APTS Notes on Statistical Inference

Jonathan Rougier

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 $\boldsymbol{X} := (X_1, \ldots, X_m)$. The joint realm is $\mathcal{X}$ and any particular specified value is $\boldsymbol{x} := (x_1, \ldots, x_m)$. 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

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

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 $j = 1, \ldots, r$. 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 1.1 (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]

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

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

[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.

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

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

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

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

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]

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

1.1.2 *The Fundamental Theorem of Prevision*

Coherence as defined in Def. 1.1 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]

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

Some terms. A *convex combination* $(w_1, \ldots, w_r)$ 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* $\boldsymbol{X} := (X_1, \ldots, X_m)$ *be a collection of random quantities with joint realm*

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

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

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

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 $\mathfrak{X}$, 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 $(w_1, \ldots, w_r)$ is a convex combination. For arbitrary function g,

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

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 $r := |\mathfrak{X}|$. 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.

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

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.

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 $\mathbb{1}_Q \le \mathbb{1}_R$, and hence $\Pr(Q) \le \Pr(R)$.

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

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

[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.

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

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

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

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

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

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.

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

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 $g_1, \ldots, g_n$ be specified functions of $\boldsymbol{x}$, and set $Y_i := g_i(\boldsymbol{X})$ for $i = 1, \ldots, n$. 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

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 $a \cdot \mathbb{1}_{|X| \geq a} \leq |X|$ 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.

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

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

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

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

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

Definition 1.2 (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.

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

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

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

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 $\Pr(P, Q) := \Pr(P \wedge Q)$. 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*,

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

where

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

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 $\mu'_X, \mu''_X, \ldots$. 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 $\mathbb{V}\mathrm{ar}$ 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 $\mathbb{V}\mathrm{ar}$ 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 $\sigma^2_X(y)$, 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^2_X 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^2_X.

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):

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

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]

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

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*

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

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*

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

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

because $\bigcup_{\boldsymbol{y} \in \mathcal{Y}} (\boldsymbol{Y} \doteq \boldsymbol{y})$ 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 1.3 (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

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

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

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

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

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

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

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

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

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

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

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

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 $\vDash \boldsymbol{X} \mid \boldsymbol{Z}$. It is straightforward to show that

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

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 $\vDash \boldsymbol{X}$, and for which

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

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*:

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

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.

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

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

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

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

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 '$\mathrm{p}(\boldsymbol{x})$' as '$\Pr(\boldsymbol{X} \doteq \boldsymbol{x})$', because this would be vacuous.

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

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

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

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

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

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

> 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)

5
Bibliography

D. Basu, 1975. Statistical information and likelihood. *Sankhyā*, **37**(1), 1–71. With discussion. 22

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`. 22

P. Billingsley, 1979. *Probability and Measure*. John Wiley & Sons, Inc., New York NY, USA, second edition. 16, 22

D.R. Cox, 2006. *Principles of Statistical Inference*. Oxford University Press. 22, 64, 65

A.P. Dawid, 2002. Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**(2), 161–190. Corrigenda vol. 70, p. 437. 19

A.P. Dawid. Beware of the DAG! In *JMLR Workshop & Conference Proceedings*, volume 6, pages 59–86, 2010. 19

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.

B. de Finetti, 1972. *Probability, Induction and Statistics*. London: John Wiley & Sons. 12, 20

B. de Finetti, 1974. *Theory of Probability*, volume 1. London: Wiley. 5, 20

B. de Finetti, 1974/75. *Theory of Probability*. London: Wiley. Two volumes (2nd vol. 1975); A.F.M. Smith and A. Machi (trs.). 9

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. 21

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. 10

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. John Wiley & Sons, Chichester, UK. 9

T. Gowers, 2002. *Mathematics: A Very Short Introduction*. Oxford University Press, Oxford, UK. 5

T. Gowers, J. Barrow-Green, and I. Leader, editors, 2008. *The Princeton Companion to Mathematics*. Princeton University Press, Princeton NJ, USA. 5, 28

I. Hacking, 1965. *The Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK. 22

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/`. 10

D. Hilbert, 1926. Über das unendliche. *Mathematische Annalen (Berlin)*, **95**, 161–190. English translation in van Heijenoort (1967). 22

J.B. Kadane, 2011. *Principles of Uncertainty*. Chapman & Hall/CRC Press, Boca Raton FL, USA. 20, 22

K. Knopp, 1956. *Infinite Sequences and Series*. Dover Publications, Inc., New York NY, USA.

F. Lad, 1996. *Operational Subjective Statistical Methods*. New York: John Wiley & Sons. 5, 9

D.V. Lindley, 1985. *Making Decisions*. London: John Wiley & Sons, 2nd edition. 13

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. 21

K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. Harcourt Brace & Co., London, UK. 9, 60, 62

J. Pearl, 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. 19

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. 22

E. Schechter, 1997. *Handbook of Analysis and its Foundations*. Academic Press, Inc., San Diego CA, USA. 21

B. van Fraassen, 1989. *Laws and Symmetry*. Oxford University Press. 14, 29

J. van Heijenoort, editor, 1967. *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard Univeristy Press, Cambridge MA, USA. 74

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. 20

P. Whittle, 2000. *Probability via Expectation*. New York: Springer, 4th edition. 5, 9, 16

D. Williams, 1991. *Probability With Martingales*. Cambridge University Press, Cambridge, UK. 5, 16, 22