for all bounded continuous g.

Here are two results which link these types of convergence.

Theorem A.7.

1. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$
2. $X_n \xrightarrow{P} c \iff X_n \xrightarrow{D} c$, *where c is a constant.*

Theorem A.8 (Slutsky's theorem)**.** *If $X_n \xrightarrow{D} c$, where c is a constant, and $Y_n \xrightarrow{P} Y$,*

1. $X_n + Y_n \xrightarrow{D} c + Y$,
2. $X_n Y_n \xrightarrow{D} cY$.

A.4.1 Convergence in probability

The *Weak Law of Large Numbers (WLLN)* is an example of convergence in probability.[2]

Weak Law of Large Numbers (WLLN)

[2] See Section A.1.2 for the definition of 'mutually uncorrelated'.

Theorem A.9 (Weak Law of Large Numbers)**.** *If $X_1, X_2, \ldots$ is an infinite sequence of mutually uncorrelated random quantities with $\mathsf{E}(X_i) = \mu$ and $Var(X_i) < \infty$, then $\bar{X}_n \xrightarrow{P} \mu$, where $\bar{X}_n := n^{-1}\sum_{i=1}^{n} X_i$, the sample mean.*

Proof. Let $\mathrm{Var}(X_i) = \sigma^2$. From the properties of the expectation and variance (Section A.1.2), $\mathsf{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. Then, by Chebyshev's Inequality (Theorem A.5),

$$\mathsf{Pr}\left\{|\bar{X}_n - \mu| \geq \varepsilon\right\} \leq \frac{\sigma^2/n}{\varepsilon^2},$$

and the righthand side tends to zero in n, for all $\varepsilon > 0$. □

A different version of the WLLN requires the stronger condition that $X_1, X_2, \ldots$ is an IID sequence,[3] but then allows for $\mathrm{Var}(X_i)$ to be infinite, as long as $\mathsf{E}(X_i)$ is finite. This is proved using characteristic functions.

[3] See Section 3.1 for the definition of IID.

Two useful properties of $\xrightarrow{\mathrm{P}}$ are given in the next result.

Theorem A.10.

1. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X + Y$.*
2. *If g is continuous and $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.*

A.4.2 Convergence in distribution

Central Limit Theorem (CLT)

The *Central Limit Theorem (CLT)* is an example of convergence in distribution. It states that if $X_1, X_2, \ldots$ is an infinite IID sequence for which $\mathsf{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1), \tag{A.2}$$

where $\bar{X}$ is the sample mean (see Theorem A.9). For convergence in distribution it is common to write a distribution on the right-hand side, as shown in (A.2), rather than a Y with a specified distribution.

The CLT can be elegantly proved using characteristic functions; see Grimmett and Stirzaker (2001, chapter 5). There is a multivariate CLT as well, for an infinite IID sequence of vectors. In this case, if X_i is a k-vector, $\mathsf{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \Sigma$, then

$$\sqrt{n}\, Q^{-T}(\bar{X}_n - \mu) \xrightarrow{\mathrm{D}} \mathrm{N}_k(\mathbf{0}, I)$$

where Q is any matrix satisfying $Q^T Q = \Sigma$.

The Normal distribution is closed under linear transformations. Therefore, although technically incorrect, it is understandable to write these two results as

$$\bar{X}_n \xrightarrow{\mathrm{D}} \mathrm{N}(\mu, \sigma^2/n) \quad \text{and} \quad \bar{X}_n \xrightarrow{\mathrm{D}} \mathrm{N}(\mu, n^{-1}\Sigma).$$

B
Bibliography

W.P. Aspinall, 2010. A route to more tractable expert advice. *Nature*, **463**, 294–295. 70

W.P. Aspinall and R.M. Cooke, 2012. 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. Forthcoming. 70

D. Basu, 1975. Statistical information and likelihood. *Sankhyā*, **37**(1), 1–71. With discussion. 10

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`. 99

J.O. Berger, 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, Inc., NY, USA, second edition. 31, 37, 38

J.O. Berger, 2006. The case for objective Bayesian analysis. *Bayesian Analysis*, **1**(3), 385–402. 45

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). 39, 72

J. Besag, 1974. Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**(2), 192–236. 53

J. Besag and P. Clifford, 1989. Generalized Monte Carlo significance tests. *Biometrika*, **76**(4), 633–642. 93

J. Besag, P. Green, D. Higdon, and K. Mengerson, 1995. Bayesian computation and stochastic systems. *Statistical Science*, **10**(1), 3–41. With discussion 42–66. 66

P. Billingsley, 1995. *Probability and Measure*. John Wiley & Sons, Inc., New York NY, USA, third edition. 5, 10

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. 70

G. Casella and R.L. Berger, 2002. *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 77, 91, 94

R. T. Clemen, 1996. *Making Hard Decisions*. Pacific Grove, CA: Duxbury Press, 2nd edition. 41

R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter, 1999. *Probabilistic Networks and Expert Systems*. New York: Springer. 51, 55, 60, 70

D.R. Cox and D.V. Hinkley, 1974. *Theoretical Statistics*. London: Chapman and Hall. 31, 77, 83

N. Cressie and C.K. Wikle, 2011. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Inc., Hoboken NJ, USA. 54

G.S. Datta and T.J. Sweeting, 2005. Probability matching priors. In *Bayesian Thinking, Modelling and Computation*, number 25 in Handbook of Statistics, pages 91–114. Elsevier B.V., Amsterdam, Netherlands. 102

A.C. Davison, 2003. *Statistical Models*. Cambridge, UK: Cambridge University Press. 49, 54, 99

A.P. Dawid, 1984. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, **147**(2), 278–290. With discussion, 290–292. 70

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, 57

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. 112

B. de Finetti, 1972. *Probability, Induction and Statistics*. London: John Wiley & Sons. 5

B. de Finetti, 1974. *Theory of Probability*, volume 1. London: Wiley. 9, 11, 12, 13

B. de Finetti, 1974/75. *Theory of Probability*. London: Wiley. Two volumes (2nd vol. 1975); A.F.M. Smith and A. Machi (trs.). 5

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. 89

M.H. DeGroot, 1986. *Probability and Statistics*. Reading, Mass.: Addison-Wesley Publishing Co., 2nd edition. 32

P. Diaconis and D. Freedman, 1986. On the consistency of bayes estimators. *The Annals of Statistics*, **14**(1), 1–26. 38

T.J. DiCiccio and B. Efron, 1996. Bootstrap confidence intervals. *Statistical Science*, **11**(3), 189–212. with discussion and rejoinder, 212–228. 101

T. L. Edwards and P.G. Challenor, 2013. Risk and uncertainty in hydrometeorological hazards. In Rougier *et al.* (2013), chapter 5, pages 100–150. 68

W. Edwards, H. Lindman, and L.J. Savage, 1963. Bayesian statistical inference for psychological research. *Psychological Review*, **70**(3), 193–242. 90

B. Efron and D.V. Hinkley, 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**(3), 457–482. 97

B. Efron and C. Morris, 1977. Stein's paradox in statistics. *Scientific American*, **236**(5), 119–127. 83

A.E. Gelfand and D. Dey, 1994. Bayesian model choice: Asymptotic and exact calculations. *Journal Royal Statistical Society B*, **56**, 501–514. 67

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, 2003. *Bayesian Data Analysis*. Boca Raton, Florida: Chapman and Hall/CRC, 2nd edition. 55, 66, 70

G. Gigerenzer, 2003. *Reckoning with Risk: Learning to Live with Uncertainty*. Penguin. 90

M. Goldstein, 2006. Subjective Bayesian analysis: Prinicples and practice. *Bayesian Analysis*, **1**(3), 403–420. 45

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. John Wiley & Sons, Chichester, UK. 5

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. 90

G.R. Grimmett and D.R. Stirzaker, 2001. *Probability and Random Processes*. Oxford, UK: Oxford University Press, 3rd edition. 5, 16, 107, 109

D. Heath and W. Sudderth, 1976. De Finetti's theorem on exchangeable variables. *The American Statistician*, **30**(4), 188–189. 57

C. Howson and P. Urbach, 2006. *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court Publishing Co., 3rd edition. 26

H. Jeffreys, 1961. *Theory of Probability*. Oxford, UK: Oxford University Press, 3rd edition. 47

J. Jesus and R.E. Chandler, 2011. Estimating functions and the generalized method of moments. *Interface Focus*, **1**, 871–885. 80

J.B. Kadane, 2011. *Principles of Uncertainty*. Chapman & Hall/CRC Press, Boca Raton FL, USA. 10

R.E. Kass and L. Wasserman, 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370. 45, 102

J.F.C. Kingman, 1978. Uses of exchangeability. *The Annals of Probability*, **6**(2), 183–197. 57

F. Lad, 1996. *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 5, 7, 10, 13, 56

F. Lindgren, H. Rue, and J. Lindström, 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, **73**(4), 423–498. 54

D.V. Lindley, 1985. *Making Decisions*. London: John Wiley & Sons, 2nd edition. 25

D.V. Lindley, 1991. Subjective probability, decision analysis and their legal consequences. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **154**(1), 83–92. 47

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. 55, 56, 66

K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. Harcourt Brace & Co., London. 72, 95

X.-L. Meng, 1994. Posterior predictive *p*-values. *The Annals of Statistics*, **22**(3), 1142–1160. 70

C.N. Morris, 1983. Parametric Empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, **78**, 47–55. With discussion. 74

J. Nocedal and S.J. Wright, 2006. *Numerical Optimization*. New York: Springer, 2nd edition. 82

Yudi Pawitan, 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press. 80

J. Pearl, 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. 58

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. 6

H. Raiffa and R. Schlaifer, 1961. *Applied Statistical Decision Theory*. MIT Press, Cambridge, Mass. 39

C.P. Robert, 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer. 31

C.P. Robert and G. Casella, 2004. *Monte Carlo Statistical Methods*. Springer, New York NY, 2nd edition. 55, 56, 66

K.J. Rothman, S. Greenland, and T.L. Lash, editors, 2008. *Modern Epidemiology*. Lippincott Williams & Wilkins, Philadelphia PA, USA, third edition. 48, 85

J.C. Rougier, R.S.J. Sparks, and L.J. Hill, editors, 2013. *Risk and Uncertainty Assessment for Natural Hazards*. Cambridge University Press, Cambridge, UK. 68, 113

D.B. Rubin, 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**(4), 1151–1172. 70, 100

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. 52, 54, 81

L.J. Savage, 1954. *The Foundations of Statistics*. Dover, New York, revised 1972 edition. 32, 78

L.J. Savage *et al.*, 1962. *The Foundations of Statistical Inference*. Methuen, London. 48, 97

M.J. Schervish, 1995. *Theory of Statistics*. New York: Springer. Corrected 2nd printing, 1997. 56, 57, 77

J.Q. Smith, 2010. *Bayesian Decision Analysis: Principle and Practice*. Cambridge University Press, Cambridge, UK. 31, 39, 40, 51

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206. University of California Press, 1956. 83

M. Stone, 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**(2), 111–147. 70

M.A. Tanner, 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer-Verlag, New York, Inc. 69, 101

C. Varin, N. Reid, and D. Firth, 2011. An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5–42. 81

W. N. Venables and B.D. Ripley, 2002. *Modern Applied Statistics with S*. New York: Springer-Verlag, fourth edition. 68

P. Whittle, 1996. *Optimal Control: Basics and Beyond*. John Wiley & Sons, Chichester, UK. 40

P. Whittle, 2000. *Probability via Expectation*. New York: Springer, 4th edition. 5, 18, 20, 33, 106

D. Williams, 1991. *Probability With Martingales*. Cambridge University Press, Cambridge, UK. 5

G. Woo, 2011. *Calculating Catastrophe*. Imperial College Press, London, UK. 68