APTS Lecture Notes on Statistical Inference

Jonathan Rougier

Contents

1
Expectation and probability

This is a summary of the main concepts and results in probability for Statistics. My objective is to give precise definitions and notation (our profession's notation is rather 'fluid'), and enough detail to reconstruct the proofs of the main results. I also want to correct a few misconceptions, which have blurred the fundamentally different natures of probability theory and statistical inference.

1.1 Random quantities and expectations

A *random quantity* represents a sequence of operations which will result in a value; real-valued functions of random quantities are also random quantities. In other words, all random quantities should have operational definitions. Statistics is about making inferences about random quantities which have not been observed, based on the values of those that have. The bridge between what we have and what we want is provided by our beliefs. Expectations and probabilities are a way of quantifying our beliefs.

A random quantity is typically denoted X, Y, or Z, often with subscripts; specified functions are typically denoted as g or h.[1] The set of possible values X can take is its *realm*, denoted $\mathcal{X} \subset \mathbb{R}$. Any particular specified value of $\mathcal{X}$ is denoted x. Where it is necessary to enumerate $\mathcal{X}$, I write

$$\mathcal{X} := \left\{x^{(1)}, \ldots, x^{(r)}\right\} \subset \mathbb{R},$$

and similarly for other letters (e.g. $\mathcal{Y}$ as the realm for Y). A random quantity whose realm contains only a single value is a *constant*, typically denoted by a lower-case letter from the top of the alphabet, such as a, b, or c.

By its operational definition, a random quantity has a finite realm, and is therefore bounded. But it is sometimes convenient to treat the realm as countably infinite, or even uncountable. In these convenient extensions it is the responsibility of the statistician to ensure that no pathologies are introduced.[2] Avoiding the pathologies of an uncountable realm is why formal probability theory is so complicated, but in most of this chapter I will treat all realms as finite, as nature intended. Generalisations are given in Sec. 1.6.

[1] The symbol 'f' is reserved for a statistical model, see Chapter 2.

[2] I term this the Principle of Excluding Pathologies, PEP.

A collection of random quantities is denoted $\boldsymbol{X} := (X_1, \dots, X_m)$. The joint realm is $\mathcal{X}$ and any particular specified value is $\boldsymbol{x} := (x_1, \dots, x_m)$. The joint realm is necessarily a subset of the product of the individual realms,

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

Where it is necessary to enumerate $\mathcal{X}$, I write

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

Assertions about random quantities are statements that hold as a consequence of their definitions; therefore they hold everywhere on the joint realm. Thus if the definition of X_1 and X_2 implies that $X_1 \leq X_2$, then $x_1^{(j)} \leq x_2^{(j)}$ for all $j = 1, \dots, r$. For our convenience, the joint realm may be extended to, say, the product of the individual realms, which would include elements for which $x_1^{(j)} > x_2^{(j)}$. In this case, our beliefs would need to be augmented to ensure that the probability attached to such elements is exactly zero (see Sec. 1.2).

1.1.1 *The axioms of Expectation*

There is a long-running debate about whether expectation or probability should be the primitive concept when quantifying beliefs about $\boldsymbol{X}$. I strongly favour the former. My *expectation* of a random quantity X, denoted $\mathrm{E}(X)$, is my 'best guess' for X, represented as a value in $\mathbb{R}$. In Statistics, unlike in probability theory, it is important to have some idea about what formal concepts actually mean, so that when I think about my "expectation of sea-level rise in 2100" this conjours up a number in my mind. 'Best guess' seems to work quite well.

I refer to my expectations about $\boldsymbol{X}$ and functions of $\boldsymbol{X}$ as as my *beliefs* about $\boldsymbol{X}$. My beliefs about $\boldsymbol{X}$ at time t depend on my *disposition* at time t: all the things I have learnt and thought about up to time t, the things I have forgotten, my general attitude, and even my current state of mind. Beliefs change from day to day—that's just the way it is, and we should not attempt to deny or conceal it. It is not, for example, a characteristic only of 'bad' scientists that their beliefs are subjective and contingent. Of course some beliefs hardly change, and, moreover, are very common. For example, the belief that the diversity of living things is due to genetic variation and heredity, and selection pressure. But the interesting scientific questions lie at the next level down: why, for example, does sexual reproduction convey a selection advantage? On this topic, the beliefs of biologists are diverse, and prone to changing.

It may be intuitive, but 'best guess' is just a heuristic for expectation. The *theory* of expectation is about a special type of 'best guess': one that is *coherent*.

Definition 1 (Coherent expectations)**.** *Expectations for X and Y are* pairwise coherent *exactly when they satisfy the two properties:*

1. Lower boundedness*:* $\mathrm{E}(X) \geq \min \mathcal{X}$*, and* $\mathrm{E}(Y) \geq \min \mathcal{Y}$*.*

2. Finite additivity*:* $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$*.*

Expectations for $\boldsymbol{X}$ *are* completely coherent *exactly when these two properties hold for all pairs of random quantities that can be defined on* $\boldsymbol{\mathcal{X}}$*.*

This is a common approach in modern mathematics: not to say what a thing is or means, but how it behaves.[3]

There are only these two axioms, but they imply a very rich set of additional constraints on expectations, and on probabilities (see Sec. 1.2). Here are some immediate important implications of complete coherence, which are straightforward to prove. First,

$$\mathrm{E}(a_1 X_1 + \cdots + a_m X_m) = a_1 \mathrm{E}(X_1) + \cdots + a_m \mathrm{E}(X_m), \qquad \text{(LIN)}$$

where $a_1, \dots, a_m$ are constants.[4] Second,

$$\mathrm{E}(a) = a \qquad \text{(Normalisation)}$$

if a is a constant. Third,

$$X \leq Y \implies \mathrm{E}(X) \leq \mathrm{E}(Y), \qquad \text{(Monotonicity)}$$

with the immediate implication that

$$\min \mathcal{X} \leq \mathrm{E}(X) \leq \max \mathcal{X}. \qquad \text{(Convexity)}$$

Fourth, *Schwartz's inequality*

$$\mathrm{E}(XY)^2 \leq \mathrm{E}(X^2)\,\mathrm{E}(Y^2), \qquad \text{(SIQ)}$$

see Williams (1991, sec. 6.8) for a short and elegant proof. Fifth, *Jensen's inequality*: if $g : \mathbb{R}^m \to \mathbb{R}$ is a convex function,[5] then

$$\mathrm{E}\{g(\boldsymbol{X})\} \geq g(\mathrm{E}\{\boldsymbol{X}\}). \qquad \text{(JEN)}$$

There is a straightforward proof based on the Supporting Hyperplane Theorem, see Thm 3.5. Schwartz's inequality (and its generalisation the Cauchy-Schwartz inequality) and Jensen's inequality are two of the most important inequalities in the whole of mathematics.[6]

1.1.2 The Fundamental Theorem of Prevision

Coherence as defined in Def. 1 has a complicated aspect. On the one hand, it is a very simple and appealing property for a pair of random quantities. On the other, who knows how much extra structure is imposed through the extention to all pairs of random quantities? Bruno de Finetti (1974, ch. 3) provided the crucial result.[7]

[3] I discuss difficulties with meaning in Sec. 1.3.1. For an excellent summary of modern mathematics, see Gowers (2002).

[4] Slightly tricky. Use additivity to prove that $\mathrm{E}(aX) = a\,\mathrm{E}(X)$ when a is a positive integer. Then use $\mathrm{E}\{(a/a)X\} = a\,\mathrm{E}\{X/a\}$ to prove that $\mathrm{E}(qX) = q\,\mathrm{E}(x)$ for any positive rational. It is straightforward to show that $\mathrm{E}(-X) = -\mathrm{E}(X)$, and so $\mathrm{E}(qX) = q\,\mathrm{E}(X)$ holds for all rationals. Then complete the argument from the rationals to the reals in the usual way.

[5] Technically, a convex function on the convex hull of $\boldsymbol{\mathcal{X}}$.

[6] Although you would have to read, say, Gowers *et al.* (2008) to substantiate this claim.

[7] See also Lad (1996, ch. 2) and Whittle (2000, ch. 15).

Some terms. A *convex combination* $(w_1, \dots, w_r)$ has $w_j \geq 0$ for each j, and $\sum_{j=1}^{r} w_j = 1$. The set of all convex combinations is the $(r-1)$-dimensional *unit simplex*, or just $(r-1)$-simplex,

$$\mathbb{S}^{r-1} := \left\{ \boldsymbol{w} \in \mathbb{R}^r : w_j \geq 0, \sum\nolimits_j w_j = 1 \right\}. \tag{1.1}$$

Theorem 1.1 (Fundamental Theorem of Prevision, FTP). *Let $\boldsymbol{X} := (X_1, \dots, X_m)$ be a collection of random quantities with joint realm*

$$\mathcal{X} := \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(r)}\} \subset \mathbb{R}^m.$$

Expectations for $\boldsymbol{X}$ are completely coherent if and only if there exists a convex combination $(w_1, \dots, w_r)$ such that

$$\forall g : \mathcal{X} \to \mathbb{R} \quad \mathrm{E}\{g(\boldsymbol{X})\} = \sum_{j=1}^{r} g(\boldsymbol{x}^{(j)}) \cdot w_j. \tag{1.2}$$

Sec. 1.6.1 gives a generalisation of the FTP to allow for non-finite realms.

Proof. The $\Leftarrow$ branch is straightforward. For $\Rightarrow$ note that $\boldsymbol{X}$ must take exactly one of the values in $\mathcal{X}$, and hence

$$1 = \sum_{j=1}^{r} \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}$$

where $\mathbb{1}_p$ is the indicator function of the first-order sentence p; see Sec. 1.2 for more details about this notation. By Normalisation and Linearity,

$$1 = \sum_{j=1}^{r} \mathrm{E}(\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}). \tag{1.3}$$

By Lower-boundedness, $\mathrm{E}(\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}) \geq 0$. Hence we can write $w_j \leftarrow \mathrm{E}(\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}})$, and $(w_1, \dots, w_r)$ is a convex combination. For arbitrary function g,

$$\begin{aligned}
\mathrm{E}\{g(\boldsymbol{X})\} &= \mathrm{E}\{g(\boldsymbol{X}) \cdot 1\} \\
&= \mathrm{E}\left\{ g(\boldsymbol{X}) \cdot \sum\nolimits_j \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} && \text{from above} \\
&= \mathrm{E}\left\{ \sum\nolimits_j g(\boldsymbol{X}) \cdot \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} \\
&= \mathrm{E}\left\{ \sum\nolimits_j g(\boldsymbol{x}^{(j)}) \cdot \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} && \text{good move!} \\
&= \sum\nolimits_j g(\boldsymbol{x}^{(j)}) \cdot \mathrm{E}(\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}) && \text{by (LIN)} \\
&= \sum\nolimits_j g(\boldsymbol{x}^{(j)}) \cdot w_j && \text{from above}
\end{aligned}$$

as required. □

Thus the FTP asserts that there is a bijection between the set of completely coherent expectations for $\boldsymbol{X}$ and the $(r-1)$-simplex $\mathbb{S}^{r-1}$, where $r := |\mathcal{X}|$. Because $\mathbb{S}^{r-1}$ is uncountably infinite, being a convex subset of $\mathbb{R}^r$, the set of completely coherent expectations for $\boldsymbol{X}$ is uncountably infinite too.

From now on I will always assume that expectations are completely coherent.

1.1.3 Moments

There are both practical and theoretical reasons for summarising beliefs about X in terms of its 'moments'. There are three types:

$$\begin{aligned} \text{'raw' moments} &:= \mathrm{E}(X^k) \\ \text{centered moments} &:= \mathrm{E}\{(X-\mu)^k\} \quad \text{where } \mu := \mathrm{E}(X) \\ \text{absolute moments} &:= \mathrm{E}(|X|^k) \end{aligned}$$

for $k = 1, 2, \ldots$. The first 'raw' moment is of course the expectation of X, and is often denoted μ, as above. Examples of the use of these moments are given in Sec. 1.2.3 and Sec. 1.6.2.

The second centered moment is termed the 'variance' of X, written 'Var(X)', and often denoted by σ^2. Its square root is termed the *standard deviation* of X, and often denoted by σ. Multiplying out shows that

$$\sigma^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2$$

from which we can infer that $\mathrm{E}(X^2) \geq \mathrm{E}(X)^2$. This is just (SIQ) with $Y \leftarrow 1$. The variance is a crucial concept because of its role in Chebyshev's inequality[8] and the Weak Law of Large Numbers, and the Central Limit Theorem.

The third and fourth centred moments are used to measure 'skewness' and 'kurtosis', but these concepts are not as popular as they used to be. For most people, it is a stretch to have quantitative beliefs about the skewness or kurtosis of X, unlike the expectation or the standard deviation.

Jensen's inequality (JEN) gives a rich set of inequalities for the moments to satisfy. For if $k \geq 1$ then $|x|^k$ is a convex function, and therefore

$$\mathrm{E}(|X|^s) = \mathrm{E}\{(|X|^r)^{s/r}\} \geq \mathrm{E}(|X|^r)^{s/r} \qquad : 0 < r \leq s.$$

Taking roots gives *Lyapunov's inequality*,

$$\mathrm{E}(|X|^s)^{1/s} \geq \mathrm{E}(|X|^r)^{1/r} \qquad : 0 < r \leq s. \tag{1.4}$$

So we do not have a free hand when specifying absolute moments: complete coherence imposes some restrictions. Raw moments can be bounded by absolute moments using

$$\mathrm{E}(|X^k|) \left\{ \begin{array}{l} \geq \mathrm{E}(|X|)^k \\ \geq |\mathrm{E}(X^k)| \end{array} \right\} \geq |\mathrm{E}(X)|^k \qquad : k \geq 1, \tag{1.5}$$

known as the *triangle inequality* when $k = 1$.

1.2 Probability

When expectation is primitive, probability is defined in terms of expectation.

[8] Chebyshev's inequality is given in (1.10).

1.2.1 *Definition, the FTP again*

Let $q(\boldsymbol{x})$ be a first order sentence; i.e. a statement about $\boldsymbol{x}$ which is either false or true. Let $\mathbb{1}_p$ denote the indicator function of the first-order sentence p; i.e. the function which is 0 when p is false and 1 when p is true. Then $Q := q(\boldsymbol{X})$ is a *random proposition*; random propositions are typically denoted P, Q, and R. The *probability* of Q is defined as

$$\Pr(Q) := \mathrm{E}(\mathbb{1}_Q). \tag{PR}$$

It is straightforward to check that if $\mathrm{E}(\cdot)$ is completely coherent, then $\Pr(\cdot)$ obeys the three axioms of probability.[9] Simple direct proofs can also be provided for some of the implications of the probability axioms. For example, if $q(\boldsymbol{x})$ and $r(\boldsymbol{x})$ are first-order sentences and $q(\boldsymbol{x})$ implies $r(\boldsymbol{x})$ for all $\boldsymbol{x}$, then $\mathbb{1}_Q \leq \mathbb{1}_R$, and hence $\Pr(Q) \leq \Pr(R)$.

Here is a heuristic for probability, in the same sense that 'best guess' is a heuristic for expectation. Imagine being offered a bet on Q, which pays £0 if Q is false, and £1 if Q is true. Then because

$$\Pr(Q) = 0 \cdot \Pr(\neg Q) + 1 \cdot \Pr(Q),$$

I can think of $\Pr(Q)$ as my 'fair price' for the bet. So this is one simple way to access beliefs about $\Pr(Q)$, I ask "What is the maximum I would be prepared to pay for such a bet?" This satisfies the obvious endpoints that if I thought Q was impossible, I would pay nothing, and if I thought Q was certain, I would pay up to £1. So the heuristic is really about a way to envisage probabilities of propositions that are neither impossible or certain.

Now we can have another look at the FTP from Thm 1.1. Let $\boldsymbol{x}^{(k)}$ be an element of $\mathcal{X}$, and define

$$q(\boldsymbol{X}) := \bigwedge_{i=1}^{m} (X_i \doteq x_i^{(k)})$$

or, in a more efficient notation, $q(\boldsymbol{X}) := (\boldsymbol{X} \doteq \boldsymbol{x}^{(k)})$.[10] Then, setting $g(\boldsymbol{X}) \leftarrow \mathbb{1}_{q(\boldsymbol{X})}$ in (1.2) shows that

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(k)}) = w_k.$$

Define the function

$$\mathrm{p}_{\boldsymbol{X}}(\boldsymbol{x}) := \Pr(\boldsymbol{X} \doteq \boldsymbol{x}), \tag{1.6}$$

known as the *probability mass function (PMF)* of $\boldsymbol{X}$. By convention, the PMF of $\boldsymbol{X}$ is defined for the whole of $\mathbb{R}^m$, and set to zero for values not in $\mathcal{X}$; the *support* of the PMF is the set

$$\operatorname{supp} \mathrm{p}_{\boldsymbol{X}} := \left\{ \boldsymbol{x} \in \mathbb{R}^m : \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{x}) > 0 \right\}, \tag{1.7}$$

which is a subset of $\mathcal{X}$. The FTP in (1.2) can now be written as

$$\forall g : \mathcal{X} \to \mathbb{R} \quad \mathrm{E}\{g(\boldsymbol{X})\} = \sum_{\boldsymbol{x} \in \mathcal{X}} g(\boldsymbol{x}) \cdot \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{FTP}$$

[9] At least, for finite disjunctions, since I have not used the stronger axiom of countable additivity; see Sec. 1.6.1.

[10] I use dots to indicate binary predicates in infix notation, so that $X_i \doteq x_i$ is the random proposition which is true when X_i is equal to x_i, and false otherwise.

or as

$$\forall g : \mathcal{X} \to \mathbb{R} \quad \mathrm{E}\{g(\boldsymbol{X})\} = \int_{\mathbb{R}^m} g(\boldsymbol{x}) \cdot \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{x}),$$

for an appropriate definition of the integral operator.

Eq. (FTP) is a theorem when expectation is taken as primitive. Probabilists, though, axiomatise $\mathrm{p}_{\boldsymbol{X}}$ and then (FTP) is the definition of expectation. My view[11] is that the probabilists' approach is back-to-front for Statistics, where we concern ourselves with our beliefs about $\boldsymbol{X}$ directly.

In notation, usual practice is to suppress the '$\boldsymbol{X}$' subscript on '$\mathrm{p}_{\boldsymbol{X}}$', on the grounds that the random quantities can be inferred from the argument to the function. I will follow this practice except where there might be ambiguity.

1.2.2 *Marginalisation*

Regardless of what is taken as primitive, the starting-point in Statistics is often a PMF for $\boldsymbol{X}$, or perhaps a family of PMFs for $\boldsymbol{X}$ (see Chapter 2). In this case it is important to know how to derive the PMF of any set of functions of $\boldsymbol{X}$.

Let $g_1, \dots, g_n$ be specified functions of $\boldsymbol{x}$, and set $Y_i := g_i(\boldsymbol{X})$ for $i = 1, \dots, n$. Then it follows from (FTP) that

$$\mathrm{p}(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{i=1}^{n} \mathbb{1}_{g_i(\boldsymbol{x}) \doteq y_i} \cdot \mathrm{p}(\boldsymbol{x}). \tag{1.8}$$

This expression uses the identity $\mathbb{1}_{A \wedge B} = \mathbb{1}_A \cdot \mathbb{1}_B$. In the case where $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)$, setting $\boldsymbol{Y} \leftarrow \boldsymbol{X}_A$ in (1.8) shows that

$$\mathrm{p}(\boldsymbol{x}_A) = \sum_{\boldsymbol{x}_B \in \mathcal{X}_B} \mathrm{p}(\boldsymbol{x}_A, \boldsymbol{x}_B). \tag{MAR}$$

This is termed *marginalising out* $\boldsymbol{X}_B$, and (MAR) is the *Marginalisation Theorem.*

In general, computing $\mathrm{p}(\boldsymbol{y})$ from $\mathrm{p}(\boldsymbol{x})$ or marginalising out $\boldsymbol{X}_B$ are both computationally expensive when $\mathcal{X}$ or $\mathcal{X}_B$ are large. One exception is when $\boldsymbol{X}$ has a *Multinormal distribution* and $\boldsymbol{Y}$ is a linear function of $\boldsymbol{X}$; see Mardia *et al.* (1979, ch. 3). Another exception for marginalisation is when

$$\mathrm{p}(\boldsymbol{x}) = \prod_{i=1}^{m} \mathrm{p}_i(x_i),$$

where often p_i is the same for all i (see Sec. 1.5). Unsurprisingly, these are both very common choices in practice. It is important to appreciate that the recurring use of these choices does not indicate a statistical regularity in our world, but the preference of statisticians for tractable computations.

Some notation. (MAR) is an example of a *functional equality*. My convention is that this expression denotes a set of equalities, one for each element in the product of the realms of the free arguments. In this case, the only free argument is $\boldsymbol{x}_A$, and so this equality holds

[11] Not mine alone! See, for example, de Finetti (1974/75), Lad (1996), Whittle (2000), and Goldstein and Wooff (2007).

for every $x_A \in \mathcal{X}_A$. Where it is necessary to restrict the domain of a free argument, the restriction will be given after a ':'. Some examples have already been given, another one is immediately below in (1.9).

1.2.3 Probabilities and expectations

A very famous and useful inequality links probabilities and expectations, *Markov's inequality*:

$$\Pr(|X| \dot{\geq} a) \leq \frac{\mathrm{E}(|X|)}{a} \qquad : a > 0. \tag{1.9}$$

This follows immediately from $a \cdot \mathbb{1}_{|X| \dot{\geq} a} \leq |X|$ and Monotonicity.

Markov's inequality is versatile, because if g is a non-negative increasing function, then

$$g(|x|) \geq g(a) \iff |x| \geq a.$$

One application of this is the *centered moment bound*,

$$\Pr(|X - \mu| \dot{\geq} a) \leq \min_{k \geq 0} \frac{\mathrm{E}(|X - \mu|^k)}{a^k} \qquad : a > 0, \tag{1.10}$$

where $\mu := \mathrm{E}(X)$. This bound shows how the absolute centered moments of X control the behaviour of the tails of the PMF of X. The special case of $k \leftarrow 2$ is termed *Chebyshev's inequality*, for which the righthand side of (1.10) is σ^2/a^2, where $\sigma^2 := \mathrm{Var}(X)$.

1.3 'Hypothetical' expectations

The material in this section is radical. I want to adjust Your viewpoint before we go any further.

1.3.1 Some reflections

> There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to think fruitfully about an idea. (Andreas Buja, quoted in Hastie *et al.*, 2009, p. xii).

Statisticians are not 'just' mathematicians. In Statistics, quantities which are abstractions from a mathematical viewpoint must be reified,[12] so that they quantify aspects of the reality which we experience together. My expectation $\mathrm{E}(X)$ has meaning to me, and this meaning informs my decision to constrain all of my expectations to be completely coherent (see Sec. 1.1). I doubt very much that You and I can agree on precisely what we each mean by 'expectation', but I hope that we have enough common ground that You consider that knowing the values of some of my expectations, and knowing that they are completely coherent by construction, is useful when You consider or revise some of Your expectations.

Although we could wish for a tighter definition of 'expectation', ideally even complete agreement between You and me regarding

[12] Verb: to make something that is abstract more concrete or real. As used in the title of Goldstein and Rougier (2009).

its meaning, nothing I have experienced in my interactions with other people leads me to think that this is possible. We humans constantly misunderstand each other. So my beliefs are mine alone, not just in the value I might attach to an expectation, but even in what I mean by 'expectation'. I don't think there is any point in constructing an elaborate theory about this, such as "my expectation of X is the value of a I would choose were I facing a penalty of $(X - a)^2$." This is a *deus ex machina*, designed to crush ambiguity, but at the expense of our humanity.

I think it is better to acknowledge from the outset basic limits to our mutual understanding. The viewpoint I want to advocate in these notes is that these limits do not imply that 'anything goes' when it comes to quantifying beliefs. You might find my beliefs useful, and You might find them more useful if they are completely coherent. You should distrust anyone who claims to have quantified 'the' expectation for X. If You are asked for 'the' expectation, You can reply, "I am happy to give you my expectation, and I hope you find it useful in quantifying yours."

This section considers the next stage of this process, what I term 'hypothetical expectations', although typically these would be termed 'conditional expectations' (see Sec. 1.3.3). Mathematicians are not obliged to attach any meaning to 'the conditional expectation of X given that Q is true'. In elementary textbooks it is defined (perhaps implicitly) as a quotient of expectations:

$$\mathrm{E}(X \mid Q) := \frac{\mathrm{E}(X \mathbb{1}_Q)}{\Pr(Q)} \quad \text{provided that } \Pr(Q) > 0.$$

Based on this definition, we can prove lots of Cool Stuff about hypothetical expectations, including relationships between hypothetical expectations with different Q's. But statisticians have to go much further. For a statistician, $\mathrm{E}(X \mid Q)$ has to have enough meaning that it could be assigned a value. For the Cool Stuff to be useful, this meaning has to be such as to make the above relation true. This is the challenge I address in Sec. 1.3.2. As far as I know, no one else has reified hypothetical expectation in the way that I do. I do not think that Sec. 1.3.2 is the last word on the meaning of hypothetical expectation. But I hope that You understand the need for what I have tried to do.

1.3.2 Definition of hypothetical expectation

Let Q be a random proposition, which may or may not be true. People are adept at thinking hypothetically, "supposing Q to be true". I can have a 'best guess' about X supposing Q to be true: this is my *hypothetical expectation* denoted as $\mathrm{E}(X \mid Q)$, and usually expressed as "my expectation of X given Q". The challenge is to give this notion enough substance that we can propose sensible properties that hypothetical expectations should possess. Here is an informal definition.

Some notation. A *partition* is a collection of mutually exclusive and exhaustive random propositions. If

$$\mathcal{Q} := \{Q^{(1)}, \dots, Q^{(k)}\}$$

is a partition, then $\Pr(Q^{(i)} \wedge Q^{(j)}) = 0$ for $i \neq j$, and $\Pr(Q^{(1)} \vee \dots \vee Q^{(k)}) = 1$.

Definition 2 (Hypothetical expectation, informal)**.** *Let $\mathcal{Q}$ be a partition. I imagine myself in the closest world in which the value of $\mathcal{Q}$ is known. The hypothetical expectation $\mathrm{E}(X \mid Q^{(j)})$ is my belief about X when $Q^{(j)}$ is true in this world.*

You can see that this is a very subtle concept—but what did You expect? The truth of $Q^{(j)}$ holds in an infinite number of imaginary worlds, and something has to be done to reduce the ambiguity. So this informal device of the 'closest world' is an attempt to mimic what we do in practice. When reasoning hypothetically, we do not consider strange new worlds in which $Q^{(j)}$ is true, but worlds that are similar to our own. Technically, the partition $\mathcal{Q}$ which defines the 'closest world' ought to be recorded along with the element $Q^{(j)}$ in the notation for hypothetical expectation, but I have suppressed it to avoid clutter.

Following (PR), I define a hypothetical probability for a random proposition as

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

It is conventional to call this a *conditional probability*, which I will do, although I could also call it a 'hypothetical probability'.

What can we say about a hypothetical expectation? And does it need to have any connection at all to 'actual' expectation? I provide a condition for each of these questions, and show how they are equivalent to a condition which directly expresses a hypothetical expectation in terms of actual expectations.

Let X be any random quantity and $\mathcal{Q}$ be any partition. The first condition is that

$$\mathrm{E}(X\mathbb{1}_{Q^{(i)}} \mid Q^{(j)}) = \begin{cases} \delta_{ij}\,\mathrm{E}(X \mid Q^{(j)}) & \Pr(Q^{(j)}) > 0 \\ \text{arbitrary} & \Pr(Q^{(j)}) = 0 \end{cases} \tag{1.11}$$

where δ_{ij} is the Kronecker delta function.[13] That is, if I am supposing $Q^{(j)}$ to be true, then I must believe that $Q^{(i)}$ is false for $i \neq j$. It is hard to disagree with this, so I call this the *sanity condition* for hypothetical expectations. Note that I make no claims at all for hypothetical expectations in what I believe to be impossible situations.

The second condition links hypothetical expectations and actual expectations. Bruno de Finetti (1972, sec. 9.5) termed it the *conglomerative property*:

$$\mathrm{E}(X) = \sum_{j=1}^{k} \mathrm{E}(X \mid Q^{(j)}) \Pr(Q^{(j)}). \tag{1.12}$$

[13] I.e. the function which is 1 when $i = j$ and zero otherwise, which can also be written as $\mathbb{1}_{i \doteq j}$.

This is a strong condition, but it has an intuitive shape. It states that I do not have a free hand when specifying all of my hypothetical expectations, because, when taken together, they must be consistent with my actual expectation. In fact, the conglomerative property represents a two-stage approach for specifying my beliefs about X. First, I think about X hypothetically, over each element of a partition, and then I combine these values according to the probability I attach to each element in the partition. Lindley (1985, sec. 3.8) termed this approach to specifying beliefs about X 'extending the conversion'.

What is interesting is that these two conditions are sufficient to define hypothetical expectation, according to the following result.

Theorem 1.2 (Hypothetical Expectations Theorem, HET)**.** *Hypothetical expectations satisfy the sanity condition and the conglomerative property if and only if they satisfy the relation*

$$\mathrm{E}(X\mathbb{1}_Q) = \mathrm{E}(X \mid Q)\,\Pr(Q) \tag{1.13}$$

for every random quantity X and every random proposition Q.

As a consequence of this result, (1.13) will be taken as the defining property of a hypothetical expectation.

Proof. Let X be a random quantity and Q be a random proposition. Where necessary, embed Q in some partition $\mathcal{Q}$.
$\Leftarrow$. Note that $\Pr(Q) = 0$ implies that $\mathrm{E}(X\mathbb{1}_Q) = 0$, by (SIQ). Then it is straightforward to check that (1.13) implies the sanity condition, substituting $X \leftarrow X\mathbb{1}_{Q^{(i)}}$ and $Q \leftarrow Q^{(j)}$. For the conglomerative property,

$$\begin{aligned}
\mathrm{E}(X) &= \mathrm{E}\left(X \cdot \textstyle\sum_j \mathbb{1}_{Q^{(j)}}\right) && \text{as } \mathcal{Q} \text{ is a partition}\\
&= \textstyle\sum_j \mathrm{E}(X\mathbb{1}_{Q^{(j)}}) && \text{by linearity}\\
&= \textstyle\sum_j \mathrm{E}(X \mid Q^{(j)})\,\Pr(Q^{(j)}) && \text{by (1.13)}
\end{aligned}$$

as required.
$\Rightarrow$.

$$\begin{aligned}
\mathrm{E}(X\mathbb{1}_{Q^{(i)}}) &= \textstyle\sum_j \mathrm{E}(X\mathbb{1}_{Q^{(i)}} \mid Q^{(j)})\,\Pr(Q^{(j)}) && \text{(conglomerative property)}\\
&= \textstyle\sum_j \delta_{ij}\,\mathrm{E}(X \mid Q^{(j)})\,\Pr(Q^{(j)}) && \text{by the sanity condition, (1.11)}\\
&= \mathrm{E}(X \mid Q^{(i)})\,\Pr(Q^{(i)})
\end{aligned}$$

as required. □

Eq. (1.13) is a good starting-point for several other useful results. Putting $X \leftarrow \mathbb{1}_P$ in (1.13) shows that the conditional probability always satisfies

$$\Pr(P, Q) = \Pr(P \mid Q)\,\Pr(Q), \tag{1.14}$$

using the common notation that $\Pr(P, Q) := \Pr(P \land Q)$. This is a result of great practical importance. It provides a two-stage approach for specifying the probability of any conjunction: first think about $\Pr(Q)$, and then about the conditional probability $\Pr(P \mid Q)$, i.e. "the probability that P is true supposing that Q is true". Note from (1.14) that $\Pr(P \mid Q)$ has the unique value

$$\Pr(P \mid Q) = \frac{\Pr(P, Q)}{\Pr(Q)} \tag{1.15}$$

when $\Pr(Q) > 0$, but is arbitrary when $\Pr(Q) = 0$.

Another useful result is that if $\mathrm{E}(\cdot)$ is completely coherent, then $\mathrm{E}(\cdot \mid Q)$ is completely coherent whenever $\Pr(Q) > 0$; this follows from the *conditional FTP*,

$$\forall g : \mathcal{X} \to \mathbb{R} \quad \mathrm{E}\{g(\boldsymbol{X}) \mid Q\} = \sum_{\boldsymbol{x} \in \mathcal{X}} g(\boldsymbol{x}) \cdot \mathrm{p}_Q(\boldsymbol{x}) \qquad : \ \Pr(Q) > 0 \tag{1.16a}$$

where

$$\mathrm{p}_Q(\boldsymbol{x}) := \Pr(\boldsymbol{X} \doteq \boldsymbol{x} \mid Q) = \frac{\mathbb{1}_{q(\boldsymbol{x})}\,\mathrm{p}(\boldsymbol{x})}{\Pr(Q)}. \tag{1.16b}$$

This result is straightforward to prove, starting from the FTP for $\mathrm{E}\{g(\boldsymbol{X})\mathbb{1}_Q\}$ and then using (1.13). I refer to (1.16b) as the *Muddy Table Theorem*, following van Fraassen (1989, ch. 7).

Eq. (1.16) and the FTP show that complete coherence implies that hypothetical expectations have a *recursive property*: every result about expectations $\mathrm{E}(\cdot)$ also holds for expectations $\mathrm{E}(\cdot \mid Q)$ if $\Pr(Q) > 0$; and every result about $\mathrm{E}(\cdot \mid Q)$ also holds for $\mathrm{E}(\cdot \mid Q, R)$ if $\Pr(Q, R) > 0$; and so on. In other words, we can drop a '$|Q$' into the back of all expectations, or a '$, R$' into the back of all hypothetical expectations, and whatever result we are interested in still holds, provided that $\Pr(Q) > 0$ or $\Pr(Q, R) > 0$; and so on.

1.3.3 'Conditional' expectations

I have been careful to write 'hypothetical' and not 'conditional' expectation for $\mathrm{E}(X \mid Q)$. This is because probability theory makes a clear distinction between the two, which is honoured in notation, but often overlooked. The hypothetical expectation $\mathrm{E}(X \mid Q)$ is a value, just like $\mathrm{E}(X)$ is a value. But the conditional expectation is a *random quantity*, not a value.

Consider two random quantities, X and Y, where the following construction generalises immediately to the case where Y is a vector of random quantities. Now

$$\mathcal{Q} := \bigcup_{y \in \mathcal{Y}} (Y \doteq y)$$

is a partition, so we will go ahead and define the function

$$\mu_X(y) := \mathrm{E}(X \mid Y \doteq y) \qquad y \in \mathcal{Y}. \tag{1.17}$$

This definition is *not* unique, because $\mathrm{E}(X \mid Y \doteq y)$ is arbitrary if $\Pr(Y \doteq y) = 0$. In general, there are an uncountable number of μ_X

functions; denote these as $\mu'_X, \mu''_X, \dots$. For each one of these, define the corresponding *conditional expectation* of X given Y,

$$\begin{aligned} \mathbb{E}'(X \mid Y) &:= \mu'_X(Y) \\ \mathbb{E}''(X \mid Y) &:= \mu''_X(Y) \\ &\vdots \end{aligned} \tag{1.18}$$

Each of these is a random quantity, being a specified function of Y, termed a *version* of the conditional expectation. But although these are different random quantities, it is straightforward to show using the FTP that they are *mean-squared equivalent*, i.e.

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

more conveniently written as $\mathbb{E}'(X \mid Y) \overset{\text{ms}}{=} \mathbb{E}''(X \mid Y)$. Therefore it is common to refer to 'the' conditional expectation $\mathbb{E}(X \mid Y)$. But, just to make the point one more time, $\mathbb{E}(X \mid Y)$ is a function of the random quantity Y, *it is not a value*.

In my notation I do not need to use two different symbols E and $\mathbb{E}$ for hypothetical expectation and conditional expectation, because the symbol to the right of the bar is clearly either a random proposition, like Q, or a random quantity, like Y. Most authors do not make a notational distinction. But I am insisting, because the difference is so fundamental, and also because it clarifies some important equalities involving hypothetical and conditional expectations.

The first one is the conglomerative property (1.12), which in this context is termed the *Tower Property* of conditional expectation:

$$\mathrm{E}(X) = \mathrm{E}\{\mathbb{E}(X \mid Y)\}, \tag{1.19}$$

also termed the Law of Iterated Expectation, see (LIE) below in Sec. 1.4. This equality holds for every version of $\mathbb{E}(X \mid Y)$. It can be developed recursively, just like a hypothetical expectation. So we could have, for example,

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

$\mathbb{E}$ behaves like an expectation, i.e. it respects the axioms of lower-boundedness and additivity, but, again, only in mean square.

The Tower Property has an elegant and useful extension, for computing variances (see Sec. 1.1.3), the *variance identity*:

$$\mathrm{Var}(X) = \mathrm{E}\{\mathbb{V}\mathrm{ar}(X \mid Y)\} + \mathrm{Var}\{\mathbb{E}(X \mid Y)\}, \tag{1.20}$$

where $\mathbb{V}$ar denotes the conditional variance,

$$\begin{aligned} \mathbb{V}\mathrm{ar}(X \mid Y) &:= \mathbb{E}[\{X - \mathbb{E}(X \mid Y)\}^2 \mid Y] \\ &= \mathbb{E}(X^2 \mid Y) - \mathbb{E}(X \mid Y)^2. \end{aligned}$$

So, like the conditional expecation, the conditional variance is a random quantity. Eq. (1.20) is straightforward to derive, using (1.19) and the definition of $\mathbb{V}$ar immediately above.

The Tower Property and the variance identity are useful because in some applications it is possible to derive a closed-form expression for $\mu_X(y)$ and for $\sigma_X^2(y)$, the hypothetical variance conditional on $Y \doteq y$. Then we have simple recipes for computing the expectation and variance of X. Note that although μ_X and σ_X^2 are not unique there is usually a 'vanilla' form. For example, the Multinormal distribution has an uncountable realm (see Sec. 1.6.3), and hence $\Pr(Y \doteq y) = 0$ for all $y \in \mathcal{Y}$. Nevertheless, it is possible to state useful expressions for μ_X and σ_X^2.

The general theory of conditional expectation was originally proposed by the great Soviet mathematician Andrey Kolmogorov, notably in his book *The Foundations of Probability*, published in 1933. Measure Theory is indispensible: see Billingsley (1979) or Williams (1991) for the details. Another view of conditional expectation is that it represents a projection; see Whittle (2000) for details.

1.4 Implications of the HET

Now we are back on-track! Regardless of where we start, (1.13) is the defining relationship for hypothetical expectations and conditional probabilities, from which the following results follow immediately.

The conglomerative property given in (1.12) is also known as the *Law of Iterated Expectation (LIE)*, and its special case for probabilities is known as the *Law of Total Probability (LTP)*

$$\mathrm{E}(X) = \sum_{Q \in \mathcal{Q}} \mathrm{E}(X \mid Q) \Pr(Q), \quad \Pr(P) = \sum_{Q \in \mathcal{Q}} \Pr(P \mid Q) \Pr(Q)$$

whenever $\mathcal{Q}$ is a partition. See below (LIE, LTP) for common expressions for these in terms of PMFs.

Here is a very useful result which I call *Taking out What is Known (TWK)*, after Williams (1991, sec. 9.7):

$$\begin{aligned} &\mathrm{E}\{g(\boldsymbol{Y}) \cdot h(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{Y} \doteq \boldsymbol{y}\} \\ &\qquad = g(\boldsymbol{y}) \cdot \mathrm{E}\{h(\boldsymbol{X}, \boldsymbol{y}) \mid \boldsymbol{Y} \doteq \boldsymbol{y}\} \qquad : \boldsymbol{y} \in \operatorname{supp} \boldsymbol{Y}; \qquad \text{(TWK)} \end{aligned}$$

recollect the definition of 'supp', the support of a PMF, given in (1.7). Conceptually, this is just an extension of the sanity condition, (1.11), since it would be weird if $\boldsymbol{Y}$ was not equal to $\boldsymbol{y}$ in the hypothetical world where $\boldsymbol{Y} \doteq \boldsymbol{y}$ was true. Eq. (TWK) can be proved using the FTP for $\mathrm{E}\{g(\boldsymbol{Y}) \cdot h(\boldsymbol{X}, \boldsymbol{Y}) \cdot \mathbb{1}_{\boldsymbol{Y} \doteq \boldsymbol{y}}\}$ and (1.13). It also holds in mean square for conditional expecations.[14]

Here are three other very important results relating probability and conditional probability, for random propositions P, Q, and R:

1. *Factorisation Theorem*, which just extends (1.14).

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

2. *Sequential Conditioning*

$$\Pr(P, Q \mid R) = \Pr(P \mid Q, R) \Pr(Q \mid R) \qquad : \Pr(R) > 0.$$

[14] For example, $\mathbb{E}(XY \mid Y) \overset{\text{ms}}{=} Y \, \mathbb{E}(X \mid Y)$.

3. *Bayes's Theorem*[15]

$$\Pr(P \mid Q) = \frac{\Pr(Q \mid P)\Pr(P)}{\Pr(Q)} \qquad : \Pr(Q) > 0.$$

Bayes's theorem also has an odds form[16]

$$\frac{\Pr(P \mid Q)}{\Pr(R \mid Q)} = \frac{\Pr(Q \mid P)}{\Pr(Q \mid R)}\,\frac{\Pr(P)}{\Pr(R)} \qquad : \Pr(Q, R) > 0.$$

This is convenient because it cancels $\Pr(Q)$. One common special case is $R \leftarrow \neg P$, where $\neg P$ denotes 'not P'.

Each of these results can be expressed in terms of PMFs, which is how statisticians usually encounter them in practice. For simplicity, I write 'supp $\boldsymbol{X}$' to denote 'supp $\mathrm{p}_{\boldsymbol{X}}$' where $\mathrm{p}_{\boldsymbol{X}}$ is the marginal PMF of $\boldsymbol{X}$, see (MAR).

0. *Law of Iterated Expectation, Law of Total Probability*

$$\mathrm{E}(X) = \sum_{\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}} \mathrm{E}(X \mid \boldsymbol{Y} \doteq \boldsymbol{y}) \cdot \mathrm{p}(\boldsymbol{y}) \tag{LIE}$$

$$\mathrm{p}(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}} \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}) \cdot \mathrm{p}(\boldsymbol{y}), \tag{LTP}$$

because $\bigcup_{\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}} (\boldsymbol{Y} \doteq \boldsymbol{y})$ is a partition.

1. *Factorisation Theorem*

$$\begin{aligned} \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) &= \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y})\,\mathrm{p}(\boldsymbol{y}) \\ \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z})\,\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{z})\,\mathrm{p}(\boldsymbol{z}), \end{aligned} \tag{FAC}$$

and so on.

2. *Sequential Conditioning*

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z})\,\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{z}) \qquad : \boldsymbol{z} \in \operatorname{supp} \boldsymbol{Z}. \tag{SEQ}$$

3. *Bayes's Theorem*

$$\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})\,\mathrm{p}(\boldsymbol{x})}{\mathrm{p}(\boldsymbol{y})} \qquad : \boldsymbol{y} \in \operatorname{supp} \boldsymbol{Y}. \tag{BAY}$$

And in odds form

$$\frac{\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y})}{\mathrm{p}(\boldsymbol{x}' \mid \boldsymbol{y})} = \frac{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})}{\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}')}\,\frac{\mathrm{p}(\boldsymbol{x})}{\mathrm{p}(\boldsymbol{x}')} \qquad : (\boldsymbol{x}', \boldsymbol{y}) \in \operatorname{supp}(\boldsymbol{X}, \boldsymbol{Y}). \tag{BOD}$$

1.5 Conditional independence

Conditional independence is the cornerstone of statistical modelling: it is the most important thing after expectation itself. Conditional independence is a property of beliefs.

Definition 3 (Conditional independence)**.**
Let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ be three collections of random quantities. My beliefs about $\boldsymbol{X}$ are conditionally independent of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ exactly when

$$\forall g : \boldsymbol{\mathcal{X}} \to \mathbb{R} \quad \mathrm{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Y} \doteq \boldsymbol{y}, \boldsymbol{Z} \doteq \boldsymbol{z}\} = \mathrm{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Z} \doteq \boldsymbol{z}\} \qquad : (\boldsymbol{y}, \boldsymbol{z}) \in \operatorname{supp}(\boldsymbol{Y}, \boldsymbol{Z}).$$

This is written $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$.

[15] I insist on "Bayes's", on the authority of Fowler's Modern English Usage, 2rd edn, p. 466. Americans do this differently.

[16] 'Odds' denotes a ratio of probabilities.

That is to say, whenever I imagine the closest world in which the values of both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are known, I find that my hypothetical beliefs about $\boldsymbol{X}$ do not depend on the value taken by $\boldsymbol{Y}$, and are the same as if $\boldsymbol{Y}$ was not known.

The definition in Def. 3 gives meaning to the notion of conditional independence as a property of beliefs, but it is unwieldy to use in practice. Happily we have the following result.

Theorem 1.3 (Equivalents to conditional independence)**.**
The following statements are equivalent:

(i) $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$

(ii) $\mathrm{p}(x \mid y, z) = \mathrm{p}(x \mid z) \qquad : (y, z) \in \operatorname{supp}(\boldsymbol{Y}, \boldsymbol{Z})$

(iii) $\mathrm{p}(x, y \mid z) = \mathrm{p}(x \mid z) \cdot \mathrm{p}(y \mid z) \qquad : z \in \operatorname{supp} \boldsymbol{Z}$

(iv) $\mathrm{E}\{g(\boldsymbol{X}) \cdot h(\boldsymbol{Y}) \mid \boldsymbol{Z} \doteq z\} = \mathrm{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Z} \doteq z\} \cdot \mathrm{E}\{h(\boldsymbol{Y}) \mid \boldsymbol{Z} \doteq z\} \qquad : z \in \operatorname{supp} \boldsymbol{Z}.$

Proof.
(i) implies (ii) after setting $g(x') \leftarrow \mathbb{1}_{x' \doteq x}$.
(ii) implies (iii). Eq. (SEQ) asserts that

$$\mathrm{p}(x, y \mid z) = \mathrm{p}(x \mid y, z) \cdot \mathrm{p}(y \mid z) \qquad : z \in \operatorname{supp} \boldsymbol{Z}. \qquad (\dagger)$$

Consider the two cases. First, $y \in \operatorname{supp}(\boldsymbol{Y} \mid \boldsymbol{Z} \doteq z)$, so that $(y, z) \in \operatorname{supp}(\boldsymbol{Y}, \boldsymbol{Z})$. In this case (ii) and (†) imply (iii). Second, $y \notin \operatorname{supp}(\boldsymbol{Y} \mid \boldsymbol{Z} \doteq z)$. In this case (†) has the form $0 = \mathrm{p}(x \mid y, z) \cdot 0$, and we may take $\mathrm{p}(x \mid y, z) \leftarrow \mathrm{p}(x \mid z)$, as required.
(iii) implies (i):

$$\begin{aligned} &\mathrm{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Y} \doteq y, \boldsymbol{Z} \doteq z\} \\ &\qquad = \textstyle\sum_x g(x) \cdot \mathrm{p}(x \mid y, z) && \text{from the CFTP, (1.16)} \\ &\qquad = \textstyle\sum_x g(x) \cdot \dfrac{\mathrm{p}(x, y \mid z)}{\mathrm{p}(y \mid z)} && (\dagger) \text{ and } (y, z) \in \operatorname{supp}(\boldsymbol{Y}, \boldsymbol{Z}) \\ &\qquad = \textstyle\sum_x g(x) \cdot \mathrm{p}(x \mid z) && \text{from (iii)} \\ &\qquad = \mathrm{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Z} \doteq z\} && \text{CFTP again.} \end{aligned}$$

(iii) implies (iv) using the CFTP. (iv) implies (iii) after setting $g(x') \leftarrow \mathbb{1}_{x' \doteq x}$ and $h(y') \leftarrow \mathbb{1}_{y' \doteq y}$. □

The definition of conditional independence can be simplified to that of *independence*, simply by dropping $\boldsymbol{Z}$. So my beliefs about $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent exactly when

$$\forall g : \mathcal{X} \to \mathbb{R} \quad \mathrm{E}\{g(\boldsymbol{X}) \mid \boldsymbol{Y} \doteq y\} = \mathrm{E}\{g(\boldsymbol{X})\} \qquad : y \in \operatorname{supp} \boldsymbol{Y}, \quad (1.21)$$

and this is written $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$. There are straightforward modifications to the equivalent conditions given in Thm 1.3.

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

We can also clarify that $X \perp\!\!\!\perp Y \not\Longrightarrow X \perp\!\!\!\perp Y \mid Z$. Suppose that X and Y are the points from two rolls of a die believed by me to be fair. In this case, I might reasonably believe that $X \perp\!\!\!\perp Y$, if I had shaken the die extensively inside a cup before each roll. But if Z is the sum of the points in the two rolls, then I can predict X exactly knowing Y and Z, but only approximately using Z alone. So Y brings information about X that augments the information in Z, and I do not believe that $X \perp\!\!\!\perp Y \mid Z$.

These two illustrations show that conditional independence is its own thing, not simply a necessary or sufficient condition for independence. My belief that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ is something I accept or reject after reflecting on how my beliefs about $\boldsymbol{X}$ in the presence of $\boldsymbol{Z}$ change on the further presence of $\boldsymbol{Y}$. The asymmetry of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is an illusion—a fascinating and deep result, which follows immediately from the symmetry of $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z})$ in (iii) of Thm 1.3. The relationship between conditional independence (symmetric) and causality (asymmetric) is very subtle; see Pearl (2000) and Dawid (2002, 2010) for discussions.

Finally, here are some additional useful concepts based on conditional independence. A collection $\boldsymbol{X}$ is *mutually conditionally independent* given $\boldsymbol{Z}$ exactly when

$$\forall A, B \quad \boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{Z} \tag{1.22}$$

where $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ are non-intersecting subsets of $\boldsymbol{X}$. I write this as $\models \boldsymbol{X} \mid \boldsymbol{Z}$. It is straightforward to show that

$$\models \boldsymbol{X} \mid \boldsymbol{Z} \iff \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{i=1}^{m} \mathrm{p}_i(x_i \mid \boldsymbol{z}), \tag{MCI}$$

using Thm 1.3. Likewise, $\boldsymbol{X}$ is *mutually independent* exactly when $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B$ for all non-intersecting $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$, written as $\models \boldsymbol{X}$, and for which

$$\models \boldsymbol{X} \iff \mathrm{p}(\boldsymbol{x}) = \prod_{i=1}^{m} \mathrm{p}_i(x_i). \tag{MI}$$

A stronger condition for mutual [conditional] independence is where p_i is the same for all i. In this case, $\boldsymbol{X}$ is [conditionally] *independent and identically distributed (IID)* [given $\boldsymbol{Z}$]. The [conditionally] IID model is the unflagging workhorse of modern applied statistics.

1.6 *Non-finite realms*

For our convenience, it will often be useful to treat the realm of a random quantity X as non-finite, or even uncountable. These are abstractions, because the realm of an operationally-defined quantity is always finite. But remember the PEP in footnote 2: we have to make sure that we do not introduce any pathologies.

Some terms. A *finite* set has a finite number of elements; otherwise it is *non-finite*. The size of a set is termed its *cardinality*, and denoted $|A|$. A finite set in a Euclidean space has a finite diameter, i.e. is bounded; a non-finite set may or may not have finite diameter. A *countable* set has the same cardinality as $\mathbb{N}$, the set of positive integers; i.e. it can be represented as $A := \{a_i : i \in \mathbb{N}\}$. An *uncountable* set has a larger cardinality than $\mathbb{N}$; typically, its cardinality would be that of the continuum, which is the cardinality of the reals in the interval $[0, 1]$. Vilenkin (1995) provides a good introduction to the complexities of 'infinity'.

1.6.1 Countable realms

Suppose that the realm of X is non-finite but countable. Since the FTP is the basic result for complete coherence, we look to its proof to check for pathologies. And there we see that the 'only if' proof breaks down at (1.3), because the righthand side is no longer the sum over finite set. The axiom of additivity makes no claims for the expectation of the sum of an infinite set of random quantities. In order to retrieve the proof and eliminate the pathology, a stronger property is required, namely that of *countable additivity*:

$$\mathrm{E}(X_1 + X_2 + \cdots) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots \qquad \text{(Countable additivity)}$$

Now the 'only if' part of the proof goes through as before.

I interpret countable additivity as protection against pathologies that might otherwise arise if the FTP did not hold for random quantities with countable realms. Other statisticians, though, make a much bigger deal about the difference between different types of additivity, on foundational/philosophical grounds. The most vociferous has been Bruno de Finetti, e.g., de Finetti (1972, ch. 5) and de Finetti (1974, ch. 3); see also Kadane (2011, sec. 3.5).

1.6.2 Unbounded realms

If we start with expectation as primitive, then infinite expectations can never arise if we do not want them, even for random quantities whose realm is unbounded. However, modern practice, which starts with a PDF[17] rather than with a set of expectations, makes it all too easy create random quantities with infinite expectations without realising it. This is because modern practice starts with a convenient choice for the PDF of X, whose tractability often arises partly from the fact that its support is unbounded: the Normal distribution, the Gamma, the Poisson, and so on. If expectation is defined as an infinite sum or an integral, then it may may 'converge' to $\pm\infty$ or it may have no well-defined limit.

The three choices given above are actually fairly safe, because they have *finite moments*, see Sec. 1.1.3. Finite moments implies

[17] I will write 'PDF' for 'PMF/PDF' in this subsection.

that all functions of X that are bounded in absolute value by a polynomial will have finite expectations.[18]

But consider the Student-t distribution with one degree of freedom, known as a Cauchy distribution, which has support $\mathbb{R}$. Even moments are infinite, and odd moments are undefined. Thus if X is Cauchy, then the expectations of some polynomials of X are infinite, and of others are undefined. The Cauchy is a very poor choice for representing beliefs about an operationally-defined random quantity. Similar problems exist for all Student-t distributions.

Here is where statisticians have to pay attention to the PEP (footnote 2). If a random quantity is treated as having an unbounded realm, then it is the statistician's responsibility to make sure that all of the moments remain finite. One elegant way to do this is to construct more complicated PDFs from mixtures of 'safe' distributions, because these mixtures will have finite moments, according to the LIE. It may not be an explicit consideration, but the practice of *hierarchical modelling* is largely about creating mixtures of this type; see Lunn *et al.* (2013, ch. 10) or Gelman *et al.* (2014, ch. 5).

1.6.3 Uncountable realms

We lapse briefly into a more abstract notation. Let $\{a_\lambda : \lambda \in \Lambda\}$ be any parameterised collection of non-negative values in $[0, \infty]$, where Λ may be uncountable. We need to define what it means to sum over these values, in such as way that if the set is countable, then we retain the usual definition. To this end, define $\sum_{\lambda \in \Lambda} a_\lambda$ as the supremum of $\sum_{\lambda \in L} a_\lambda$, for all finite sets $L \subset \Lambda$. Now consider the case where $\sum_{\lambda \in \Lambda} a_\lambda = 1$, as it would be were the a_λ's probabilities on the realm Λ. In this case it is straightforward to show that only a *countable* number of the a_λ's can be non-zero. This argument is taken directly from Schechter (1997, sec. 10.40).

So, returning to more concrete notions, no matter what the realm of $\boldsymbol{X}$, finite, countable, or uncountable, at most a countable number of the elements of $\mathcal{X}$ will have non-zero probabilities. If $\mathcal{X}$ is uncountable, we can always 'thin' it to countable set, without changing our beliefs. Of course a countable set is still very large. The set of rationals in $[0, 1]$ is countable, but comprises an inconceivably minute proportion of the set of reals in $[0, 1]$, which has the cardinality of the continuum.

But this does present a new difficulty, if we proceed without first thinning $\mathcal{X}$ to a countable set. If the realm of $\boldsymbol{X}$ is uncountable and the distribution function $F(\boldsymbol{x}) := \Pr(\boldsymbol{X} \dot{\leq} \boldsymbol{x})$ is continuous, then the probability of $\boldsymbol{X}$ taking any specified value $\boldsymbol{x}$ is zero. To be clear, in a tiny ball around $\boldsymbol{x}$ there may be a countable number of elements with non-zero probability, but a single point selected arbitrarily from the continuum will always fall between the points of a countable subset of the continuum. So we cannot continue to define 'p$(\boldsymbol{x})$' as 'Pr$(\boldsymbol{X} \doteq \boldsymbol{x})$', because this would be vacuous.

$\boldsymbol{X}$ is a *continuous* random quantity (it maybe a vector) exactly

[18] However, this result *cannot* be extended to real analytic functions, except in the case when the realm of X is bounded.

when its distribution function F is continuous. It is an *absolutely continuous* random quantity exactly when F is differentiable.[19] Statisticians wanting to tap the continuum for their convenience almost always choose absolutely continuous random quantities. For an absolutely continuous $\boldsymbol{X}$, 'p' is defined to be the *probability density function (PDF)*, satisfying

$$\Pr(x \dot{<} \boldsymbol{X} \dot{\leq} x + \mathbf{d}x) = \mathrm{p}(x)\,\mathbf{d}x. \tag{1.23}$$

It is undoubtedly confusing to use the same symbol 'p' for *probability* in the case where $\boldsymbol{X}$ has a finite or countable realm, and *probability density* where $\boldsymbol{X}$ has an uncountably infinite realm, but this convention does make sense in the more general treatment of probability using Measure Theory, in which sums over $\mathcal{X}$ are treated formally as Lebesgue integrals (Billingsley, 1979; Williams, 1991).

Measure Theory is only required to handle uncountable realms, for which pathologies can and do arise.[20] But uncountable realms are 'unnatural', a view reiterated many times since Cantor's early work on non-finite sets. This is not just statistical parochialism. David Hilbert, one of the great mathematicians and an admirer of Cantor's work, stated

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

And later in the same essay,

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

For similar sentiments from eminent statisticians, see, e.g., Hacking (1965, ch. 5), Basu (1975), Berger and Wolpert (1984, sec. 3.4), or Cox (2006, sec. 1.6). All of these statisticians acknowledge the convenience of uncountable realms, but there is no *necessity* for uncountable realms. Thus Statistics would have entirely missed its mark if it could only be developed using Measure Theory. It has been a deliberate decision on my part not to use Measure Theory in these notes. Let me finish with a telling quote taken from Kadane (2011, start of ch. 4):

> Does anyone believe that the difference between the Lebesgue and Riemann integrals can have physical significance, and that whether say, an airplane would or would not fly could depend on this difference? If such were claimed, I should not care to fly on that plane. (Richard Wesley Hamming)

[19] There are also hybrid random quantities where the distribution is mostly continuous, but has vertical jumps, at what are termed 'atoms'.

[20] See, for example, the Borel paradox, discussed in Poole and Raftery (2000).

2
Modes of statistical inference

This chapter is an overview of statistical inference, from its origins in populations and (random) samples, to its modern practice. Over the last thirty years, the applications of Statistics have diversified enormously, reflecting the availability of new large datasets, more powerful computers and better algorithms, improved statistical models, and a growing societal concern to assess and manage uncertainty and risk (see, e.g., Smith, 2010, ch. 1). But although the pace of change has been and will remain rapid, the template of a modern statistical inference has become fairly stable. Moreover, the adoption of statistical methods in other communities, such as Information Theory and Machine Learning, has broadly conformed to the same template (see, e.g., MacKay, 2003; Murphy, 2012).

In this chapter I distinguish between Bayesian statisticians and modern statisticians. It is convenient to think of these as two different tribes, but in reality a professional applied statistician is adaptable. In my experience, some application communities are more comfortable with one approach than with the other and, where possible I try to accommodate this. Having said that, it will be apparent from my assessment that my sympathies lie with the modern Bayesian approach.

2.1 *The origins of Frequentist inference*

The origins of the modern theory of Statistics is found in populations and samples; more specifically, what can be inferred about a population from a sample, suitably collected? In many ways, some of which will become evident in this chapter, Statistics has struggled to escape from these origins.

Consider a population of some kind, of size m. This is simply a set that can be enumerated in some fashion, although it is helpful and not misleading to think of people. Each element of the population has some measurable characteristics. Denote the characteristics of the ith element in the population as X_i, with common realm

$$\mathcal{X} := \{x^{(1)}, \dots, x^{(r)}\}.$$

Questions about the population are ultimately questions about the proportion of the population for which $X_i = x^{(j)}$, for $j = 1, \dots, r$.

In this approach it is not possible to address the question "What are the characteristics of element i?", but only "What is the frequency of characteristic $x^{(j)}$ in the population?". For this reason, it is known as the *Frequentist approach* to inference.[1]

Incomplete knowledge about the population is a form of *epistemic uncertainty*. To represent her epistemic uncertainty, the Frequentist statistician proposes a *model* for the population. 'Model' in this context has a precise statistical meaning: it is a family of distributions, where by 'family' is meant 'set', and where the index of the family is termed the *parameter*:

$$\text{model:} \quad \big\{ f(\cdot\,; \theta) : \theta \in \Omega \big\}$$

where θ is the parameter and Ω is the *parameter space*. The model asserts that $f(x^{(j)}; \theta)$ is the proportion of the population with $X_i = x^{(j)}$ in family member θ. Hence

$$\big(f(x^{(1)}; \theta), \dots, f(x^{(r)}; \theta) \big) \in \mathbb{S}^{r-1}$$

for each $\theta \in \Omega$, where $\mathbb{S}^{r-1}$ is the $(r-1)$-dimensional unit simplex, defined in (1.1).

What does the introduction of a model bring? In the vacuous model, $\Omega = \mathbb{S}^{r-1}$ and $f(x^{(j)}; \theta) = \theta_j$ for $j = 1, \dots, r$. But if the dimension of Ω is less than $r-1$, then the family is restricted in some way. This restriction is used to represent beliefs about the population. As a canonical example, it is common to choose a Normal (or 'Gaussian') model for a scalar X_i, which represents the belief that the population proportions are bell-shaped. In this case, the only variations in the shape are where it is centred, and how wide it is. This gives us a two-dimensional parameter $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++} = \Omega$, and the model[2]

$$f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\} \mathrm{d}x.$$

The two-dimensional parameter has restricted the vacuous model, which has a notionally uncountable number of 'degrees of belief', to a model with only two degrees of belief. This is a *lot* of dimensional reduction or, to put it another way, it represents very strong beliefs indeed.

Now a sample is collected, denoted $\boldsymbol{Y} := (Y_1, \dots, Y_n)$, where each $Y_i \in \mathcal{X}$. It is necessary to describe how the sample is selected from the population. One type of selection mechanism has a dramatic effect in simplifying the inference, which is that the elements in the sample are selected without reference to the values $X_1, \dots, X_m$. A sample of this type is termed *ignorable*; see Gelman *et al.* (2014, ch. 8) for more details.

The ubiquitous example of an ignorable selection mechanism is *random sampling*, without or with repacement. In random sampling the entire population is enumerated, a subset of size n is identified using a random mechanism, and then the values of X_i are collected for this subset. If the subset is selected in such a way that every one

[1] Although the scope of this label will be generalised below, notably in Sec. 2.5.

[2] A notational point. In order to treat the model like a PMF, I include the infinitesimal 'dx' when the realm is uncountable; see Sec. 1.6.3.

of the $\binom{m}{n}$ subsets is equally probable, then this is random sampling *without replacement*. If the sampling fraction n/m is small, then the difference between random sampling without replacement and random sampling *with replacement* is small; see Freedman (1977) for details. Random sampling with replacement is typically taken to represent the more general notion of an ignorable selection mechanism.

Under random sampling with replacement, the PMF of the sampled values is

$$\mathrm{p}(\boldsymbol{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta) \quad \text{for some } \theta \in \Omega. \tag{2.1}$$

In terms of Sec. 1.5, $\boldsymbol{Y}$ is mutually independent (MI) for each $\theta \in \Omega$. In fact, even stronger, $\boldsymbol{Y}$ is independent and identically distributed (IID) for each $\theta \in \Omega$, often written as

$$Y_1, \dots, Y_n \overset{\text{iid}}{\sim} f(\cdot\,;\theta) \quad \text{for some } \theta \in \Omega.$$

This is a very tractable PMF, and the basis for some very elegant mathematical results.

We should pause for a moment. Except in trivial experiments, I doubt that anyone in the history of the world has sampled a population of people at random without/with replacement. This presupposes that the entire population can be enumerated and contacted, and that every element of the population who is contacted will respond. Neither of these conditions holds in practice, and there are many examples of woeful experiments with a large *sampling bias*, where the sample collected is not representative of the population in the way that a random sample would be representative. The very first question a statistician should ask about a sample is whether or not it has sampling bias. This question is more pertinent than ever in this era of Big Data, where samples are often 'samples of opportunity' rather than samples collected through a carefully designed selection mechanism. Harford (2014) has an excellent discussion.

The finishing line for Frequentist inference in now in sight. If we knew θ, then we would know the population proportions exactly, and we would also know the answer to our question, which is some function of the proportions, and hence some function of θ, say $g(\theta)$. One approach is to estimate θ from the sample observations $\boldsymbol{y}^{\text{obs}}$, and then plug in the estimate to derive a point estimate of the population proportions. The most popular estimate is the *Maximum Likelihood (ML)* estimate,

$$\hat{\theta}(\boldsymbol{y}^{\text{obs}}) := \operatorname*{argmax}_{t \in \Omega} \mathrm{p}(\boldsymbol{y}^{\text{obs}};t) = \operatorname*{argmax}_{t \in \Omega} f(\boldsymbol{y}^{\text{obs}};t). \tag{2.2}$$

Then $g\big(\hat{\theta}(\boldsymbol{y}^{\text{obs}})\big)$ is the ML estimate of $g(\theta)$. Most statisticians would be concerned to quantify the variation induced by the random sampling (this variation will decrease as n increases), and would prefer to compute and present a *95% Confidence Interval* for $g(\theta)$ based on $\boldsymbol{y}^{\text{obs}}$.

However, this approach is not as clear-cut as it seems, because there is an uncountable number of point estimators and confidence procedures for θ and $g(\theta)$, and their properties vary widely, and depend on the model and on g. In fact it would not be an over-simplification to say that much of C20th Frequentist statistics has been about proposing, examining, and sometimes rejecting different approaches to narrowing the set of possible point estimators and confidence procedures. Some proposals have been more-or-less completely abandoned (e.g. the notion that an estimator should be 'unbiased'); others continue to be used despite having unattractive features (e.g. the ML estimator, which is sometimes inadmissible). Much more detail about choosing point estimators and confidence procedures is given in Chapter 3 and Chapter 4.

Also, many modern statistical inferences do not have the simplifying features of population inference. In particular, the concept of a population and a model is less well-defined, inference is not just about parameters, and the PMF of the observations does not have the tractable product form given in (2.1). So, really, nothing is clear-cut! Statistics is the antithesis of what many non-statisticians would like it to be: a more-or-less belief-free approach to assessing the evidence for a hypothesis, based on some observations. Instead, there are beliefs encoded in the model, beliefs encoded in the choice of inferential procedure, and beliefs encoded in the computation. Statistical training is about recognising the presence of all of these beliefs, and developing the knowledge and experience to make good choices, or at least to avoid bad ones.

2.2 *Exchangeability*

Exchangeability is a qualitative belief about the population. It is important because it provides a stepping-stone from the Frequentist approach to population inference, to the much more general approaches presented in the following sections.

Beliefs about a sequence $\boldsymbol{X} := (X_1, \dots, X_m)$ are *exchangeable* exactly when

$$\forall g : \mathcal{X} \to \mathbb{R}, \forall \boldsymbol{\pi} \quad \mathrm{E}\{g(X_1, \dots, X_m)\} = \mathrm{E}\{g(X_{\pi_1}, \dots, X_{\pi_m})\} \tag{2.3}$$

where $\boldsymbol{\pi}$ is a permutation of $(1, \dots, m)$. An equivalent condition is that the PMF of $\boldsymbol{X}$ is a symmetric function of $\boldsymbol{x}$. Put simply, if beliefs about $\boldsymbol{X}$ are exchangeable then only values count: the question of who had which value is immaterial.

The power of exchangeability comes from the *Exchangeability Representation Theorem (ERT)*. One version of the ERT is as follows. Let $\boldsymbol{X}$ be an exchangeable sequence of length m, where m is large, and let $\boldsymbol{Y} := (Y_1, \dots, Y_n)$ be a sample from $\boldsymbol{X}$ selected without reference to the values in $\boldsymbol{X}$ (i.e. 'ignorably'). If $n \ll m$ then there exists a statistical model f and a *prior distribution* π_θ for which, to a good approximation (exact as $m \to \infty$)

$$\mathrm{p}(\boldsymbol{y}) = \int \prod_{i=1}^{n} f(y_i; t) \cdot \pi_\theta(t) \, \mathrm{d}t. \tag{2.4}$$

In other words the model and the prior distribution are jointly implicit in any exchangeable belief about $\boldsymbol{X}$. It is very striking that the qualitative belief of exchangeability for $\boldsymbol{X}$ translates into such a tightly constrained PMF for $\boldsymbol{Y}$.

In his original article, de Finetti (1937) proved the ERT for the special case where $\mathcal{X} = \{0, 1\}$; see also Heath and Sudderth (1976). This result was extended to complete generality by Hewitt and Savage (1955), and several other proofs have been given. I like Kingman (1978), although it is technical.

Eq. (2.4) is a joint PMF for $(\boldsymbol{Y}, \theta)$ in which θ has been marginalised out. If we condition the joint distribution on $\theta \doteq t$ then

$$\mathrm{p}(\boldsymbol{y} \mid t) = \prod_{i=1}^{n} f(y_i; t), \quad t \in \operatorname{supp} \pi_\theta.$$

In terms of Sec. 1.5, $\boldsymbol{Y}$ is mutually conditionally independent (MCI) given θ. This highlights a fundamental difference between (2.1) and (2.4): in the first case θ is an index of the set Ω, and in the second case it is a random variable.[3] It is confusing to give the Frequentist parameter and the exchangeable random variable the same symbol θ. But this practice is totally entrenched in our profession's notation, and it is too late to change.

It is important to appreciate that the PMF for $\boldsymbol{Y}$ in (2.4) does not imply any particular PMF for $\boldsymbol{X}$. In fact, there is an uncountable number of PMFs for $\boldsymbol{X}$ which are exchangeable, and which have (2.4) as the marginal PMF for $\boldsymbol{Y}$. However, one particular candidate is very attractive, namely

$$\mathrm{p}(\boldsymbol{x}) = \int \prod_{i=1}^{m} f(x_i; t) \cdot \pi_\theta(t) \, \mathrm{d}t. \tag{2.5}$$

It seems to be only short step to *start* with (2.5) as beliefs about $\boldsymbol{X}$, from which (2.4) is the marginal PMF for any ignorable sample $\boldsymbol{Y}$. But this is a huge step conceptually, because it proposes a PMF for the population rather than just the sample, and opens up a whole new vista for inference. It is now possible to make inferences directly about each X_i, and the observations $\boldsymbol{y}^{\text{obs}}$ can be incorporated into beliefs about unobserved X_i's by conditioning on the proposition $\boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}$.

The application of the ERT to the Frequentist inference of Sec. 2.1 suggests that proposing a prior distribution π_θ and conditioning on $\boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}$ is an alternative to choosing an estimator or a confidence procedure for θ. This was the basis for the *Neo-Bayesian* movement which was spearheaded in the 1950s by L.J. Savage (and an honorable mention for I.J. Good), based in part on the trenchant criticisms of the Frequentist approach by Bruno de Finetti and Harold Jeffreys.

[3] I am being pedantic here, and not calling θ a 'random quantity', because it need not be operationally defined.

The Neo-Bayesian approach fixed a number problems with the Frequentist approach, notably that it is often *inadmissible*, discussed in detail in Chapter 3.[4] But, with the benefit of hindsight, I think this misses the big picture. What the ERT really did was enable a much more general concept of the purpose and practice of statistical modelling, and usher in the 'modern Bayesian' era.

2.3 *The modern Bayesian approach*

Here is the general situation. We have random quantities of interest, $\boldsymbol{X}$. Another set of random quantities are *observables*, $\boldsymbol{Y}$, for which we acquire *observations*, $\boldsymbol{y}^{\text{obs}}$. Sometimes $\boldsymbol{Y} \subset \boldsymbol{X}$, but often not. Our basic objective is to update our beliefs about $\boldsymbol{X}$ using the values $\boldsymbol{y}^{\text{obs}}$.

One particular approach has proved powerful in practice, and has strong theoretical support in Philosophy and in Computer Science as a model for a rational approach to learning (see, e.g., van Fraassen, 1989; Paris, 1994). This approach is to represent beliefs about $(\boldsymbol{X}, \boldsymbol{Y})$ as a PMF, and to update beliefs about $\boldsymbol{X}$ by conditioning on $\boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}$. This approach is known by the ugly word *conditionalization*. There are interesting computational questions about how best to implement the conditioning, but these are really questions for Probability Theory (many theoretical statisticians are applied probabilists). For the applied statistician, the crucial question is how to construct the PMF for $(\boldsymbol{X}, \boldsymbol{Y})$.

Exchangeability provides the template for statistical modelling. Our beliefs about $(\boldsymbol{X}, \boldsymbol{Y})$ are complicated, because of all the things we know about how elements of $(\boldsymbol{X}, \boldsymbol{Y})$ are similar to each other; e.g., how the reading ability of one child of a specified age relates to that of the same child at a different age, or another child of the same or a different age. It would be very challenging to write down a PMF directly. Instead, the trick is to sneak up on it. We introduce additional *random variables* $\theta \in \Omega$. I will write t for a representative value for θ, and I will treat Ω, the realm of θ, as uncountably infinite: this explains the presence of 'dt' in the expressions below; see Sec. 1.6.3.[5] What θ represents is entirely up to the statistician: it is simply a device to allow her to specify a joint distribution for $(\boldsymbol{X}, \boldsymbol{Y}, \theta)$ using the simplifications that arise from *conditional independence*.[6]

(FAC) asserts that we can always decompose the joint PMF as

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}, t)\, \mathrm{d}t = \begin{cases} \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}, t) \cdot \mathrm{p}(\boldsymbol{y} \mid t) \cdot \mathrm{p}(t)\, \mathrm{d}t, \text{ or} \\ \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, t) \cdot \mathrm{p}(\boldsymbol{x} \mid t) \cdot \mathrm{p}(t)\, \mathrm{d}t. \end{cases} \tag{2.6}$$

Conditional independence allows us to simplify one or more of the PMFs on the righthand side, to the point, one hopes, where a specific choice can be made. It is worth stressing again that the statistician chooses her parameters θ in order to make these choices as simple as possible.

[4] A good snapshot of our profession in transition is given by Savage *et al.* (1962).

[5] Typically, Ω is a convex subset of a finite-dimensional Euclidean space.

[6] Now would be a good time to look back at Sec. 1.5.

Here are some elementary but useful examples, from Sec. 2.2. In the special case where $\boldsymbol{Y} \subset \boldsymbol{X}$, we have

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, t) = \mathbb{1}_{\boldsymbol{y} \doteq x_{1:n}},$$

where I have assumed that $\boldsymbol{Y}$ corresponds to the first n elements of $\boldsymbol{X}$, without loss of generality. This very specific case implies that $\boldsymbol{Y} \perp\!\!\!\perp \theta \mid \boldsymbol{X}$, because the righthand side is invariant to t.

For the exchangeable models of Sec. 2.2, the conditional independence is $\vDash \boldsymbol{X} \mid \theta$, or, equivalently,

$$\mathrm{p}(\boldsymbol{x} \mid t) = \prod_{i=1}^{m} \mathrm{p}(x_i \mid t).$$

Then the statistician chooses $\mathrm{p}(x_i \mid t) \leftarrow f(x_i; t)$. This expression illustrates a useful convention, that 'p' is used to denote generic PMFs, and other symbols like 'f' and 'π_θ' are used to indicate specified PMFs. So 'p' is a function which obeys the rules laid down in Sec. 1.4, but '$f(x;t)$' is a specific choice of function and evaluates to a number.[7]

So, following the second branch of (2.6), the exchangeable model has

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}, t)\,\mathrm{d}t = \mathbb{1}_{\boldsymbol{y} \doteq x_{1:n}} \cdot \prod_{i=1}^{m} f(x_i; t) \cdot \pi_\theta(t)\,\mathrm{d}t \tag{2.7}$$

where $\mathrm{p}(t) \leftarrow \pi_\theta(t)$. It is a pragmatic choice by the statistician, to introduce θ and to specify $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}, t)$ in terms of f and π_θ, rather than to specify $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y})$ directly. But this approach is so powerful that it deserves to be presented as a principle.

Definition 4 (Principle of statistical modelling, PSM)**.** *Represent the complicated joint beliefs you want for $(\boldsymbol{X}, \boldsymbol{Y})$ by introducing additional random variables θ, specifying a relatively simple joint distribution for $(\boldsymbol{X}, \boldsymbol{Y}, \theta)$ using conditional independence, and then marginalising out θ.*

The crucial thing about the PSM is that the parameters are entirely instrumental: they need have no 'meaning', because their purpose is simply to induce an appropriate PMF for the random quantities $(\boldsymbol{X}, \boldsymbol{Y})$ after the parameters have been marginalised out. Ultimately, all of our inferences concern $(\boldsymbol{X}, \boldsymbol{Y})$. There may be situations in which expectations of specified functions of $(\boldsymbol{X}, \boldsymbol{Y})$ are expressible in terms of specified functions of θ, but this should not be presupposed.

From this point of view, Frequentist inference, which focuses largely on inferences about specified functions of the parameters, is rather limited. It has proved very hard to extend Frequentist inference to the general case of inferences about functions of $(\boldsymbol{X}, \boldsymbol{Y})$. The usual approach, given below in Sec. 2.5, is acknowledged by all statisticians to be deficient, although is not possible to assert categorically that the Bayesian approach is better, since inferences in the Bayesian approach can be influenced by the choice of prior distribution for θ.

[7] In more general cases, f is an algorithm from which we can simulate an X with the PMF $f(\cdot\,; t)$, for each $t \in \Omega$.

* * *

Acceptance of the PSM is the hallmark of *modern Bayesians*, in contrast to the neo-Bayesians mentioned above, who were more concerned to correct deficiencies in the Frequentist approach. Dennis Lindley, who had worked closely with L.J. Savage, was one of the early modern Bayesians (see, e.g., Lindley and Smith, 1972; Lindley, 1980). A modern statistical model for $(\boldsymbol{X}, \boldsymbol{Y})$ is constructed from notions of exchangeability and conditional independence, typically represented as a *hierarchical model*. Such a model has several layers of parameters, quite different from the 'flat' parameter space implied by the Frequentist approach. The potential for strong dependencies among parameters has challenged simplistic notions about the dimension of the parameter space (see, e.g., Spiegelhalter *et al.*, 2002, 2014).

2.4 *Computation for 'conditionalization'*

The hypothetical expectation or PMF when conditioning on $\boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}$ is indicated by an asterisk. Throughout this section I will assume that $\boldsymbol{y}^{\text{obs}} \in \operatorname{supp} \boldsymbol{Y}$, since at this stage anything else would be daft. Let g be any specified function of $(\boldsymbol{x}, \boldsymbol{y})$. Then

$$\begin{aligned} \mathrm{E}^*\{g(\boldsymbol{X}, \boldsymbol{Y})\} &:= \mathrm{E}\{g(\boldsymbol{X}, \boldsymbol{Y}) \mid \boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}\} \\ &= \mathrm{E}\{g(\boldsymbol{X}, \boldsymbol{y}^{\text{obs}}) \mid \boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}\} && \text{by (TWK)} \\ &= \sum_{\boldsymbol{x}} \int g(\boldsymbol{x}, \boldsymbol{y}^{\text{obs}}) \cdot \mathrm{p}^*(\boldsymbol{x}, t)\, \mathrm{d}t && \text{by the CFTP, (1.16).} \end{aligned} \tag{2.8}$$

Examples of $\mathrm{p}^*(\boldsymbol{x}, t)\, \mathrm{d}t$ are given later in this section. I do not rule out including θ among the arguments to g, but useful inferences are usually about operationally defined random quantities, rather than statistically convenient random variables.

In much modern practice, a *Monte Carlo* technique is used to generate a finite sequence of values for $(\boldsymbol{X}, \theta)$ based on $\mathrm{p}^*(\boldsymbol{x}, t)$, and then the sum/integral in (2.8) is replaced by the arithmetic mean over the sequence; one sequence can serve for any number of different choices of g. This practice is justified asymptotically under a number of different Monte Carlo sampling schemes, although the most popular scheme by far is *Markov chain Monte Carlo (MCMC)*. There are many textbooks on this topic, see, e.g., Robert and Casella (2004). For shorter introductions, see Besag *et al.* (1995) and Besag (2004).

The power of MCMC techniques such as the Metropolis-Hastings algorithm is that they can be used to compute hypothetical expectations in cases where $\mathrm{p}^*(\boldsymbol{x}, t)$ is only known only up to a multiplicative constant. Every $\mathrm{p}^*(\boldsymbol{x}, t)$ has $\mathrm{p}(\boldsymbol{y}^{\text{obs}})$ in the demoninator, and this is typically an expensive quantity to evaluate.[8] The growth of applications of conditionalization has gone hand-in-hand with more powerful computers and better MCMC algorithms.

[8] See, e.g., (2.9b) below.

I'd like to finish this brief comment about MCMC on a cautionary note. Implementing an efficient MCMC algorithm involves quite a lot of mathematics and programming; it is easy to make a mistake. Here is my scheme for minimising code errors. For a given model, start by writing a `rmodel` function which generates a random $(\boldsymbol{X}, \boldsymbol{Y})$ for specified parameters. Then write the sampler. Then test the sampler using the `rmodel` function and the method of Cook *et al.* (2006).[9] This testing will require you to perform the inferential calculation many times, so you may want to reduce the size of $(\boldsymbol{X}, \boldsymbol{Y})$ while testing. It would be embarrassing to describe some of the errors this approach has saved me from.

* * *

At this point, if you have done a Bayesian Statistics course you may be looking at (2.8) and asking "Hang on—where is the 'posterior distribution', and where is Bayes's theorem?" The answer is that they sometimes appear in $\mathrm{p}^*(\boldsymbol{x}, t)$ because of the way that we have chosen to factorise $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}, t)$. But they are not an essential feature of conditionalization.

To illustrate the situation where the posterior distribution appears, consider the exchangeable model based on f and π_θ, given in (2.7). Then

$$\begin{aligned}
\mathrm{p}^*(\boldsymbol{x}, t)\,\mathrm{d}t &= \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}^{\text{obs}}, t)\,\mathrm{d}t \Big/ \mathrm{p}(\boldsymbol{y}^{\text{obs}}) \\
&= \mathbb{1}_{\boldsymbol{y}^{\text{obs}} \doteq \boldsymbol{x}_{1:n}} \cdot \prod_{i=1}^{m} f(x_i; t) \cdot \pi_\theta(t)\,\mathrm{d}t \Big/ \mathrm{p}(\boldsymbol{y}^{\text{obs}}) \\
&= \mathbb{1}_{\boldsymbol{y}^{\text{obs}} \doteq \boldsymbol{x}_{1:n}} \cdot \prod_{i=n+1}^{m} f(x_i; t) \prod_{i=1}^{n} f(y_i^{\text{obs}}; t) \cdot \pi_\theta(t)\,\mathrm{d}t \Big/ \mathrm{p}(\boldsymbol{y}^{\text{obs}}) \\
&= \mathbb{1}_{\boldsymbol{y}^{\text{obs}} \doteq \boldsymbol{x}_{1:n}} \cdot \prod_{i=n+1}^{m} f(x_i; t) \cdot \pi_\theta^*(t)\,\mathrm{d}t
\end{aligned} \tag{2.9a}$$

where π_θ^* is the *posterior distribution* of θ (which follows by Bayes's theorem), and

$$\mathrm{p}(\boldsymbol{y}^{\text{obs}}) = \int \mathrm{p}(\boldsymbol{y}^{\text{obs}}, t)\,\mathrm{d}t = \int \prod_{i=1}^{n} f(y_i^{\text{obs}}; t) \cdot \pi_\theta(t)\,\mathrm{d}t. \tag{2.9b}$$

Eq. (2.9) represents an algorithm for evaluating $\mathrm{p}^*(\boldsymbol{x}, t)$ for any choice of f and π_θ. As already explained, MCMC methods allow us to ignore the value of $\mathrm{p}(\boldsymbol{y}^{\text{obs}})$ if that is more convenient, so that only the numerator of (2.9a) is required. Obviously this is a huge advantage, because (2.9b) can be very expensive to compute if the parameter space is large.

Here is another very important illustration, which tells a different story. In many applications in spatial statistics X_i represents random quantities from region i, and Y_i represents measurements made on the random quantities in region i, where only a subset of the regions (the first n) are measured. The natural conditional independence here is

$$Y_i \perp\!\!\!\perp \text{e.e.} \mid X_i \qquad i = 1, \dots, n,$$

[9] Make sure you understand why this method works because—to be frank—the paper could be clearer.

where 'e.e.' denotes 'everything else'. This gives a PMF which factorises as

$$\begin{aligned} \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}, t)\,\mathrm{d}t &= \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, t) \cdot \mathrm{p}(\boldsymbol{x} \mid t) \cdot \mathrm{p}(t)\,\mathrm{d}t \\ &= \prod_{i=1}^{n} \mathrm{p}(y_i \mid x_i) \cdot \mathrm{p}(\boldsymbol{x} \mid t) \cdot \mathrm{p}(t)\,\mathrm{d}t. \end{aligned}$$

In this factorisation $\mathrm{p}(y_i \mid x_i)$ represents measurement error in region i. Unlike the exchangeable case, there are no natural conditional independence beliefs under which $\mathrm{p}(\boldsymbol{x} \mid t)$ factorises; statisticians typically choose an off-the-shelf model such as a *Gauss Markov random field*, see Rue and Held (2005). For more information on spatial and spatial-temporal modelling, see Cressie and Wikle (2011).

Using the specific choices

$$\mathrm{p}(y_i \mid x_i) \leftarrow f_1(y_i \mid x_i),\ \mathrm{p}(\boldsymbol{x} \mid t) \leftarrow f_2(\boldsymbol{x}; t),\ \text{ and }\ \mathrm{p}(t) \leftarrow \pi_\theta(t)$$

gives

$$\mathrm{p}^*(\boldsymbol{x}, t)\,\mathrm{d}t \propto \prod_{i=1}^{n} f_1(y_i^{\mathrm{obs}} \mid x_i) \cdot f_2(\boldsymbol{x}; t) \cdot \pi_\theta(t)\,\mathrm{d}t$$

where I have suppressed the constant $1/\,\mathrm{p}(\boldsymbol{y}^{\mathrm{obs}})$. This expression is an algorithm for evaluating $\mathrm{p}^*(\boldsymbol{x}, t)$ up to an unknown constant, for any choice of f_1, f_2, and π_θ. As such, it is all that is required for an MCMC evaluation of any hypothetical expectation E^*. But notice that it does not re-arrange into a simple expression involving the posterior distribution of θ. This is because there is no simple closed-form expression for $\mathrm{p}(\boldsymbol{y} \mid t)$. So the notion of a posterior distribution, although it always exists in theory, does not always exist as a simple closed-form expression, even ignoring the multiplicative constant.

Personally, I do not find the notion of a posterior distribution for θ useful. It will show up as a simple closed-form expression, ignoring the multiplicative constant, if $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y} \mid t)$ factorises as $\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}, t) \cdot \mathrm{p}(\boldsymbol{y} \mid t)$, according to the specific choices that have been made in the statistical model. Otherwise, it is implicit. It is not important for the inference, which focuses on E^* for specified functions of $(\boldsymbol{x}, \boldsymbol{y})$. To compute this inference, MCMC methods can get along perfectly well without it. Its presence in textbooks is a legacy of the Neo-Bayesian approach, and also of the preference of authors to use simple examples based on exchangeability. For me, modern Bayesian inference is about conditionalization, the Principle of Statistical Modelling, and the use of numerical methods like MCMC for computing expectations like (2.8).

2.5 *Modern Frequentist inference*

The modern Frequentist asserts that beliefs about parameters are not the kinds of things one could or should describe in terms of probabilities. This one difference rules out the use of conditionalization to incorporate the observations $\boldsymbol{y}^{\text{obs}}$ into updated beliefs about specified functions of $(\boldsymbol{X}, \boldsymbol{Y})$.

In fact, it is more nuanced than this, because there are some parameters that Frequentists are happy to describe with probabilities, such as *random effects*. But not all of them. When pressed on this reluctance, naïve Frequentists may well state that they are troubled by the 'subjectivity' of the prior distribution—as though the model is somehow 'non-subjective'. Or they may claim that different prior distributions will give different outcomes—as though this were not true of different choices of estimator or confidence procedure. Many of their estimators and procedures are not guaranteed to be admissible according to standard loss functions; since inadmissibility is itself a Frequentist construction, this is not encouraging.[10] All in all, it is not easy to be a modern Frequentist, and it is not surprising that, as statistical inference has moved away from samples and populations into more complicated territory, Frequentist inference has given way to Bayesian inference.

Nevertheless, we must still cover Frequentist inference, because most applied statistics is being done by people who are not conversant with modern statistical methods, and are carrying out the kinds of analyses which would not have looked modern in the 1970s. What is worse, much of it is being done using bad methods from the 1970s. I personally have no objection to Frequentist inference and sometimes use it myself. But it should be done properly. That is why much of Chapter 3 and all of Chapter 4 is devoted to it. There are no follow-up chapters on Bayesian inference because the two previous sections of this chapter have said all that can be said at this level of generality.

So our modern Frequentist statistician constructs a model $\{f, \Omega\}$ in much the same way as a modern Bayesian statistician. The difference is that she is reluctant to take the final step and specify a '$\mathrm{p}(t)$' at the end of (2.6). As already stated, this rules out conditionalization as the way of updating beliefs about $\boldsymbol{X}$ using $\boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}$, and other approaches must be found. I will discuss just one here; for more detail, see the material on confidence procedures in Chapter 4.

Going back to Sec. 2.1, one possibility is simply to replace the unknown index θ with an estimate based on $\boldsymbol{y}^{\text{obs}}$. For any given value $t \in \Omega$, the model provides the PMF $f(\boldsymbol{x}, \boldsymbol{y}; t)$. Based on the model, we can compute the conditional PMF

$$f^*(\boldsymbol{x}; t) := f(\boldsymbol{x} \mid \boldsymbol{y}^{\text{obs}}; t) = \frac{f(\boldsymbol{x}, \boldsymbol{y}^{\text{obs}}; t)}{\sum_{\boldsymbol{x}'} f(\boldsymbol{x}', \boldsymbol{y}^{\text{obs}}; t)}.$$

And then beliefs about $g(\boldsymbol{X}, \boldsymbol{Y})$ can be expressed as a function of t,

$$\mathrm{E}^*\{g(\boldsymbol{X}, \boldsymbol{Y}); t\} := \sum_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y}^{\text{obs}}) \cdot f^*(\boldsymbol{x}; t)$$

following the same reasoning as at the start of Sec. 2.4. One possibility at this point is to compute and report the lower and upper

[10] Bayesian procedures carry this guarantee; see Chapter 3 and in particular Sec. 3.3.

limits

$$\inf_{t \in \Omega} \mathrm{E}^*\{g(\boldsymbol{X},\boldsymbol{Y});t\} \quad \text{and} \quad \sup_{t \in \Omega} \mathrm{E}^*\{g(\boldsymbol{X},\boldsymbol{Y});t\}.$$

But in practice the range of values is far to large to be useful, unless Ω is already very tightly constrained. Instead, t is replaced by an estimate based on $\boldsymbol{y}^{\text{obs}}$. Usually this is the ML estimate introduced in (2.2), which gives the point estimate

$$\hat{\mathrm{E}}^*\{g(\boldsymbol{X},\boldsymbol{Y})\} := \mathrm{E}^*\{g(\boldsymbol{X},\boldsymbol{Y});\hat{\theta}(\boldsymbol{y}^{\text{obs}})\}.$$

This is termed a *plug-in estimate*, because $\hat{\theta}(\boldsymbol{y}^{\text{obs}})$ has been 'plugged-in' for the unknown θ. No one thinks it is a good idea to collapse the entire parameter space down to one point. This is why the use of plug-in estimates should be discouraged and replaced by confidence intervals.

One more issue needs to be cleared up. The ML estimate looks like a simple thing to compute, using numerical optimisation. And in some cases it is. These are exactly the cases where the model factorises as $f(\boldsymbol{x},\boldsymbol{y};t) = f_1(\boldsymbol{x} \mid \boldsymbol{y};t) \cdot f_2(\boldsymbol{y};t)$. This is the case for the exchangeable model beloved of textbooks, which gives the impression, quite wrongly, that maximising the probability of the observations with respect to $t \in \Omega$ is straightforward. In many cases, it is practically impossible. Take, for example, the spatial illustration from the end of Sec. 2.4. We need to find

$$\mathrm{p}(\boldsymbol{y}^{\text{obs}};t) = \sum_{\boldsymbol{x}} \mathrm{p}(\boldsymbol{x},\boldsymbol{y}^{\text{obs}};t) = \sum_{\boldsymbol{x}} \prod_{i=1}^{n} f_1(y_i \mid x_i) \cdot f_2(\boldsymbol{x};t)$$

which does not simplify any further. So every evaluation of $\mathrm{p}(\boldsymbol{y}^{\text{obs}};t)$ requires a sum/integral over the whole of $\mathcal{X}$, which could be massive.

This computational problem was first addressed in Besag (1974) who proposed to replace $\mathrm{p}(\boldsymbol{y}^{\text{obs}};t)$ with a more tractable approximation.[11] A more complete answer arrived with Dempster *et al.* (1977), who described the *EM algorithm*. This is a maximisation method for models which include *latent variables*, which would be $\boldsymbol{x}$ in this case. Like any numerical optimisation method, the convergence of the EM algorithm is typically to a local maximum (or occasionally a saddlepoint). See Robert and Casella (2004, ch. 5) for more details and useful variants, and Murphy (2012) for the widespread use of the EM algorithm in Machine Learning.

2.6 *Model checking*

Remember that all beliefs are inherently subjective—they pertain to a person, and can differ from person to person. As I stressed in Chapter 1, this is not something to to conceal, but to recognise and acknowledge. Sooner rather than later the statistician has to confront the issue of whether her beliefs about $\boldsymbol{X}$ are acceptably represented by her E^*.[12] In other words, she has to ask herself

[11] This is one of the iconic papers in Statistics.

[12] I will focus on the modern Bayesian approach, but the same questions could be asked of the Frequentist belief $\hat{\mathrm{E}}^*$, and the same method can be used.

whether she is prepared to sign her name to beliefs constructed in this way.

To set the scene, consider this observation from McWilliams (2007):

> [Atmospheric and Ocean Simulation] models yield space-time patterns remeniscent of nature (e.g., visible in semiquantitative, high-resolution satellite images), thus passing a meaningful kind of *Turing test* between the artifical and the actual. (p. 8709, emphasis added)

This is not a vacuous comparison in climate modelling, because the ocean simulation is not conditioned on the satellite observations. Rubin (1984, sec. 5) proposed a method for making the same kind of comparison in statistical inference. This involves 'cloning' the observations, so that the comparison can be between the actual observations and the clones.[13] The inference passes its Turing test if the statistician cannot spot the observations among the clones.

We have to use our imagination to implement this test. Consider a model which factorises as

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y} \mid t) = \mathrm{p}_{\boldsymbol{Y}|\boldsymbol{X},\theta}(\boldsymbol{y} \mid \boldsymbol{x}, t) \cdot \mathrm{p}_{\boldsymbol{X}|\theta}(\boldsymbol{x} \mid t) \leftarrow f_1(\boldsymbol{y} \mid \boldsymbol{x}; t) \cdot f_2(\boldsymbol{x}; t),$$

which is the common situation. Now imagine creating a clone of the observations, denoted $\boldsymbol{Y}'$, with the two properties

$$\boldsymbol{Y}' \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{X}, \theta \quad \text{and} \quad \mathrm{p}_{\boldsymbol{Y}'|\boldsymbol{X},\theta}(\boldsymbol{y}' \mid \boldsymbol{x}, t) \leftarrow f_1(\boldsymbol{y}' \mid \boldsymbol{x}; t). \qquad (2.10)$$

These are not arbitrary choices, but designed to come as close as possible to the idea of siblings. It is as though $(\boldsymbol{X}, \theta)$ are the genes, and $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ are siblings.[14] Then we see whether we can spot $\boldsymbol{y}^{\mathrm{obs}}$ from among its siblings, *conditioned on* $\boldsymbol{Y} \doteq \boldsymbol{y}^{obs}$. This last part is crucial, if we want to use E^* rather than E to represent our beliefs.

Starting from (2.10), the conditional PMF is

$$\mathrm{p}^*(\boldsymbol{x}, \boldsymbol{y}', t)\,\mathrm{d}t = f_1(\boldsymbol{y}' \mid \boldsymbol{x}; t) \cdot \mathrm{p}^*(\boldsymbol{x}, t)\,\mathrm{d}t.$$

In practice, an MCMC sampler targeting $\mathrm{p}^*(\boldsymbol{x}, t)$ is all that is required. Every now and then, take the current values of $(\boldsymbol{X}, \theta)$ in the chain, and use them to simulate a $\boldsymbol{Y}'$ using f_1. This will give a collection of $\boldsymbol{Y}'$'s that are approximately IID from $\mathrm{p}^*(\boldsymbol{y}')$.

To run the Turing test, the statistician generates a set of clones, and then visualises these and the actual observations $\boldsymbol{y}^{\mathrm{obs}}$, to see if she can spot the actual observations among the clones. The more clones the better of course, in terms of the power of the test, but the statistician's time and patience are also an issue. Often the observations will be recognisable to a statistician with long exposure to the application, so that anonymising them among the clones is not possible. Even in this case, the Turing test can be a powerful way to assess the model. If the clones are appear different from the observations then the nature of the difference is a useful pointer to further model development. There is lots of good advice on model checking and on visualisation in Gelman *et al.* (2014, ch. 6).

[13] Prof. Rubin did not write 'clones', of course. See the next footnote.

[14] So I should really have said 'sibling' instead of 'clone', but the latter is cooler. Also there is not a verb to describe the artifical construction of siblings.

Just for fun, Figure 2.1 is a Turing test for a model of volcanism, with three named volcanoes and nine clones (from currently unpublished work). There is no immediately apparent difference between the observations and the clones (labeled as 'REP'), although on a detailed inspection it looks as though the minimum repose period for the observations might be longer than for the clones; the repose period is the time between eruptions. I would be happy to proceed with this model.

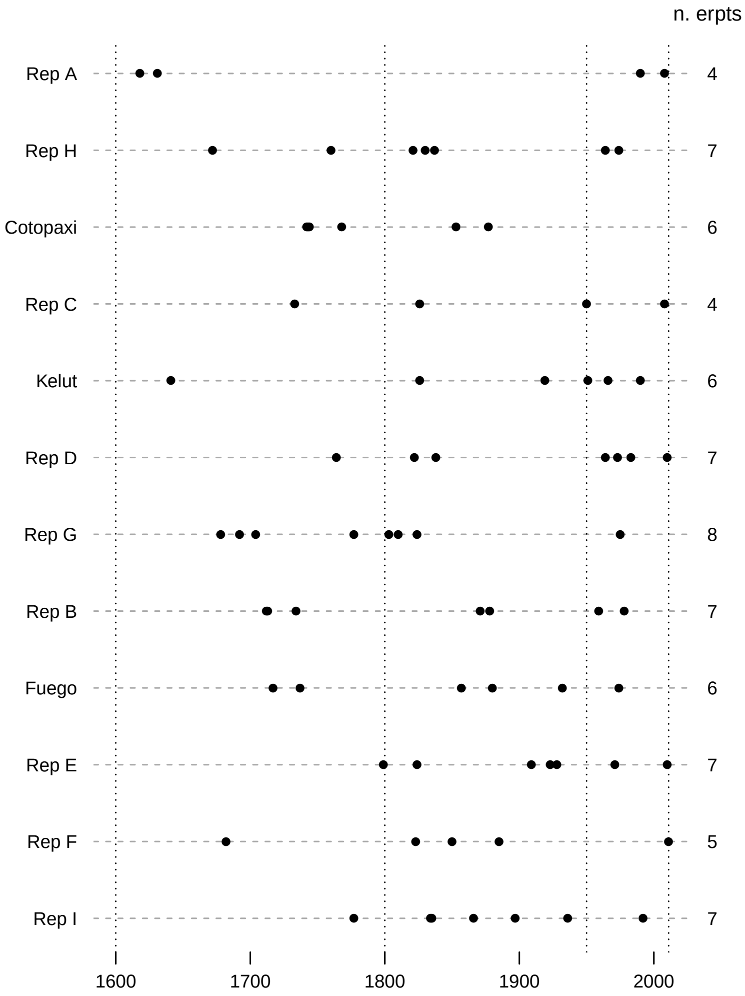

Figure 2.1: 'Turing test' for a model of volcanism, where each dot is a recorded large explosive eruption. The full dataset was much larger. This is a summary image designed to capture some of its key features. In each case, the full dataset is simulated, and then the volcano with the largest number of recorded eruptions is visualised, shown as 'REP'. For the actual dataset, three volcanoes had the maximum six recorded eruptions, and they are shown by name. The vertical dashed lines indicate periods assigned different recording rates, with the recording rate since 1950 being 1.

3
Statistical Decision Theory

The basic premise of Statistical Decision Theory is that we want to make inferences about the parameter of a family of distributions. So the starting point of this chapter is a family of distributions for the observables $\boldsymbol{Y} := (Y_1, \dots, Y_n)$, of the general form

$$\boldsymbol{Y} \sim f(\cdot\,; \theta) \quad \text{for some } \theta \in \Omega,$$

where f is the 'model', θ is the 'parameter', and Ω the 'parameter space', just as in Chapter 2. Nothing in this chapter depends on whether $\boldsymbol{Y}$ is a scalar or a vector, and so I will write Y throughout. The parameter space Ω may be finite or non-finite, possibly non-countable; generally, though, I will treat it as finite, since this turns out to be much simpler. The value $f(y; \theta)$ denotes the probability of $Y \doteq y$ under family member θ. I will assume throughout this chapter that $f(y; \theta)$ is easily computed.

These basic premises, (i) that we are interested in the value of the parameter θ, and (ii) that $f(y; t)$ is easily computed, are both restrictive, as was discussed in Chapter 2. But in this chapter and the next we are exploring the challenges of Frequentist inference, which operates in a more restrictive domain than modern Bayesian inference.

3.1 General Decision Theory

There is a general theory of decision-making, of which Statistical Decision Theory is a special case. Here I outline the general theory, subject to one restriction which always holds for Statistical Decision Theory (to be introduced below). In general we should imagine the statistician applying decision theory on behalf of a client, but for simplicity of exposition I will assume the statistician is her own client.

There is a set of random quantities X with domain $\mathcal{X}$; as above I treat these as a scalar quantity, without loss of generality. The statistician contemplates a set of *actions*, $a \in \mathcal{A}$. Associated with each action is a consequence which depends on X. This is quantified in terms of a *loss function*, $L : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, with larger values indicating worse consequences. Thus $L(a, x)$ is the loss incurred by the statistician if action a is taken and X turns out to be x.

Before making her choice of action, the statistician will observe $Y \in \mathcal{Y}$. Her choice should be some function of the value of Y, and this is represented as a *decision rule*, $\delta : \mathcal{Y} \to \mathcal{A}$. Of the many ways in which she might choose δ, one possibility is to minimise her expected loss, and this is termed the *Bayes rule*,

$$\delta^* := \operatorname*{argmin}_{\delta \in \mathcal{D}} \mathrm{E}\{L(\delta(Y), X)\},$$

where $\mathcal{D}$ is the set of all possible rules. The value $\mathrm{E}\{L(\delta(Y), X)\}$ is termed the *Bayes risk* of decision rule δ, and therefore the Bayes rule is the decision rule which minimises the Bayes risk.

There is a justly famous result which gives the explicit form for a Bayes rule. I will give this result under the restriction anticipated above, which is that the PMF $\mathrm{p}(x \mid y)$ does not depend on the choice of action. Decision theory can handle the more general case, but it is seldom appropriate for Statistical Decision Theory.

Theorem 3.1 (Bayes Rule Theorem, BRT). *A Bayes rule satisfies*

$$\delta^*(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, X) \mid Y \doteq y\} \tag{3.1}$$

whenever $y \in \operatorname{supp} Y$.[1]

This astounding result indicates that the minimisation of expected loss over the space of all functions from $\mathcal{Y}$ to $\mathcal{A}$ can be achieved by the pointwise minimisation over $\mathcal{A}$ of the expected loss conditional on $Y \doteq y$. It converts an apparently intractable problem into a simple one.

Proof. As usual, we take expectations to be completely coherent. Then the FTP (Thm 1.1) asserts the existence of a PMF for (X, Y), which we can factorise as

$$\mathrm{p}(x,y) = \mathrm{p}(x \mid y)\,\mathrm{p}(y)$$

using the notation and concepts from Chapter 1. Now take any $\delta \in \mathcal{D}$, for which

$$\begin{aligned}
\mathrm{E}\{L(\delta(Y), X)\} &= \textstyle\sum_y \sum_x L(\delta(y), x) \cdot \mathrm{p}(x \mid y)\,\mathrm{p}(y) && \text{by the FTP} \\
&\geq \textstyle\sum_y \Big\{ \operatorname{argmin}_a \sum_x L(a, x)\,\mathrm{p}(x \mid y) \Big\}\,\mathrm{p}(y) \\
&= \textstyle\sum_y \Big\{ \sum_x L(\delta^*(y), x)\,\mathrm{p}(x \mid y) \Big\}\,\mathrm{p}(y) && \text{from (3.1) and the CFTP, (1.16)} \\
&= \textstyle\sum_y \sum_x L(\delta^*(y), x) \cdot \mathrm{p}(x \mid y)\,\mathrm{p}(y) \\
&= \mathrm{E}\{L(\delta^*(Y), X)\} && \text{FTP again.}
\end{aligned}$$

Hence δ^* provides a lower bound on the expected loss, over all possible decision rules. Note that the sum over y can actually be over $\operatorname{supp} Y$ if there are y for which $\mathrm{p}(y) = 0$, which ensures that the conditional expectation inside the curly brackets is always well-defined. □

[1] Recollect that $\operatorname{supp} Y$ is the subset of $\mathcal{Y}$ for which $\mathrm{p}(y) > 0$, termed the 'support' of Y.

3.2 *Inference about parameters*

Now consider the special case of Statistical Decision Theory, in which inference is not about some random quantities X, but about the parameter θ. For simplicity I will assume that the parameter space is finite.[2] Furthermore, because nothing in this chapter depends on whether each element of the parameter space is a scalar or a vector, I will treat θ as a scalar and write

$$\Omega := \{\theta_1, \dots, \theta_k\},$$

rather than my usual notation for elements of sets, which is to use superscripts in parentheses (i.e. I will write θ_j rather than $\theta^{(j)}$). A word about notation. I will write 'θ_j' to indicate one of the elements of Ω, and 'θ' to indicate the unknown index of Ω (Frequentist) or the random variable with realm Ω (Bayesian). This is clearer than letting one symbol represent several different things, which is unfortunately a common practice.

The three types of inference about θ are (i) point estimation, (ii) set estimation, and (iii) hypothesis testing. It is a great conceptual and practical simplification that Statistical Decision Theory distinguishes between these three types simply according to their action sets, which are:

Type of inference	Action set $\mathcal{A}$
Point estimation	The parameter space, Ω. See Sec. 3.4.
Set estimation	The set of all subsets of Ω, denoted 2^Ω. See Sec. 3.5.
Hypothesis testing	A specified partition of Ω, denoted $\mathcal{P}$ below. See Sec. 3.6.

One challenge for Statistical Decision Theory is that finding the Bayes rule requires specifying a *prior distribution* over Ω, which I will denote

$$\boldsymbol{\pi} := (\pi_1, \dots, \pi_k) \in \mathbb{S}^{k-1}$$

where $\mathbb{S}^{k-1}$ is the $(k-1)$-dimensional unit simplex, see (1.1). Applying the BRT (Thm 3.1),

$$\begin{aligned} \delta^*(y) &= \underset{a \in \mathcal{A}}{\operatorname{argmin}} \operatorname{E}\{L(a, \theta) \mid Y \doteq y\} \\ &= \underset{a \in \mathcal{A}}{\operatorname{argmin}} \textstyle\sum_j L(a, \theta_j) \cdot \operatorname{p}(\theta_j \mid y) \qquad \text{by the CFTP,} \end{aligned}$$

where the conditional PMF is

$$\operatorname{p}(\theta_j \mid y) = \frac{f(y; \theta_j) \cdot \pi_j}{\Pr(Y \doteq y)} = \frac{f(y; \theta_j) \cdot \pi_j}{\sum_{j'} f(y; \theta_{j'}) \cdot \pi_{j'}} \tag{3.2}$$

by Bayes's Theorem. So the Bayes rule will not be an attractive way to choose a decision rule for Frequentist statisticians, who are reluctant to specify a prior distribution for θ. These statisticians need a different approach to choosing a decision rule.

[2] See the comment after Thm 3.2 for extensions.

The accepted approach for Frequentist statisticians is to narrow the set of possible decision rules by ruling out those that are obviously bad. Define the *risk function* for rule δ as

$$\begin{aligned} R(\delta, \theta_j) &:= \mathrm{E}\{L(\delta(Y), \theta_j); \theta_j\} \\ &= \textstyle\sum_y L(\delta(y), \theta_j) \cdot f(y; \theta_j). \end{aligned} \tag{3.3}$$

That is, $R(\delta, \theta_j)$ is the expected loss from rule δ when $\theta = \theta_j$. A decision rule δ *dominates* another rule δ' exactly when

$$R(\delta, \theta_j) \leq R(\delta', \theta_j) \quad \text{for all } \theta_j \in \Omega,$$

with a strict inequality for at least one $\theta_j \in \Omega$. If you had both δ and δ', you would never want to use δ'.[3] A decison rule is *admissible* exactly when it is not dominated by any other rule; otherwise it is *inadmissible*. So the accepted approach is to reduce the set of possible decision rules under consideration by only using admissible rules.

It is hard to disagree with this approach, although one wonders how big the set of admissible rules will be, and how easy it is to enumerate the set of admissible rules in order to choose between them. This is the subject of Sec. 3.3. To summarise,

Theorem 3.2 (Wald's Complete Class Theorem, CCT)**.** *In the case where both the action set $\mathcal{A}$ and the parameter space Ω are finite, a decision rule δ is admissible if and only if it is a Bayes rule for some prior distribution $\boldsymbol{\pi}$ with strictly positive values.*

There are generalisations of this theorem to non-finite realms for Y, non-finite action sets, and non-finite parameter spaces; however, the results are highly technical. See Schervish (1995, ch. 3), Berger (1985, chs 4, 8), and Ghosh and Meeden (1997, ch. 2) for more details and references to the original literature.

So what does the CCT say? First of all, if you select a Bayes rule according to some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$ then you cannot ever choose an inadmissible decision rule.[4] So the CCT states that there is a very simple way to protect yourself from choosing an inadmissible decision rule. Second, if you cannot produce a $\boldsymbol{\pi} \gg \mathbf{0}$ for which your proposed rule δ is a Bayes Rule, then you cannot show that δ is admissible.

But here is where you must pay close attention to logic. Suppose that δ' is inadmissible and δ is admissible. It does not follow that δ dominates δ'. So just knowing of an admissible rule does not mean that you should abandon your inadmissible rule δ'. You can argue that although you know that δ' is inadmissible, you do not know of a rule which dominates it. All you know, from the CCT, is the family of rules within which the dominating rule must live: it will be a Bayes rule for some $\boldsymbol{\pi} \gg \mathbf{0}$. This may seem a bit esoteric, but it is crucial in understanding modern parametric inference. Statisticians sometimes use inadmissible rules according to standard loss functions. They can argue that yes, their rule δ is or

[3] Here I am assuming that all other considerations are the same in the two cases: e.g. $\delta(y)$ and $\delta'(y)$ take about the same amount of resource to compute.

[4] Here I am using a fairly common notion for vector inequalities. If all components of $\boldsymbol{x}$ are non-negative, I write $\boldsymbol{x} \geq \mathbf{0}$. It in addition at least one component is positive, I write $\boldsymbol{x} > \mathbf{0}$. If all components are positive I write $\boldsymbol{x} \gg \mathbf{0}$. For comparing two vectors, $\boldsymbol{x} \geq \boldsymbol{y}$ exactly when $\boldsymbol{x} - \boldsymbol{y} \geq \mathbf{0}$, and so on.

may be inadmissible, which is unfortunate, but since the identity of the dominating rule is not known, it is not wrong to go on using δ. Nevertheless, it would be better to use an admissible rule.

3.3 The Complete Class Theorem

This section can be skipped once the previous section has been read. But it describes a very beautiful result, Thm 3.2 above, originally due to an iconic figure in Statistics, Abraham Wald.[5] I assume throughout this section that all sets are finite: the realm $\mathcal{Y}$, the action set $\mathcal{A}$, and the parameter space Ω.

The CCT is if-and-only-if. Let $\boldsymbol{\pi}$ be any prior distribution on Ω. Both branches use a simple result that relates the Bayes Risk of a decision rule δ to its Risk Function:

$$\begin{aligned} \mathrm{E}\{L(\delta(Y),\theta)\} &= \textstyle\sum_j \mathrm{E}\{L(\delta(Y),\theta_j);\theta_j\} \cdot \pi_j \quad \text{by (LIE) and (TWK)} \\ &= \textstyle\sum_j R(\delta,\theta_j) \cdot \pi_j. \qquad (\dagger) \end{aligned}$$

The first branch is easy to prove.

Theorem 3.3. *If δ is a Bayes rule for prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$, then it is admissible.*

Proof. By contradiction. Suppose that the Bayes rule δ is not admissible; i.e. there exists a rule δ' which dominates it. In this case

$$\begin{aligned} \mathrm{E}\{L(\delta(Y),\theta)\} &= \textstyle\sum_j R(\delta,\theta_j) \cdot \pi_j && \text{from } (\dagger) \\ &> \textstyle\sum_j R(\delta',\theta_j) \cdot \pi_j && \text{if } \boldsymbol{\pi} \gg \mathbf{0} \\ &= \mathrm{E}\{L(\delta'(Y),\theta)\} \end{aligned}$$

and hence δ cannot have been a Bayes rule, because δ' has a smaller expected loss. The strict inequality holds if δ' dominates δ *and* $\boldsymbol{\pi} \gg \mathbf{0}$. Without it, we cannot deduce a contradiction. □

The second branch of the CCT is harder to prove. The proof uses one of the great theorems in Mathematics, the Supporting Hyperplane Theorem (SHT, given below in Thm 3.5).

Theorem 3.4. *If δ is admissible, then it is a Bayes rule for some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$.*

For a given loss function L and model f, construct the *risk matrix*,

$$R_{ij} := R(\delta_i, \theta_j)$$

over the set of all decision rules. If there are m decision rules altogether (m is finite because $\mathcal{Y}$ and $\mathcal{A}$ are both finite), then R represents m points in k-dimensional space, where k is the cardinality of Ω.

Now consider *randomised rules*, indexed by $\boldsymbol{w} \in \mathbb{S}^{m-1}$. For randomised rule $\boldsymbol{w}$, actual rule δ_i is selected with probability w_i.

[5] For his tragic story, see `https://en.wikipedia.org/wiki/Abraham_Wald`.

The risk for rule $\boldsymbol{w}$ is

$$\begin{aligned} R(\boldsymbol{w}, \theta_j) &:= \textstyle\sum_i \mathrm{E}\{L(\delta_i(Y), \theta_j); \theta_j\} \cdot w_i \qquad \text{by the (LIE)} \\ &= \textstyle\sum_i R(\delta_i, \theta_j) \cdot w_i. \end{aligned}$$

If we also allow randomised rules—and there is no reason to disallow them, as the original rules are all still available as special cases—then the set of risks for all possible randomised rules is the *convex hull* of the rows of the risk matrix R, denoted $[R] \subset \mathbb{R}^k$, and termed the *risk set*.[6] We can focus on the risk set because every point in $[R]$ corresponds to at least one choice of $\boldsymbol{w} \in \mathbb{S}^{m-1}$.

Only a very small subset of the risk set will be admissible. A point $\boldsymbol{r} \in [R]$ is admissible exactly when it is on the lower boundary of $[R]$. More formally, define the 'quantant' of $\boldsymbol{r}$ to be the set

$$Q(\boldsymbol{r}) := \left\{ \boldsymbol{x} \in \mathbb{R}^k : \boldsymbol{x} \le \boldsymbol{r} \right\}$$

(see footnote 4). By definition, $\boldsymbol{r}$ is dominated by every $\boldsymbol{r}'$ for which $\boldsymbol{r}' \in Q(\boldsymbol{r}) \setminus \{\boldsymbol{r}\}$. So $\boldsymbol{r} \in [R]$ is admissible exactly when $[R] \cap Q(\boldsymbol{r}) = \{\boldsymbol{r}\}$. The set of $\boldsymbol{r}$ for satisfying this condition is the lower boundary of $[R]$, denoted $\lambda(R)$.

Now we have to show that every point in $\lambda(R)$ is a Bayes rule for some $\boldsymbol{\pi} \gg \mathbf{0}$. For this we use the SHT, the proof of which can be found in any book on convex analysis.

Theorem 3.5 (Supporting Hyperplane Theorem, SHT)**.** *Let $[R]$ be a convex set in $\mathbb{R}^k$, and let $\boldsymbol{r}$ be a point on the boundary of $[R]$. Then there exists an $\boldsymbol{a} \in \mathbb{R}^k$ not equal to $\mathbf{0}$ such that*

$$\boldsymbol{a}^T \boldsymbol{r} = \min_{\boldsymbol{r}' \in [R]} \boldsymbol{a}^T \boldsymbol{r}'.$$

So let $\boldsymbol{r} \in \lambda(R)$ be any admissible risk. Let $\boldsymbol{a} \in \mathbb{R}^k$ be the coefficients of its supporting hyperplane. Because $\boldsymbol{r}$ is on the lower boundary of $[R]$, $\boldsymbol{a} \gg \mathbf{0}$.[7] Set

$$\pi_j := \frac{a_j}{\sum_{j'} a_{j'}} \qquad j = 1, \dots, k,$$

so that $\boldsymbol{\pi} \in \mathbb{S}^{k-1}$ and $\boldsymbol{\pi} \gg \mathbf{0}$. Then the SHT asserts that

$$\textstyle\sum_j r_j \cdot \pi_j \le \sum_j r'_j \cdot \pi_j \quad \text{for all } \boldsymbol{r}' \in [R]. \qquad (\ddagger)$$

Let $\boldsymbol{w}$ be any randomised strategy with risk $\boldsymbol{r}$. Since $\sum_j r_j \cdot \pi_j$ is the expected loss of $\boldsymbol{w}$ (see †), (‡) asserts that $\boldsymbol{w}$ is a Bayes rule for prior distribution $\boldsymbol{\pi}$. Because $\boldsymbol{r}$ was an arbitrary point on $\lambda(R)$, and hence an arbitrary admissible rule, this completes the proof of Thm 3.4.

3.4 *Point estimation*

For point estimation the action space is $\mathcal{A} = \Omega$, and the loss function $L(\theta_j, \theta_{j'})$ represents the (negative) consequence of choosing θ_j

[6] If $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}$ are m points in $\mathbb{R}^k$, then the convex hull of these points is the set of $\boldsymbol{x} \in \mathbb{R}^k$ for which $\boldsymbol{x} = w_1 \boldsymbol{x}^{(1)} + \cdots + w_m \boldsymbol{x}^{(m)}$ for some $\boldsymbol{w} \in \mathbb{S}^{m-1}$.

[7] Proof: because if $\boldsymbol{r}$ is on the lower boundary, the slightest decrease in any component of $\boldsymbol{r}$ must move $\boldsymbol{r}$ outside $[R]$.

as a point estimate of θ, when the 'true' value of θ is $\theta_{j'}$. Note that this is questionable, if θ does not correspond to an operationally-defined quantity such as the population mean. If θ is a convenient abstraction, then there is no 'true' value.

There will be situations where an obvious loss function $L : \Omega \times \Omega \to \mathbb{R}$ presents itself. But not very often. Hence the need for a generic loss function which is acceptable over a wide range of situations. A natural choice in the very common case where Ω is a convex subset of $\mathbb{R}^d$ is a *convex loss function*,[8]

$$L(\theta_j, \theta_{j'}) \leftarrow h(\theta_j - \theta_{j'}) \tag{3.4}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is a smooth non-negative convex function with $h(\mathbf{0}) = 0$. This type of loss function asserts that small errors are much more tolerable than large ones. One possible further restriction would be that h is an even function.[9] This would assert that under-prediction incurs the same loss as over-prediction. There are many situations where this is *not* appropriate, but in these cases a generic loss function should be replaced by a more specific one.

Proceeding further along the same lines, an even, differentiable and strictly convex loss function can be approximated by a *quadratic loss function*,

$$h(\boldsymbol{x}) \propto \boldsymbol{x}^T Q \, \boldsymbol{x} \tag{3.5}$$

where Q is a symmetric positive-definite $d \times d$ matrix. This follows directly from a Taylor series expansion of h around $\mathbf{0}$:

$$h(\boldsymbol{x}) = 0 + 0 + \tfrac{1}{2}\boldsymbol{x}^T \nabla^2 h(\mathbf{0}) \, \boldsymbol{x} + 0 + O(\|\boldsymbol{x}\|^4)$$

where the first 0 is because $h(\mathbf{0}) = 0$, the second 0 is because $\nabla h(\mathbf{0}) = 0$ since h is minimised at $\boldsymbol{x} = \mathbf{0}$, and the third 0 is because h is an even function. $\nabla^2 h$ is the *hessian matrix* of second derivatives, and it is symmetric by construction, and positive definite at $\boldsymbol{x} = \mathbf{0}$, if h is strictly convex and minimised at $\mathbf{0}$.

In the absence of anything more specific the quadratic loss function is the generic loss function for point estimation. Hence the following result is widely applicable.

Theorem 3.6. *Under a quadratic loss function, the Bayes rule for point prediction is the conditional expectation*

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

A Bayes rule for a point estimation is known as a *Bayes estimator*. Note that although the matrix Q is involved in defining the quadratic loss function in (3.5), it does not influence the Bayes estimator. Thus the Bayes estimator is the same for an uncountably large class of loss functions. Depending on your point of view, this is either its most attractive or its most disturbing feature.

Proof. Here is a proof that does not involve differentiation. The BRT (Thm 3.1) asserts that

$$\delta^*(y) = \operatorname*{argmin}_{t \in \Omega} \mathrm{E}\{L(t, \theta) \mid Y \doteq y\}. \tag{3.6}$$

[8] If Ω is convex then it is uncountable, and hence definitely not finite. But this does not have any disturbing implications for the following analysis.

[9] I.e. $h(-x) = h(x)$.

So let $\psi(y) := \operatorname{E}(\theta \mid Y \doteq y)$. For simplicity, treat θ as a scalar. Then

$$\begin{aligned} L(t,\theta) &\propto (t-\theta)^2 \\ &= (t-\psi(y)+\psi(y)-\theta)^2 \\ &= (t-\psi(y))^2 + 2(t-\psi(y))(\psi(y)-\theta) + (\psi(y)-\theta)^2. \end{aligned}$$

Take expectations conditional on $Y \doteq y$ to get

$$\operatorname{E}\{L(t,\theta) \mid Y \doteq y\} \propto (t-\psi(y))^2 + \operatorname{E}\{(\psi(y)-\theta)^2 \mid Y \doteq y\}. \quad (\dagger)$$

Only the first term contains t, and this term is minimised over t by setting $t \leftarrow \psi(y)$, as was to be shown.

The extension to vector θ with loss function (3.5) is straightforward, but involves more ink. It is crucial that Q in (3.5) is positive definite, because otherwise the first term in (†), which becomes $(t-\psi(y))^T Q\,(t-\psi(y))$, is not minimised if and only if $t = \psi(y)$. □

Note that the same result holds in the more general case of a point prediction of random quantities X based on observables Y: under quadratic loss, the Bayes estimator is $\operatorname{E}(X \mid Y \doteq y)$.

* * *

Now apply the CCT (Thm 3.2) to this result. For quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$.[10] Among the casualties of this conclusion is the Maximum Likelihood Estimator (MLE),

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

Stein's paradox showed that under quadratic loss, the MLE is not admissible in the case of a Multinormal distribution with known variance, by producing an estimator which dominated it. This result caused such consternation when first published that it might be termed 'Stein's bombshell'. See Efron and Morris (1977) for more details, and Samworth (2012) for an accessible proof. Interestingly, the MLE is still the dominant point estimator in applied statistics, even though its admissibility under quadratic loss is questionable.

3.5 Set estimators

For set estimation the action space is $\mathcal{A} = 2^\Omega$, and the loss function $L(C, \theta_j)$ represents the (negative) consequences of choosing $C \subset \Omega$ as a set estimate of θ, when the 'true' value of θ is θ_j. The points made at the start of Sec. 3.4 also apply here.

There are two contrary requirements for set estimators of θ. We want the sets to be small, but we also want them to contain θ. There is a simple way to represent these two requirements as a loss function, which is to use

$$L(C,t) \leftarrow |C| + \kappa \cdot (1 - \mathbb{1}_{t \in C}) \quad \text{for some } \kappa > 0 \qquad (3.7a)$$

where $|C|$ is the cardinality of C.[11] The value of κ controls the

[10] This is under the conditions of Thm 3.2, or with appropriate extensions of them in the non-finite cases.

[11] Here and below I am treating Ω as countable, for simplicity.

trade-off between the two requirements. If $\kappa \downarrow 0$ then minimising the expected loss will always produce the empty set. If $\kappa \uparrow \infty$ then minimising the expected loss will always produce Ω. For κ in-between, the outcome will depend on beliefs about Y and the value y.

It is important to note that the crucial result, Thm 3.7 below, continues to hold for the much more general set of loss functions

$$L(C, t) \leftarrow g(|C|) + h(1 - \mathbb{1}_{t \in C}) \tag{3.7b}$$

where g is non-decreasing and h is strictly increasing. This is a large set of loss functions, which should satisfy most statisticians who do not have a specific loss function already in mind.

For point estimators there was a simple characterisation of the Bayes rule for quadratic loss functions (Thm 3.6). For set estimators the situation is not so simple. However, for loss functions of the form (3.7) there is a simple necessary condition for a rule to be a Bayes rule.

Theorem 3.7. *Under a loss function of the form* (3.7), $\delta : \mathcal{Y} \to 2^\Omega$ *is a Bayes rule* only if*:*

$$\forall y, \forall \theta_j \in \delta(y) \quad \theta_{j'} \notin \delta(y) \implies \mathrm{p}(\theta_{j'} \mid y) \le \mathrm{p}(\theta_j \mid y) \tag{3.8}$$

where $\mathrm{p}(\theta_j \mid y)$ *was defined in* (3.2)*.*

Proof. The proof is by contradiction. Fix y and let $C \leftarrow \delta(y)$. We show that if (3.8) does not hold, then C does not minimise the expected loss conditional on $Y \doteq y$, as required by the BRT (Thm 3.1). Now,

$$\mathrm{E}\{L(C, \theta) \mid Y \doteq y\} = |C| + \kappa \cdot (1 - \Pr\{\theta \in C \mid Y \doteq y\}) \tag{†}$$

using (3.7a), for simplicity. Let $\theta_j \in C$, and let $\theta_{j'} \notin C$, but with $\mathrm{p}(\theta_{j'} \mid y) > \mathrm{p}(\theta_j \mid y)$, contradicting (3.8). In this case, θ_j and $\theta_{j'}$ could be swapped in C, leaving the first term in (†) the same, but decreasing the second. Hence C could not have minimised the expected loss conditional on $Y \doteq y$, and δ could not have been a Bayes rule. □

To give condition (3.8) a simple name, I will refer to it as the 'level set' property, since it almost asserts that $\delta(y)$ must always be a level set of the probabilities $\{\mathrm{p}(\theta_j \mid Y \doteq y) : \theta_j \in \Omega\}$.[12] Chapter 4 provides a tighter definition of this property.

Now relate this result to the CCT (Thm 3.2). First, Thm 3.7 asserts that δ having the level set property for all y is necessary (but not sufficient) for δ to be a Bayes rule for loss functions of the form (3.7). Second, the CCT asserts that being a Bayes rule is a necessary (but not sufficient) condition for δ to be admissible.[13] So unless δ has the level set property for all y then it is impossible for δ to be admissible for loss functions of the form (3.7). This result is embodied in Bayesian approaches to set estimation for θ.

[12] I can only say 'almost' because the property is ambiguous about the inclusion of θ_j and $\theta_{j'}$ for which $\mathrm{p}(\theta_j \mid Y \doteq y) = \mathrm{p}(\theta_{j'} \mid Y \doteq y)$, while a level set is unambiguous.

[13] As before, terms and conditions apply in the non-finite cases.

Definition 5 (High Posterior Probability (HPP) set)**.** *The rule* $\delta : \mathcal{Y} \to 2^{\Omega}$ *is a level-*$(1-\alpha)$ *HPP set exactly when it is the smallest set for which* $\Pr(\theta \in \delta(y) \mid Y \doteq y) \geq 1-\alpha$*.*

This definition acknowledges that for a given level, say $(1-\alpha) \leftarrow 0.95$, it might not be possible to find a set C for which $\Pr(\theta \in C \mid Y \doteq y) = 0.95$, so instead we settle for the smallest set whose probability is at least 0.95.[14] The requirement that $\delta(y)$ is the smallest set automatically ensures that it satisfies the level set property.

Now it is *not* the case that the collection of, say, level 0.95 HPP sets (taken over all $y \in \mathcal{Y}$) is consistent with the Bayes rule for (3.7) for some specified κ. So the level 0.95 HPP sets cannot claim to be a Bayes rule for (3.7). But they satisfy the necessary condition to be admissible for (3.7), which is a good start. Moreover, the level of an HPP set is much easier to interpret than the value of κ.

Things are trickier for Frequentist approaches, which must proceed without a prior distribution for $\theta \in \Omega$, and thus cannot compute $\text{p}(\theta_j \mid Y \doteq y)$. Frequentist approaches to set estimation are based on confidence procedures, which are covered in detail in Chapter 4. We can make a strong recommendation based on Thm 3.7. Denote the Frequentist model as $\{f, \Omega\}$, for which a prior distribution $\boldsymbol{\pi}$ would imply

$$\text{p}(\theta_j \mid Y \doteq y) = \frac{f(y;\theta_j) \cdot \pi_j}{\sum_{j'} f(y;\theta_{j'}) \cdot \pi_{j'}}.$$

Clearly, if $\pi_j = 1/k$ for all j, then $\text{p}(\theta_j \mid Y \doteq y) \propto f(y;\theta_j)$, which which implies that they have the same level sets. So the recommendation is

- Base confidence procedures on level sets of $\{f(y;\theta_j) : \theta_j \in \Omega\}$.

This recommendation ensures that confidence procedures satisfy the necessary condition to be admissible for (3.7). I will be adopting this recommendation in Chapter 4.

3.6 Hypothesis tests

For hypothesis tests, the action space is a partition of Ω, denoted

$$\mathcal{H} := \{H_0, H_1, \dots, H_d\}.$$

Each element of $\mathcal{H}$ is termed a *hypothesis*; it is traditional to number the hypotheses from zero. The loss function $L(H_i, \theta_j)$ represents the (negative) consequences of choosing element H_i, when the 'true' value of θ is θ_j. It would be usual for the loss function to satisfy

$$\theta_j \in H_i \implies L(H_i, \theta_j) = \min_{i'} L(H_{i'}, \theta_j)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice.

I will be quite cavalier about hypothesis tests. If the statistician has a complete loss function, then the CCT (Thm 3.2) applies,

[14] If Ω is uncountable, then it is usually possible to hit 0.95 exactly, in which case C is an 'exact' 95% *High Posterior Density (HPD)* set.

a $\boldsymbol{\pi} \gg \mathbf{0}$ must be found, and there is nothing more to be said. The famous *Neyman-Pearson (NP) Lemma* is of this type. It has $\Omega = \{\theta_0, \theta_1\}$, with $H_i = \{\theta_i\}$, and loss function

L	θ_0	θ_1
H_0	0	ℓ_1
H_1	ℓ_0	0

with $\ell_0, \ell_1 > 0$. The NP Lemma asserts that a decision rule for choosing between H_0 and H_1 is admissible if and only if it has the form

$$\frac{f(y;\theta_0)}{f(y;\theta_1)} \begin{cases} < c & \text{choose } H_1 \\ = c & \text{toss a coin} \\ > c & \text{choose } H_0 \end{cases}$$

for some $c > 0$. This is just the CCT (Thm 3.2).[15]

The NP Lemma is particularly simple, corresponding to a choice in a family with only two elements. In situations more complicated than this, it is extremely challenging and time-consuming to specify a loss function. And yet statisticians would still like to choose between hypotheses, in decision problems whose outcome does not seem to justify the effort required to specify the loss function.[16]

There is a generic loss function for hypothesis tests, but it is hardly defensible. The *0-1 ('zero-one') loss function* is

$$L(H_i, \theta_j) \leftarrow 1 - \mathbb{1}_{\theta_j \in H_i},$$

i.e., zero if θ_j is in H_i, and one if it is not. Its Bayes rule is to select the hypothesis with the largest conditional probability. It is hard to think of a reason why the 0-1 loss function would approximate a wide range of actual loss functions, unlike in the cases of generic loss functions for point estimation and set estimation. This is not to say that it is wrong to select the hypothesis with the largest conditional probability; only that the 0-1 loss function does not provide a very compelling reason.

* * *

There is another approach which has proved much more popular. In fact, it is the dominant approach to hypothesis testing. This is to co-opt the theory of set estimators, for which there *is* a defensible generic loss function, which has strong implications for the selection of decision rules (see Sec. 3.5). The statistician can use her set estimator $\delta : \mathcal{Y} \to 2^{\Omega}$ to make at least some distinctions between the members of $\mathcal{H}$, on the basis of the value of the observable, y^{obs}:

- 'Accept' H_i exactly when $\delta(y^{\text{obs}}) \subset H_i$,
- 'Reject' H_i exactly when $\delta(y^{\text{obs}}) \cap H_i = \emptyset$,
- 'Undecided' about H_i otherwise.

Note that these three terms are given in scare quotes, to indicate that they acquire a technical meaning in this context. We do not use

[15] In fact, $c = (\pi_1/\pi_0) \cdot (\ell_1/\ell_0)$, where (π_0, π_1) is the prior probability for which $\pi_1 = 1 - \pi_0$.

[16] Just to be clear, *important* decisions should not be based on cut-price procedures: an important decision warrants the effort required to specify a loss function.

the scare quotes in practice, but we always bear in mind that we are not "accepting H_i" in the vernacular sense, but simply asserting that $\delta(y^{\text{obs}}) \subset H_i$ for our particular choice of δ.

One very common special case is where $\mathcal{H} = \{H_0, H_1\}$, and one of the elements, say H_0, is a a very small set, even possibly a singleton.[17] This special case is known as *Null Hypothesis Significance Testing (NHST)*, where H_0 is known as the 'null hypothesis'. In this case is is virtually impossible to accept H_0, because set estimators hardly ever shrink down to the size of H_0. So instead we either reject H_0 and accept H_1, or, if we are undecided, we 'fail to reject' H_0.

This type of hypothesis testing is practiced mainly by Frequentist statisticians, and so I will continue in a Frequentist vein. In the Frequentist approach, it is conventional to use a 95% confidence set as the set estimator for hypothesis testing. Other levels, notably 90% and 99%, are occasionally used. If H_0 is rejected using a 95% confidence set, then this is reported as "H_0 is rejected at a significance level of 5%" (occasionally 10% or 1%). Confidence sets are covered in detail in Chapter 4.

This seems quite clear-cut, but we must end on a note of caution. First, the statistician has not solved the decision problem of choosing an element of $\mathcal{H}$. She has solved a different problem. Based on a set estimator, she may reject H_0 on the basis of y^{obs}, but that does not mean she should proceed as though H_0 is false. This would require her to solve the correct decision problem, for which she would have to supply a loss function. So, first caution:

- Rejecting H_0 is not the same as deciding that H_0 is false. Significance tests do not solve decision problems.

Second, loss functions of the form (3.7) may be generic, but that does not mean that there is only one 95% confidence procedure.[18] As Chapter 4 will show, there are an uncountable number of ways of constructing a 95% confidence procedure. In fact, there are an uncountable number of ways of constructing a 95% confidence procedure based on level sets of the likelihood function. So the statistician still needs to make and to justify two subjective choices, leading to the second caution:

- Accepting or rejecting a hypothesis is contingent on the choice of confidence procedure, as well as on the level.

[17] Where H_i is a singleton, it is known as a *simple hypothesis*; otherwise it is a *composite hypothesis*.

[18] The same point can be made about 95% HPP sets, for which there is one for each prior distribution over Ω.

4
Confidence sets

This chapter is a continuation of Chapter 3, and the same conditions hold; re-read the introduction to Chapter 3 if necessary, and the start of Sec. 3.2. In brief, interest focuses on the parameter θ in the model

$$Y \sim f(\cdot\,;\theta) \quad \text{for some } \theta \in \Omega, \tag{4.1}$$

where Y are observables and $f(y;\theta)$ is assumed to be easily computed. The parameter space is denoted

$$\Omega := \{\theta_1, \dots, \theta_k\}$$

for simplicity, even though the parameter may be vector-valued, and the parameter space may be uncountable; typically the parameter space is a convex subset of a finite-dimensional Euclidean space. An element of Ω is denoted θ_j, while θ is used to denote the unknown 'true' index of Ω.[1] The observed value of Y is denoted y^{obs}.

New notation. In this chapter we have the tricky situation in which a specified function $g : \mathcal{Y} \times \Omega \to \mathbb{R}$ becomes a random quantity when Y is a random quantity. Then the distribution of $g(Y, \theta_j)$ depends on the value of θ. Often the value of θ will be the same value as the second argument to g, but this is not implied by simply writing $g(Y, \theta_j)$. So it is best to make the value of θ explicit, when writing about the distribution of $g(Y, \theta_j)$. Hence I write $g(Y, \theta_j)\big|_{\theta=\theta_j}$ to indicate the random quantity $g(Y, \theta_j)$ when $Y \sim f(\cdot\,;\theta_j)$.

4.1 Confidence procedures and confidence sets

A confidence procedure is a special type of decision rule for the problem of set estimation. Hence it is a function of the form $C : \mathcal{Y} \to 2^{\Omega}$, where 2^{Ω} is the set of all sets of Ω.[2] Decision rules for set estimators were discussed in Sec. 3.5.

[1] This is a *façon de parler*. There is no requirement for (4.1) to be true, thank goodness!

[2] In this chapter I am using 'C' for a confidence procedure, rather than 'δ' for a decision rule.

Definition 6 (Confidence procedure). *$C : \mathcal{Y} \to 2^{\Omega}$ is a level-$(1-\alpha)$ confidence procedure exactly when*

$$\Pr\{\theta_j \in C(Y); \theta_j\} \geq 1 - \alpha \quad \textit{for all } \theta_j \in \Omega.$$

If the probability equals $(1-\alpha)$ for all θ_j, then C is an exact *level-$(1-\alpha)$ confidence procedure.*[3]

The value $\Pr\{\theta_j \in C(Y); \theta_j\}$ is termed the *coverage* of C at θ_j. Thus a 95% confidence procedure has coverage of at least 95% for all θ_j, and an exact 95% confidence procedure has coverage of exactly 95% for all θ_j. The diameter of $C(y)$ can grow rapidly with its coverage.[4] In fact, the relation must be extrememly convex when coverage is nearly one, because, in the case where $\Omega = \mathbb{R}$, the diameter at coverage $= 1$ is unbounded. So an increase in the coverage from, say 95% to 99%, could correspond to a doubling or more of the diameter of the confidence procedure. For this reason, exact confidence procedures are highly valued, because a conservative 95% confidence procedure can deliver sets that are much larger than an exact one.

But, immediately a note of caution. It seems obvious that exact confidence procedures should be preferred to conservative ones, but this is easily exposed as a mistake. Suppose that $\Omega = \mathbb{R}$. Then the following procedure is an exact level-$(1-\alpha)$ confidence procedure for θ. First, draw a random variable U with a standard uniform distribution.[5] Then set

$$C(y) := \begin{cases} \mathbb{R} & U \leq 1 - \alpha \\ \{0\} & \text{otherwise.} \end{cases} \tag{†}$$

This is an exact level-$(1-\alpha)$ confidence procedure for θ, but also a meaningless one because it does not depend on y. If it is objected that this procedure is invalid because it includes an auxiliary random variable, then this rules out the method of generating approximately exact confidence procedures using bootstrap calibration (Sec. 4.3.3). And if it is objected that confidence procedures must depend on y, then (†) could easily be adapted so that y is the seed of a numerical random number generator for U. So something else is wrong with (†). In fact, it fails a necessary condition for admissibility that was derived in Sec. 3.5. This will be discussed in Sec. 4.2.

It is helpful to distinguish between the confidence procedure C, which is a function of y, and the result when C is evaluated at $y \leftarrow y^{\text{obs}}$, which is a set in Ω. I like the terms used in Morey *et al.* (2015), which I will also adapt to P-values in Sec. 4.5.

Definition 7 (Confidence set). *$C(y^{obs})$ is a level-$(1-\alpha)$ confidence set exactly when C is a level-$(1-\alpha)$ confidence procedure.*

So a confidence procedure is a function, and a confidence set is a set. If $\Omega \subset \mathbb{R}$ and $C(y^{\text{obs}})$ is convex, i.e. an interval, then a confidence set (interval) is represented by a lower and upper

[3] Exact is a special case. But when it necessary to emphasize that C is not exact, the term 'conservative' is used.

[4] The diameter of a set in a metric space such as Euclidean space is the maximum of the distance between two points in the set.

[5] See footnote 7.

value. We should write, for example, "using procedure C, the 95% confidence interval for θ is $[0.55, 0.74]$", inserting "exact" if the confidence procedure C is exact.

4.2 *Families of confidence procedures*

The trick with confidence procedures is to construct one with a specified level, or, failing that, a specified lower bound on the level. One could propose an arbitrary $C : \mathcal{Y} \to 2^{\Omega}$, and then laboriously compute the coverage for every $\theta_j \in \Omega$. At that point one would know the level of C as a confidence procedure, but it is unlikely to be 95%; adjusting C and iterating this procedure many times until the minimum coverage was equal to 95% would be exceedingly tedious. So we need to go backwards: start with the level, e.g. 95%, then construct a C guaranteed to have this level.

Define a *family of confidence procedures* as $C : \mathcal{Y} \times [0, 1] \to 2^{\Omega}$, where $C(\cdot; \alpha)$ is a level-$(1 - \alpha)$ confidence procedure for each α. If we start with a family of confidence procedures for a specified model, then we can compute a confidence set for any level we choose.

It turns out that families of confidence procedures all have the same form. The key concept is *stochastic dominance*. Let X and Y be two scalar random quantities. Then X stochastically dominates Y exactly when

$$\Pr(X \dot{\leq} v) \leq \Pr(Y \dot{\leq} v) \quad \text{for all } v \in \mathbb{R}.$$

Visually, the distribution function for X is never to the left of the distribution function for Y.[6] Although it is not in general use, I define the following term.

Definition 8 (Super-uniform). *The random quantity X is* super-uniform *exactly when it stochastically dominates a standard uniform random quantity.*[7]

In other words, X is super-uniform exactly when $\Pr(X \leq u) \leq u$ for all $0 \leq u \leq 1$. Note that if X is super-uniform then its support is bounded below by 0, but not necessarily bounded above by 1. Now here is a representation theorem for families of confidence procedures.[8]

Theorem 4.1 (Families of Confidence Procedures, FCP). *Let $g : \mathcal{Y} \times \Omega \to \mathbb{R}$. Then*

$$C(y; \alpha) := \{\theta_j \in \Omega : g(y, \theta_j) > \alpha\} \tag{4.2}$$

is a family of level-$(1 - \alpha)$ confidence procedures if and only if $g(Y, \theta_j)\big|_{\theta=\theta_j}$ is super-uniform for all $\theta_j \in \Omega$. $C(\cdot; \alpha)$ is exact if and only if $g(Y, \theta_j)\big|_{\theta=\theta_j}$ is uniform for all θ_j.

Proof.
($\Leftarrow$). Let $g(Y, \theta_j)\big|_{\theta=\theta_j}$ be super-uniform for all θ_j. Then, for arbitrary

[6] Recollect that the distribution function of X has the form $F(x) := \Pr(X \dot{\leq} x)$ for $x \in \mathbb{R}$.

[7] A standard uniform random quantity being one with distribution function $F(u) = \max\{0, \min\{u, 1\}\}$.

[8] Look back to 'New notation' at the start of the Chapter for the definition of $g(Y; \theta_j)\big|_{\theta=\theta_j}$.

θ_j,

$$\begin{aligned}\Pr\{\theta_j \in C(Y;\alpha);\theta_j\} &= \Pr\{g(Y,\theta_j) \dot{>} \alpha;\theta_j\}\\ &= 1 - \Pr\{g(Y,\theta_j) \dot{\leq} \alpha;\theta_j\}\\ &= 1 - (\leq \alpha) \geq 1 - \alpha\end{aligned}$$

as required. For the case where $g(Y,\theta_j)\big|_{\theta=\theta_j}$ is uniform, the inequality is replaced by an equality.

($\Rightarrow$). This is basically the same argument in reverse. Let $C(\cdot;\alpha)$ defined in (4.2) be a level-$(1-\alpha)$ confidence procedure. Then, for arbtrary θ_j,

$$\Pr\{g(Y,\theta_j) \dot{>} \alpha;\theta_j\} \geq 1 - \alpha.$$

Hence $\Pr\{g(Y,\theta_j) \dot{\leq} \alpha;\theta_j\} \leq \alpha$, showing that $g(Y,\theta_j)\big|_{\theta=\theta_j}$ is super-uniform as required. Again, if $C(\cdot;\alpha)$ is exact, then the inequality is replaced by a equality, and $g(Y,\theta_j)\big|_{\theta=\theta_j}$ is uniform. □

Families of confidence procedures have the very intuitive *nesting property*, that

$$\alpha < \alpha' \implies C(y;\alpha) \supset C(y;\alpha'). \tag{4.3}$$

In other words, higher-level confidence sets are always supersets of lower-level confidence sets from the same family. This has sometimes been used as part of the definition of a family of confidence procedures (see, e.g., Cox and Hinkley, 1974, ch. 7), but I prefer to see it as an unavoidable consequence of the fact that all families must be defined using (4.2) for some g.

* * *

Sec. 3.5 made a recommendation about set estimators for θ, which was that confidence procedures should be based on level sets of $\{f(y;\theta_j) : \theta_j \in \Omega\}$. This was to satisfy a necessary condition to be admissible under the loss function (3.7). Here I restate that recommendation as a property.

Definition 9 (Level Set Property, LSP)**.** *A confidence procedure C has the Level Set Property exactly when*

$$C(y) = \{\theta_j \in \Omega \text{ such that } f(y;\theta_j) > c\}$$

for some c which may depend on y. A family of confidence procedures has the LSP exactly when $C(\cdot\,;\alpha)$ has the LSP for all α, for which c may depend on y and α.

A family of confidence procedures does not necessarily have the LSP. So it is not obvious, but highly gratifying, that it is possible to construct families of confidence procedures with the LSP. Three different approaches are given in the next section.

4.3 *Methods for constructing confidence procedures*

All three of these methods produce families of confidence procedures with the LSP. This is a long section, and there is a summary in Sec. 4.3.4.

4.3.1 *Markov's inequality*

Here is a result that has pedagogic value, because it can be used to generate an uncountable number of families of confidence procedures, each with the LSP.

Theorem 4.2. *Let h be any PMF for Y. Then*

$$C(y;\alpha) := \{\theta_j \in \Omega : f(y,\theta_j) > \alpha \cdot h(y)\} \tag{4.4}$$

is a family of confidence procedures, with the LSP.

Proof. Define $g(y,\theta_j) := f(y;\theta_j)/h(y)$, which may be ∞. Then the result follows immediately from Thm 4.1 because $g(Y,\theta_j)\big|_{\theta=\theta_j}$ is super-uniform for each θ_j:

$$\begin{aligned}
\Pr\{f(Y;\theta_j)/h(Y) \dot{\leq} u;\theta_j\} &= \Pr\{h(Y)/f(Y;\theta_j) \dot{\geq} 1/u;\theta_j\} \\
&\leq \frac{\mathrm{E}\{h(Y)/f(Y;\theta_j);\theta_j\}}{1/u} && \text{Markov's inequality, (1.9)} \\
&\leq \frac{1}{1/u} = u.
\end{aligned}$$

For the final inequality,

$$\begin{aligned}
\mathrm{E}\{h(Y)/f(Y;\theta_j);\theta_j\} &= \sum_{y \in \operatorname{supp} f(\cdot\,;\theta_j)} \frac{h(y)}{f(y;\theta_j)} \cdot f(y;\theta_j) && \text{FTP, Thm 1.1} \\
&= \sum_{y \in \operatorname{supp} f(\cdot\,;\theta_j)} h(y) \\
&\leq 1.
\end{aligned}$$

If $\operatorname{supp} h \subset \operatorname{supp} f(\cdot\,;\theta_j)$, then this inequality is an equality. □

Among the interesting choices for g, one possibility is $g \leftarrow f(\cdot\,;\theta_i)$, for $\theta_i \in \Omega$. Note that with this choice, the confidence set of (4.4) always contains θ_i. So we know that we can construct a level-$(1-\alpha)$ confidence procedure whose confidence sets will always contain θ_i, for any $\theta_i \in \Omega$.

This is another illustration of the fact that the definition of a confidence procedure given in Def. 6 is too broad to be useful. But now we see that insisting on the LSP is not enough to resolve the issue. Two statisticians can both construct 95% confidence sets for θ which satisfy the LSP, using different families of confidence procedures. Yet the first statistician may reject the null hypothesis that $H_0 : \theta = \theta_i$ (see Sec. 3.6), and the second statistician may fail to reject it, for any $\theta_i \in \Omega$.

Actually, the situation is not as grim as it seems. Markov's inequality is very slack (refer to its proof at eq. 1.9), and so the coverage of the family of confidence procedures defined in Thm 4.2 is likely to be much larger than $(1-\alpha)$, e.g. much larger than 95%. Remembering the comment about the rapid increase in the diameter of the confidence set as the coverage increases, from Sec. 4.1, a more likely outcome is that $C(y;0.05)$ is large for many different choices of h, in which case no one rejects the null hypothesis.

All in all, it would be much better to use an exact family of confidence procedures, if one existed. And, for perhaps the most popular model in the whole of Statistics, this is the case.

4.3.2 The Linear Model

The Linear Model (LM) is commonly expressed as

$$Y \overset{\mathrm{D}}{=} X\beta + \epsilon \quad \text{where } \epsilon \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 I_n) \tag{4.5}$$

where Y is an n-vector of observables, X is a specified $n \times p$ matrix of *regressors*, β is a p-vector of *regression coefficients*, and ϵ is an n-vector of *residuals*.[9] The parameter is $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_{++}$.

'$\mathrm{N}_n(\cdot)$' denotes the n-dimensional *Multinormal distribution* with specified expectation vector and variance matrix (see, e.g., Mardia *et al.*, 1979, ch. 3). The symbol '$\overset{\mathrm{D}}{=}$' denotes 'equal in distribution'; this notation is useful here because the Multinormal distribution is closed under affine transformations. Hence Y has a Multinormal distribution, because it is an affine transformation of ϵ. So the LM must be restricted to applications for which Y can be thought of, at least approximately, as a collection of n random quantities each with realm $\mathbb{R}$, and for each of which our uncertainty is approximately symmetric. Many observables fail to meet these necessary conditions (e.g. applications in which Y is a collection of counts); for these applications, we have *Generalized Linear Models (GLMs)*. GLMs retain many of the attractive properties of LMs.

Wood (2015, ch. 7) provides an insightful summary of the LM, while Draper and Smith (1998) give many practical details.

Now I show that the Maximum Likelihood Estimator (MLE) of (4.5) is

$$\begin{aligned} \hat{\beta}(y) &= (X^T X)^{-1} X^T y \\ \widehat{\sigma^2}(y) &= n^{-1}(y - \hat{y})^T (y - \hat{y}) \end{aligned}$$

where $\hat{y} := X\hat{\beta}(y)$.

Proof. For a LM, it is more convenient to minimise $-2 \log f(y; \beta_j, \sigma_j^2)$ over (β_j, σ_j^2) than to maximise $f(y; \beta_j, \sigma_j^2)$.[10] Then

$$-2 \log f(y; \beta_j, \sigma_j^2) = n \log(2\pi\sigma_j^2) + \frac{1}{\sigma_j^2}(y - X\beta_j)^T (y - X\beta_j)$$

from the PDF of the Multinormal distribution. Now use a simple device to show that this is minimised at $\beta_j = \hat{\beta}(y)$ for all values of σ_j^2. I will write $\hat{\beta}$ rather than $\hat{\beta}(y)$:

$$\begin{aligned} &(y - X\beta_j)^T (y - X\beta_j) \\ &\quad = (y - X\hat{\beta} + X\hat{\beta} - X\beta_j)^T (y - X\hat{\beta} + X\hat{\beta} - X\beta_j) \\ &\quad = (y - \hat{y})^T (y - \hat{y}) + 0 + (X\hat{\beta} - X\beta_j)^T (X\hat{\beta} - X\beta_j) \end{aligned} \tag{†}$$

where multiplying out shows that the cross-product term in the middle is zero. Only the final term contains β_j. Writing this term as

$$(\hat{\beta} - \beta_j)^T (X^T X)(\hat{\beta} - \beta_j)$$

[9] Usually I would make Y and ϵ bold, being vectors, and I would prefer not to use X for a specified matrix, but this is the standard notation.

[10] Note my insistence that (β_j, σ_j^2) be considered as an element of the parameter space, *not* as the 'true' value.

shows that if X has full column rank, so that $X^T X$ is positive definite, then (†) is minimised if and only if $\beta_j = \hat{\beta}$. Then

$$-2 \log f(y; \hat{\beta}, \sigma_j^2) = n \log(2\pi\sigma_j^2) + \frac{1}{\sigma_j^2}(y - \hat{y})^T (y - \hat{y}).$$

Solving the first-order condition gives the MLE for $\widehat{\sigma^2}(y)$, and it is easily checked that this is a global minimum. □

Now suppose we want a confidence procedure for β. For simplicity, I will assume that σ^2 is specified, and for practical purposes I would replace it by $\widehat{\sigma^2}(y^{\text{obs}})$ in calculations. This is known as *plugging in* for σ^2. The LM extends to the case where σ^2 is not specified, but, as long as $n/(n-p) \approx 1$, it makes little difference in practice to plug in.[11]

With β_j representing an element of the β-parameter space $\mathbb{R}^p$, and σ^2 specified, we have, from the results above,

$$-2 \log \left(\frac{f(y; \beta_j, \sigma^2)}{f(y; \hat{\beta}(y), \sigma^2)} \right) = \frac{1}{\sigma^2} \{\hat{\beta}(y) - \beta_j\}^T (X^T X) \{\hat{\beta}(y) - \beta_j\}. \quad (4.6)$$

Now suppose we could prove the following.

Theorem 4.3. *With σ^2 specified,*

$$\frac{1}{\sigma^2} \{\hat{\beta}(Y) - \beta_j\}^T (X^T X) \{\hat{\beta}(Y) - \beta_j\} |_{\beta = \beta_j}$$

has a χ_p^2 distribution.

We could define the decision rule:

$$C(y; \alpha) := \left\{ \beta_j \in \mathbb{R}^p : -2 \log \left(\frac{f(y; \beta_j, \sigma^2)}{f(y; \hat{\beta}(y), \sigma^2)} \right) < \chi_p^{-2}(1 - \alpha) \right\}. \quad (4.7)$$

where $\chi_p^{-2}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the χ_p^2 distribution. Under Thm 4.3, (4.6) shows that C in (4.7) would be an exact level-$(1-\alpha)$ confidence procedure for β; i.e. it provides a family of exact confidence procedures. Also note that it satisfies the LSP from Def. 9.

After that build-up, it will come as no surprise to find out that Thm 4.3 is true. Substituting Y for y in the MLE of β gives

$$\hat{\beta}(Y) \overset{\text{D}}{=} (X^T X)^{-1} X^T (X\beta + \epsilon) \overset{\text{D}}{=} \beta + (X^T X)^{-1} X^T \epsilon,$$

writing σ for $\sqrt{\sigma^2}$. So the distribution of $\hat{\beta}(Y)$ is another Multinormal distribution

$$\hat{\beta}(Y) \sim \text{N}_p(\beta, \Sigma) \quad \text{where } \Sigma := \sigma^2 (X^T X)^{-1}.$$

Now apply a standard result for the Multinormal distribution to deduce

$$\{\hat{\beta}(Y) - \beta_j\}^T \Sigma^{-1} \{\hat{\beta}(Y) - \beta_j\} |_{\beta = \beta_j} \sim \chi_p^2 \quad (†)$$

(see Mardia *et al.*, 1979, Thm 2.5.2). This proves Thm 4.3 above. Let's celebrate this result!

[11] As an eminent applied statistician remarked to me: it if matters to your conclusions whether you use a standard Normal distribution or a Student-*t* distribution, then you probability have bigger things to worry about.

Theorem 4.4. *For the LM with σ^2 specified, C defined in* (4.7) *is a family of exact confidence procedures for β, which has the LSP.*

Of course, when we plug-in for σ^2 we slightly degrade this result, but not by much if $n/(n-p) \approx 1$.

This happy outcome where we can find a family of exact confidence procedures with the LSP is more-or-less unique to the regression parameters in the LM. but it is found, approximately, in the large-n behaviour of a much wider class of models, including GLMs, as explained next.

4.3.3 Wilks confidence procedures

There is a beautiful theory which explains how the results from Sec. 4.3.2 generalise to a much wider class of models than the LM. The theory is quite strict, but it almost-holds over relaxations of some of its conditions. Stated informally, if $Y := (Y_1, \dots, Y_n)$ and

$$f(y; \theta_j) = \prod_{i=1}^{n} f_1(y_i; \theta_j) \quad \text{for some } \theta \in \Omega, \tag{4.8}$$

(see Sec. 2.1) and f_1 is a *regular model*, and the parameter space Ω is a convex subset of $\mathbb{R}^p$ (and invariant to n), then

$$-2 \log \left(\frac{f(Y; \theta_j)}{f(Y; \hat\theta(Y))} \right) \bigg|_{\theta = \theta_j} \xrightarrow{\mathrm{D}} \chi^2_p \tag{4.9}$$

where $\hat\theta$ is the Maximum Likelihood Estimator (MLE) of θ, and '$\xrightarrow{\mathrm{D}}$' denotes 'convergence in distribution' as n increases without bound. Eq. (4.9) is sometimes termed *Wilks's Theorem*, hence the name of this subsection.

The definition of 'regular model' is quite technical, but a working guideline is that $f_1(y_i; \theta_j)$ must be smooth and differentiable in θ_j for each y_i; in particular, supp Y_i must not depend on θ_j. Cox (2006, ch. 6) provides a summary of this result and others like it, and more details can be found in Casella and Berger (2002, ch. 10), or, for the full story, in van der Vaart (1998).

This result is true for the LM, because we showed that it is exactly true for any n provided that σ^2 is specified, and the ML plug-in for σ^2 converges on the true value as $n/(n-p) \to 1$.[12] In general, we can use it the same way as in the LM, to derive a decision rule:

$$C(y; \alpha) := \left\{ \theta_j \in \Omega : -2 \log \left(\frac{f(Y; \theta_j)}{f(Y; \hat\theta(Y))} \right) < \chi_p^{-2}(1-\alpha) \right\}. \tag{4.10}$$

As already noted, this C satisfies the LSP. Further, under the conditions for which (4.9) is true, C is also a family of approximately exact confidence procedures.

Eq. (4.10) can be written differently, perhaps more intuitively. Define

$$L(\theta_j; y) := f(y; \theta_j)$$

[12] This is a general property of the MLE, that it is *consistent* when f has the product form given in (4.8).

known as the *likelihood function* of θ_j; sometimes the y argument is suppressed, notably when $y \leftarrow y^{\text{obs}}$. Let $\ell := \log L$, the *log-likelihood function*. Then (4.10) can be written

$$C(y; \alpha) = \left\{ \theta_j \in \Omega : \ell(\theta_j; y) > \ell(\hat{\theta}(y); y) - \kappa(\alpha) \right\} \tag{4.11}$$

where $\kappa(\alpha) := \chi_p^{-2}(1 - \alpha)/2$. In this procedure we keep all $\theta_j \in \Omega$ whose log-likelihood values are within $\kappa(\alpha)$ of the maximum log-likelihood. In the common case where $\Omega \subset \mathbb{R}$, (4.11) gives *'Allan's Rule of Thumb'*:[13]

- For an approximate 95% confidence procedure for a scalar parameter, keep all values of $\theta_j \in \Omega$ for which the log-likelihood is within 2 of the maximum log-likelihood.

The value 2 is from $\chi_1^{-2}(0.95)/2 = 1.9207\ldots \approx 2$.

Bootstrap calibration. The pertinent question, as always with methods based on asymptotic properties for particular types of model, is whether the approximation is a good one. The crucial concept here is *level error*. The coverage that we want is at least $(1 - \alpha)$ everywhere, which is termed the 'nominal level'. But were we to evaluate a confidence procedure such as (4.11) for a general model (not a LM) we would find that, over all $\theta_j \in \Omega$, that the minimum coverage was not $(1 - \alpha)$ but something else; usually something less than $(1 - \alpha)$. This is the 'actual level'. The difference is

$$\text{level error} := \text{nominal level} - \text{actual level}.$$

Level error exists because the conditions under which (4.11) provides an exact confidence procedure are not met in practice, outside the LM. Although it is tempting to ignore level error, experience suggests that it can be large, and that we should attempt to correct for level error if we can.

One method for making this correction is *bootstrap calibration*, described in DiCiccio and Efron (1996). Here are the steps, based on (4.11), although with a generic κ in place of the function $\kappa(\alpha)$:

$$C(y; \kappa) = \left\{ \theta_j \in \Omega : \ell(\theta_j; y) > \ell(\hat{\theta}(y); y) - \kappa \right\}. \tag{4.12}$$

1. Compute a point estimate for θ, say $\hat{\theta}^{\text{obs}} := \hat{\theta}(y^{\text{obs}})$ the ML estimate. Other estimates are also possible, see Sec. 3.4.

2. For $i = 1, \ldots, m$:

   Sample $y^{(i)} \sim f(\cdot\,; \hat{\theta}^{\text{obs}})$, compute and record $\hat{\theta}^{(i)} := \hat{\theta}(y^{(i)})$, and $\hat{\ell}^{(i)} := \ell(\hat{\theta}^{(i)}; y^{(i)})$.

So, at the end of this process we have $\hat{\theta}^{\text{obs}}$ and the sample of values $\{y^{(i)}, \hat{\theta}^{(i)}, \hat{\ell}^{(i)}\}$ for $i = 1, \ldots, m$. Computing the ML estimate has to be a quick procedure because m needs to be large, say 1000s.

[13] After Allan Seheult, who first taught it to me.

Now if we choose a particular value for κ, an empirical estimate of the coverage at $\theta = \hat{\theta}^{\text{obs}}$ is

$$\begin{aligned}\widehat{\text{cvg}}(\kappa) &:= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\hat{\theta}^{\text{obs}} \in C(y^{(i)};\kappa)\} \\ &= \frac{1}{m}\sum\nolimits_i \mathbb{1}\{\ell(\hat{\theta}^{\text{obs}};y^{(i)}) > \hat{\ell}^{(i)} - \kappa\} \\ &= \frac{1}{m}\sum\nolimits_i \mathbb{1}\{\hat{\ell}^{(i)} - \ell(\hat{\theta}^{\text{obs}};y^{(i)}) < \kappa\}.\end{aligned}$$

Therefore to set the empirical coverage to $(1-\alpha)$, κ needs to be the $(1-\alpha)$-quantile of the values

$$\{\hat{\ell}^{(i)} - \ell(\hat{\theta}^{\text{obs}};y^{(i)})\}_{i=1}^{m}.$$

So the final step is to find this value, call it $\kappa^*(\alpha)$, and then compute the confidence set $C(y^{\text{obs}};\kappa^*(\alpha))$ from (4.12).

This is a very complicated procedure, and it is hard to be precise about the reduction in level error that occurs (see DiCiccio and Efron, 1996, for more details). One thing that is definitely informative is the discrepancy between $\kappa^*(\alpha)$ and $\kappa(\alpha)$, which is an indicator of how well the asymptotic conditions hold. Put simply, if the discrepancy is small then either threshold will do. But if the discrepancy is large, then $\kappa(\alpha)$ will not do, and one is forced to use $\kappa^*(\alpha)$, or nothing. A large sample is required, for $(1-\alpha) = 0.95$: accurately estimating the 95th percentile is going to require about $m = 1000$ samples.[14]

4.3.4 *Summary*

With the Linear Model (LM) described in Sec. 4.3.2, we can construct a family of exact confidence procedures, with the LSP, for the parameters β. Additionally—I did not show it but it follows directly—we can do the same for all affine functions of the parameters β, including individual components.

In general we are not so fortunate. It is not that we cannot construct families of confidence procedures with the LSP: Sec. 4.3.1 shows that we can, in an uncountable number of different ways. But their levels will be conservative, and hence they are not very informative. A better alternative, which ought to work well in large-n simple models like (4.8) is to use Wilks's Theorem to construct a family of approximately exact confidence procedures, which have the LSP, see Sec. 4.3.3.

The Wilks approximation can be checked and—one hopes—improved, using bootstrap calibration. Bootstrap calibration is a necessary precaution for small n or more complicated models (e.g. time series or spatial applications). But in these cases a Bayesian approach is likely to be a better choice, which is reflected in modern practice.

[14] See Harrell and Davis (1982) for a simple estimator for quantiles.

4.4 Marginalisation

Suppose that $g : \theta \mapsto \phi$ is some specified function, and we would like a confidence procedure for ϕ. If C is a level-$(1-\alpha)$ confidence procedure for ϕ then it must have ϕ-coverage of at least $(1-\alpha)$ for all $\theta_j \in \Omega$. The most common situation is where $\Omega \subset \mathbb{R}^p$, and g extracts a single component of θ: for example, $\theta = (\mu, \sigma^2)$ and $g(\theta) = \mu$. So I call the following result the Confidence Procedure Marginalisation Theorem.

Theorem 4.5 (Confidence Procedure Marginalisation, CPM). *Suppose that $g : \theta \mapsto \phi$, and that C is a level-$(1-\alpha)$ procedure for θ. Then $g \circ C$ is a level-$(1-\alpha)$ confidence procedure for ϕ.*[15]

Proof. Follows immediately from the fact that $\theta_j \in C(y)$ implies that $\phi_j \in (g \circ C)(y)$ for all y, and hence

$$\Pr\{\theta_j \in C(Y); \theta_j\} \le \Pr\{\phi_j \in (g \circ C)(Y); \theta_j\}$$

for all $\theta_j \in \Omega$. So if C has θ-coverage of at least $(1-\alpha)$, then $g \circ C$ has ϕ-coverage of at least $(1-\alpha)$ as well. □

This result shows that we can derive level-$(1-\alpha)$ confidence procedures for functions of θ directly from level-$(1-\alpha)$ confidence procedures for θ. But it also shows that the coverage of such derived procedures will typically be more than $(1-\alpha)$, even if the original confidence procedure is exact.

There is an interesting consequence of this result based on the confidence procedures defined in Sec. 4.3.2 and Sec. 4.3.3. Taking the latter more general case, consider the family of approximately exact confidence procedures defined in (4.12). Let $g^{-1} \subset \Omega$ be the inverse image of g. Then

$$\begin{aligned} &\phi_j \in (g \circ C)(y; \alpha) \\ &\qquad \iff \exists \theta_j : \phi_j = g(\theta_j) \wedge \theta_j \in C(y; \alpha) \\ &\qquad \iff \max_{\theta_j \in g^{-1}(\phi_j)} \ell(\theta_j; y) > \ell(\hat{\theta}(y); y) - \kappa(\alpha) \end{aligned}$$

The expression on the left of the final inequality is the *profile log-likelihood*,

$$\ell_g(\phi_j; y) := \max_{\theta_j \in g^{-1}(\phi_j)} \ell(\theta_j; y). \tag{4.13}$$

It provides a simple rule for computing a log-likelihood for any function of θ_j. Because $g \circ C$ is conservative, we would expect to be able to reduce the threshold below $\kappa(\alpha)$ if g is not bijective. However, this is not an area where the asymptotic theory is very reliable (i.e. it takes a long time to 'kick in'). A better option here is to use bootstrap calibration to derive a $\kappa^*(\alpha)$ for g, as described in Sec. 4.3.3.

4.5 P-values

There is a general theory for P-values, also known as *significance levels*, which is outlined in Sec. 4.5.2. But first I want to focus on

[15] $g \circ C := \left\{\phi_j : \phi_j = g(\theta_j) \text{ for some } \theta_j \in C\right\}$.

P-values as used in Null Hypothesis Signficance Tests, which is a very common situation.

As discussed in Sec. 4.3, we have methods for constructing families of good confidence procedures, and the knowledge that there are also families of confidence procedures which are poor (including completely uninformative). In this section I will take it for granted that a family of good confidence procedures has been used.

4.5.1 *P-values and confidence sets*

Null Hypothesis Signficance Tests (NHST) were discussed in Sec. 3.5. In a NHST the parameter space is partitioned as

$$\Omega = \{H_0, H_1\},$$

where typically H_0 is a very small set, maybe even a singleton. We 'reject' H_0 at a significance level of α exactly when a level-$(1 - \alpha)$ confidence set $C(y^{\text{obs}}; \alpha)$ does not intersect H_0; otherwise we 'fail to reject' H_0 at a significance level of α.

In practice, then, a hypothesis test with a significance level of 5% (or any other specified value) returns one bit of information, 'reject', or 'fail to reject'. We do not know whether the decision was borderline or nearly conclusive; i.e. whether, for rejection, H_0 and $C(y^{\text{obs}}; 0.05)$ were close, or well-separated. We can increase the amount of information if C is a family of confidence procedures, in the following way.

Definition 10 (P-value, confidence set). *Let $C(\cdot\,; \alpha)$ be a family of confidence procedures. The P-value of H_0 is the smallest value α for which $C(y^{obs}; \alpha)$ does not intersect H_0.*

The picture for determining the P-value is to dial up the value of α from 0 and shrink the set $C(y^{\text{obs}}; \alpha)$, until it is just clear of H_0. Of course we do not have to do this in practice. From the Representation Theorem (Thm 4.1) we know that $C(y^{\text{obs}}; \alpha)$ is synonymous with a function $g : \mathcal{Y} \times \Omega \to \mathbb{R}$, and $C(y^{\text{obs}}; \alpha)$ does not intersect with H_0 if and only if

$$\forall \theta_j \in H_0 : g(y^{\text{obs}}, \theta_j) \leq \alpha.$$

Thus the p-value is computed as

$$p(y^{\text{obs}}; H_0) := \max_{\theta_j \in H_0} g(y^{\text{obs}}, \theta_j), \tag{4.14}$$

for a specified family of confidence procedures (represented by the choice of g). Here is an interesting and suggestive result.[16] This will be the basis for the generalisation in Sec. 4.5.2.

Theorem 4.6. *Under Def. 10 and* (4.14), *$p(Y; H_0)\big|_{\theta=\theta_j}$ is super-uniform for every $\theta_j \in H_0$.*

[16] Recollect the definition of 'super-uniform' from Def. 8.

Proof. $p(y; H_0) \leq u$ implies that $g(y, \theta_j) \leq u$ for all $\theta_j \in H_0$. Hence

$$\Pr\{p(Y; H_0) \leq u; \theta_j\} \leq \Pr\{g(Y, \theta_j) \leq u; \theta_j\} \leq u \qquad : \theta_j \in H_0$$

where the final inequality follows because $g(Y, \theta_j)\big|_{\theta=\theta_j}$ is super-uniform for all $\theta_j \in \Omega$, from Thm 4.1. □

If interest concerns H_0, then $p(y^{\text{obs}}; H_0)$ definitely returns more information than a hypothesis test at any fixed significance level, because $p(y^{\text{obs}}; H_0) \leq \alpha$ implies 'reject H_0' at significance level α, and $p(y^{\text{obs}}; H_0) > \alpha$ implies 'fail to reject H_0' at signficance level α. But a p-value of, say, 0.045 would indicate a borderline 'reject H_0' at $\alpha = 0.05$, and a p-value of 0.001 would indicate nearly conclusive 'reject H_0' at $\alpha = 0.05$. So the following conclusion is rock-solid:

- When performing a NHST, a p-value is more informative than a simple 'reject H_0' or 'fail to reject H_0' at a specified significance level (such as 0.05).

4.5.2 *The general theory of P-values*

Thm 4.6 suggests a more general definition of a p-value, which does not just apply to hypothesis tests for parametric models, but which holds much more generally, for any PMF or model for Y.

Definition 11 (Significance procedure). *Let $Y \sim f$ for specified PMF f. Then $p : \mathcal{Y} \to \mathbb{R}$ is a* significance procedure *for f exactly when $p(Y)$ is super-uniform under f; if $p(Y)$ is uniform under $Y \sim f$, then p is an* exact *significance procedure for f. The value $p(y^{obs})$ is a* significance level *or* p-value *for f exactly when p is a significance procedure for f.*

This definition can be extended to a set of PMFs for Y by requiring that p is a significance procedure for every element in the set; this is consistent with the definition of $p(y; H_0)$ in Sec. 4.5.1. The usual extension would be to take the maximum of the p-values over the set.[17]

For any specified f, there are a lot of confidence procedures: an uncountable number, actually, because *every test statistic $t : \mathcal{Y} \to \mathbb{R}$ induces a significance procedure*. For a specified t define

$$p(y; t) := \Pr\{t(Y) \geq t(y); f\}.$$

Then it follows from the *Probability Integral Transform* that $p(Y; t)$ is super-uniform under $Y \sim f$; see Casella and Berger (2002, section 8.3.4). Many of these significance procedures are useless, just like many confidence procedures are useless.

Sec. 4.5.1 made the case for reporting an NHST in terms of a p-value. But what can be said about the more general use of p-values to 'score' the model f? This is a question with a simple answer that many people wish was different:

- A p-value is not a useful guide to whether f is a good model for Y.

[17] Although Berger and Boos (1994) have an interesting suggestion for parametric models.

There is a huge literature on this topic: start at Greenland and Poole (2013) and work backwards, not neglecting Goodman (1999). There is also, unfortunately, a growing literature showing that people 'cheat' when p-values are used for NHSTs; see, e.g., Masicampo and Lalande (2012).

To put the issues in a nutshell, the f that we propose as a model for Y is an artifact: nature herself does not generate y^{obs} according to f or, really, according to any process which we can represent as a PMF or a family of PMFs. So the answer to the question "Is f the right model for Y?" is always "No", and, with a large enough sample, we would expect the p-value for $H_0 : Y \sim f$ to be very small. On the other hand, if we have a very small sample or a bad choice of t then we would expect the p-value to have a similar distribution under f and many other models a bit like f; i.e. to be uninformative about f. So what can we conclude about the 'goodness' of f as a model for Y, from a p-value? It depends on the sample size and the choice of test statistic, but in a way that is opaque to us. In particular, the idea that a single cut-off value such as $p(y^{\text{obs}}) \leq 0.05$ would serve in a wide variety of applications is hopelessly naïve.[18]

Unfortunately, the same issues apply to NHSTs, which is not surprising given the duality between confidence procedures and significance procedures shown in Thm 4.1. With a large sample and a good choice of confidence procedure we would expect to reject any H_0. But the challenge for science is never to find an effect (i.e. to reject H_0). The challenge is to attribute that effect to a specific cause in order to provide an explanation, or a defensible prediction. Statistics has a lot to say about this topic, but that would appear in a chapter on Experimental Design, not one on confidence procedures.[19]

To finish, here is a quote from Psychology:

> No aspect of classical statistics has been so popular with psychologists and other scientists as hypothesis testing, though some classical statisticians agree with us that the topic has been overemphasized. A statistician of great experience told us, "I don't know much about tests, because I have never had occasion to use one." Our devotion of most of the rest of this paper to tests would be disproportionate, if we were not writing for an audience accustomed to think of statistics largely as testing. (Edwards *et al.*, 1963, p. 213)

Note the date: it seems as though very little has changed in fifty years, in Psychology. Otherwise, there would be no need for papers such as Morey *et al.* (2015). I have never had occasion to use a hypothesis test either—at least, not since becoming a 'proper' statistician!

[18] See Cowles and Davis (1982) for the origins of the threshold 0.05.

[19] See, e.g., Cox (1958), for an excellent introduction to Experimental Design.

5 Bibliography

D. Basu, 1975. Statistical information and likelihood. *Sankhyā*, **37**(1), 1–71. With discussion. 24

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`. 24

J.O. Berger, 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, Inc., NY, USA, second edition. 42

J.O. Berger and D.D. Boos, 1994. *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, **89**, 1012–1016. 63

J. Besag, 1974. Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**(2), 192–236. 36

J. Besag, 2004. Markov Chain Monte Carlo methods for statistical inference. Available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.320.9631`. 32

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. 32

P. Billingsley, 1979. *Probability and Measure*. John Wiley & Sons, Inc., New York NY, USA, second edition. 18, 24

G. Casella and R.L. Berger, 2002. *Statistical Inference*. Pacific Grove, CA: Duxbury, 2nd edition. 58, 63

S.R. Cook, A. Gelman, and D.B. Rubin, 2006. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, **15**(3), 675–692. 33

M. Cowles and C. Davis, 1982. On the origins of the .05 level of statistical significance. *American Psychologist*, **37**(5), 553–558. 64

D. R. Cox, 1958. *Planning of Experiments*. New York: John Wiley & Sons, Inc. 64

D.R. Cox, 2006. *Principles of Statistical Inference*. Oxford University Press. 24, 58

D.R. Cox and D.V. Hinkley, 1974. *Theoretical Statistics*. London: Chapman and Hall. 54

N. Cressie and C.K. Wikle, 2011. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Inc., Hoboken NJ, USA. 34

A.P. Dawid, 2002. Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**(2), 161–190. Corrigenda vol. 70, p. 437. 21

A.P. Dawid. Beware of the DAG! In *JMLR Workshop & Conference Proceedings*, volume 6, pages 59–86, 2010. 21

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). 29

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. 66

B. de Finetti, 1972. *Probability, Induction and Statistics*. London: John Wiley & Sons. 14, 22

B. de Finetti, 1974. *Theory of Probability*, volume 1. London: Wiley. 7, 22

B. de Finetti, 1974/75. *Theory of Probability*. London: Wiley. Two volumes (2nd vol. 1975); A.F.M. Smith and A. Machi (trs.). 11

A.P. Dempster, N.M. Laird, and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–38. 36

T.J. DiCiccio and B. Efron, 1996. Bootstrap confidence intervals. *Statistical Science*, **11**(3), 189–212. with discussion and rejoinder, 212–228. 59, 60

N.R. Draper and H. Smith, 1998. *Applied Regression Analysis*. New York: John Wiley & Sons, 3rd edition. 56

W. Edwards, H. Lindman, and L.J. Savage, 1963. Bayesian statistical inference for psychological research. *Psychological Review*, **70**(3), 193–242. 64

B. Efron and C. Morris, 1977. Stein's paradox in statistics. *Scientific American*, **236**(5), 119–127. 46

D. Freedman, 1977. A remark on the difference between sampling with and without replacement. *Journal of the American Statistical Association*, **72**, 681. 27

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin, 2014. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton FL, USA, 3rd edition. Online resources at `http://www.stat.columbia.edu/~gelman/book/`. 23, 26, 37

M. Ghosh and G. Meeden, 1997. *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London, UK. 42

M. Goldstein and J.C. Rougier, 2009. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, **139**, 1221–1239. With discussion, pp. 1243–1256. 12

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. John Wiley & Sons, Chichester, UK. 11

S. Goodman, 1999. Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals of Internal Medicine*, **130**, 995–1004. 64

T. Gowers, 2002. *Mathematics: A Very Short Introduction*. Oxford University Press, Oxford, UK. 7

T. Gowers, J. Barrow-Green, and I. Leader, editors, 2008. *The Princeton Companion to Mathematics*. Princeton University Press, Princeton NJ, USA. 7

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. 64

I. Hacking, 1965. *The Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK. 24

T. Harford, 2014. Big data: Are we making a big mistake? *Financial Times Magazine*. Published online Mar 28, 2014. Available at `http://on.ft.com/P0PVBF`. 27

F. Harrell and C. Davis, 1982. A new distribution-free quantile estimator. *Biometrika*, **69**, 635–640. 60

T. Hastie, R. Tibshirani, and J. Friedman, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition. Available online at `http://statweb.stanford.edu/~tibs/ElemStatLearn/`. 12

D. Heath and W. Sudderth, 1976. De Finetti's theorem on exchangeable variables. *The American Statistician*, **30**(4), 188–189. 29

E. Hewitt and L.J. Savage, 1955. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, **80**, 470–501. 29

D. Hilbert, 1926. Über das unendliche. *Mathematische Annalen (Berlin)*, **95**, 161–190. English translation in van Heijenoort (1967). 24

J.B. Kadane, 2011. *Principles of Uncertainty*. Chapman & Hall/CRC Press, Boca Raton FL, USA. 22, 24

J.F.C. Kingman, 1978. Uses of exchangeability. *The Annals of Probability*, **6**(2), 183–197. 29

F. Lad, 1996. *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 7, 11

D.V. Lindley, 1980. L.J. Savage—his work in probability and statistics. *The Annals of Statistics*, **8**(1), 1–24. 32

D.V. Lindley, 1985. *Making Decisions*. London: John Wiley & Sons, 2nd edition. 15

D.V. Lindley and A.F.M. Smith, 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**(1), 1–41. with discussion, ??–?? 32

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. 23

D. MacKay, 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press. 25

K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. Harcourt Brace & Co., London, UK. 11, 56, 57

E.J. Masicampo and D.R. Lalande, 2012. A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*. 64

J.C. McWilliams, 2007. Irreducible imprecision in atmospheric and oceanic simulations. *Proceedings of the National Academy of Sciences*, **104**(21), 8709–8713. 37

R.D. Morey, R. Hoekstra, J.N. Rouder, M.D. Lee, and E-J Wagenmakers, 2015. The fallacy of placing confidence in confidence intervals. *Psychonomic Bullentin & Review*. Forthcoming, see `https://learnbayes.org/papers/confidenceIntervalsFallacy/` for more resources. 52, 64

K.P. Murphy, 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge MA, USA. 25, 36

J.B. Paris, 1994. *The Uncertain Reasoner's Companion: A Mathematical Perspective*. Cambridge: Cambridge University Press. 30

J. Pearl, 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. 21

D. Poole and A.E. Raftery, 2000. Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association*, **95**, 1244–1255. 24

C.P. Robert and G. Casella, 2004. *Monte Carlo Statistical Methods*. Springer, New York NY, 2nd edition. 32, 36

D.B. Rubin, 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**(4), 1151–1172. 37

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. 34

R.J. Samworth, 2012. Stein's paradox. *Eureka*, **62**, 38–41. Available online at `http://www.statslab.cam.ac.uk/~rjs57/SteinParadox.pdf`. Careful readers will spot a typo in the maths. 46

L.J. Savage *et al.*, 1962. *The Foundations of Statistical Inference*. Methuen, London, UK. 30

E. Schechter, 1997. *Handbook of Analysis and its Foundations*. Academic Press, Inc., San Diego CA, USA. 23

M.J. Schervish, 1995. *Theory of Statistics*. Springer, New York NY, USA. Corrected 2nd printing, 1997. 42

J.Q. Smith, 2010. *Bayesian Decision Analysis: Principle and Practice*. Cambridge University Press, Cambridge, UK. 25

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**(4), 583–616. With discussion, pp. 616–639. 32

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, 2014. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, **76**(3), 485–493. 32

A.W. van der Vaart, 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press. 58

B. van Fraassen, 1989. *Laws and Symmetry*. Oxford University Press. 16, 30

J. van Heijenoort, editor, 1967. *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard Univeristy Press, Cambridge MA, USA. 67

N. Ya. Vilenkin, 1995. *In Search of Infinity*. Birkhäuser Boston, Cambridge MA, USA. English translation by Abe Shenitzer. Currently available online, `http://yakovenko.files.wordpress.com/2011/11/vilenkin1.pdf`. 22

P. Whittle, 2000. *Probability via Expectation*. New York: Springer, 4th edition. 7, 11, 18

D. Williams, 1991. *Probability With Martingales*. Cambridge University Press, Cambridge, UK. 7, 18, 24

S. Wood, 2015. *Core Statistics*. Cambridge University Press, Cambridge, UK. 56