2 *Probability*

{chap:PR}

It is very common to start with probabilities, and then define expectations in terms of probabilities (see eq. 2.1 below). I have done the opposite, because I think that expectations are a better 'primitive'. I find that I can often have beliefs about a collection of random quantities $\boldsymbol{X}$ that do not involve probabilities, but which obey the axioms of Expectation.[1] It is also the case that many of the core topics in Statistics, such as Decision Theory (**??**), are naturally expressed in terms of expectations rather than probabilities.

2.1 Definition

{sec:PR-def}

If we start with expectations, then we need to define probabilities in terms of expectations. It turns out that there is no choice in how to do this, if the resulting probabilities are to obey the Laws of Probability. The nature of these Laws is explored in Sec. 2.4.

{def:lawP}

Definition 2.1 (Laws of Probability)**.**

1. For any proposition P, $\Pr(P) \geq 0$;
2. If P if certain, then $\Pr(P) = 1$;
3. If P and Q are mutually exclusive, then $\Pr(P \vee Q) = \Pr(P) + \Pr(Q)$.

Theorists have a slightly stronger requirement for (3.). As it stands, (3.) can be extended to finite disjunctions of mutually-exclusive propositions, by recursion, termed *Finite Additivity*. But theorists require a stronger property, to account for non-finite disjunctions, termed *Countable Additivity*; see Sec. 1.6. A few people get worked up about the different between these two conditions, and one, Bruno de Finetti, was famous for rejecting Countable Additivity (see, e.g. de Finetti, 1972, 1974/75). Others have risen to the challenge of working within the more general but less tractable framework of Finite Additivity (e.g., Dubins and Savage, 1965, but this is not an easy read). I doubt it matters at our level of generality, but I personally have a preference for Finite Additivity, when reasoning about the real world.

[1] JCR: A later chapter, not yet written, takes this notion much further.

Here is the definition of probability in terms of expectation, which ensures that the Laws of Probability hold. Probability is defined on the domain of random propositions. The definition makes this clear.

{def:defP}

Definition 2.2 (Probability, 'Pr'). Let $\boldsymbol{X}$ be a set of random quantities. Let $q(\boldsymbol{x})$ be any sentence from first-order logic[2], termed a *proposition*. Define $Q := q(\boldsymbol{X})$, termed a *random proposition*. Then

$$\Pr(Q) := \mathrm{E}(\mathbb{1}_Q),$$

where $\mathbb{1}$ is the indicator function.[3]

It is straightforward to check that complete coherence implies that probabilities defined in this way satisfy the Laws of Probability. (1.) follows by Lower-boundedness, because $\mathbb{1}_P \geq 0$. (2.) follows by Normalisation, because $\mathbb{1}_P = 1$ if P is certain. (3.) follows by Additivity, because if P and Q are mutually exclusive, then $\mathbb{1}_{P \vee Q} = \mathbb{1}_P + \mathbb{1}_Q$. Complete coherence is required to ensure that these pairwise properties hold for all possible propositions.

Here is the result which shows that this is the only way to define probability in terms of expectation.

{thm:defP1}

Theorem 2.1. *Suppose that* $\Pr(Q) = \mathrm{E}\{g(Q)\}$, *where*

$$g : \{\textit{FALSE}, \textit{TRUE}\} \to \mathbb{R},$$

for some choice of g. The only choice of g which is compatible with both complete coherence and the Laws of Probability is $g(Q) := \mathbb{1}_Q$.

Proof. For complete coherence, the FTP (Thm 1.1) asserts that there is a $p \in \mathcal{P}$ such that

$$\Pr(Q) = \mathrm{E}\{g(Q)\} = \sum_{\omega \in \Omega} g(q(\boldsymbol{x}(\omega))) \cdot p(\omega)$$

for every first-order sentence $q(\boldsymbol{x})$. The Laws of Probability imply that if $q(\boldsymbol{x}(\omega)) = \mathsf{FALSE}$ for all ω then $\Pr(Q) = 0$, and if $q(\boldsymbol{x}(\omega)) = \mathsf{TRUE}$ then $\Pr(Q) = 1$. Since $p(\omega) \geq 0$ and $\sum_\omega p(\omega) = 1$, it follows that $g(\mathsf{FALSE}) = 0$ and $g(\mathsf{TRUE}) = 1$, i.e. $g(Q) = \mathbb{1}_Q$, as was to be shown. □

2.2 Probability Mass Functions

{sec:PR-PMF}

The definition of probability provides a straightforward interpretation of $p \in \mathcal{P}$ from the FTP (Thm 1.1).[4]

{thm:pomega}

Theorem 2.2. *If expectations are completely coherent, then*

$$\Pr(\boldsymbol{X} \doteq \boldsymbol{x}) = \begin{cases} p(\omega^{-1}(\boldsymbol{x})) & \boldsymbol{x} \in \mathcal{X} \\ 0 & \textit{otherwise.} \end{cases}$$

[2] That is, a statement about $\boldsymbol{x}$ that evaluates to either FALSE or TRUE.

[3] That is, the function of the proposition p for which $\mathbb{1}_p = 0$ if p is FALSE, and $\mathbb{1}_p = 1$ if p is TRUE.

[4] Notation. Where an equality or an inequality is being used as a binary predicate in a first order sentence, I indicate this with a dot. This disambiguates the use of these binary predicates in infix notation. So, for example, '$x \leq y$' in free text is typically an assertion taken to be true, while '$x \mathrel{\dot{\leq}} y$' is a first order sentence which is either FALSE or TRUE.

Proof. If $x \notin \mathcal{X}$, then $\mathbb{1}_{X \doteq x} = 0$, and $\Pr(X \doteq x) = 0$ by Normalisation. In the case where $x \in \mathcal{X}$,

$$\begin{aligned} \Pr(X \doteq x) &= \textstyle\sum_\omega \mathbb{1}_{x(\omega) \doteq x} \cdot p(\omega) && \text{by the FTP} \\ &= \textstyle\sum_\omega \mathbb{1}_{\omega \doteq \omega^{-1}(x)} \cdot p(\omega) && \text{because } \omega \mapsto x \text{ is bijective} \\ &= p\big(\omega^{-1}(x)\big). && \square \end{aligned}$$

Thus, the properties of $p \in \mathcal{P}$ translate directly into probabilities for X. These probabilities are crucial in modern statistical practice, for reasons discussed at the end of Sec. 2.4. Therefore the following notation is very handy:

$$\mathrm{p}_X(x) := \Pr(X \doteq x) \qquad x \in \mathbb{R}^m.$$

This is termed the *Probability Mass Function (PMF)* of X. Often the subscript X is suppressed, when it is obvious from the argument what the underlying random quantities must be. The above PMF would usually be written '$\mathrm{p}(x)$'.

The trick with the PMF is that it holds over the whole of $\mathbb{R}^m$, which makes many operations easy to represent, notably those which simplify the realm of X. The effect on the FTP is dramatic:

$$\begin{aligned} \mathrm{E}\{g(X)\} &= \sum_{\omega \in \Omega} g(x(\omega)) \cdot p(\omega) \\ &= \sum_{x \in \mathcal{X}} g(x) \cdot p\big(\omega^{-1}(x)\big) \\ &= \sum_{x \in \mathcal{X}} g(x) \cdot \mathrm{p}_X(x) \\ &= \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_m \in \mathcal{X}_m} g(x) \cdot \mathrm{p}_X(x). \end{aligned} \tag{2.1}$$

p_X calmly assigns zero probability to those elements of the product of the realms which are not in the joint realm. Expressed in terms of PMFs, the necessary and sufficient condition for complete coherence are $\mathrm{p}_X(x) \geq 0$ and $\sum_{x \in \mathcal{A}} \mathrm{p}_X(x) = 1$ whenever $\mathcal{A}$ is a countable superset of $\mathcal{X}$.

Textbooks which treat probability as primitive must define expectation in terms of probabilities. So in these textbooks (1.4) is a definition, not a theorem. Likewise, for this next result.

{thm:MAR}

Theorem 2.3 (Marginalisation Theorem, MAR). *Let $X := (Y, Z)$ where Y and Z are themselves finite collections of random quantities. If expectations are completely coherent, then*

$$\mathrm{p}_Y(y) = \sum_{z \in \mathcal{Z}} \mathrm{p}_X(y, z), \tag{2.2}$$

where $\mathcal{Z}$ is the joint realm of Z, or any countable superset of it.

Proof. We apply the definition of a probability, but use PMFs in the

FTP:

$$\begin{aligned} \mathrm{p}_{\boldsymbol{Y}}(\boldsymbol{y}) &= \Pr(\boldsymbol{Y} \doteq \boldsymbol{y}) \\ &= \textstyle\sum_{\boldsymbol{y}'} \sum_{\boldsymbol{z}} \mathbb{1}_{\boldsymbol{y}' \doteq \boldsymbol{y}} \cdot \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{y}', \boldsymbol{z}) && \text{from (1.4)} \\ &= \textstyle\sum_{\boldsymbol{z}} \sum_{\boldsymbol{y}'} \mathbb{1}_{\boldsymbol{y}' \doteq \boldsymbol{y}} \, \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{y}', \boldsymbol{z}) \\ &= \textstyle\sum_{\boldsymbol{z}} \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{y}, \boldsymbol{z}). && \square \end{aligned}$$

Functional equalities. Eq. (2.2) is an example of a functional equality. My convention is that functional equalities represents sets of equalities, one for each element in the product of the domains of the free arguments. In other words, in (2.2), for which the free argument is $\boldsymbol{y}$, I regard it as superfluous to write 'for all $\boldsymbol{y} \in \mathbb{R}^m$'. Where the domain of a free argument needs to be constrained in order for the equality to hold, I provide the constraint after a colon; the first example of this is (2.3) immediately below.

2.3 *Inequalities*

{sec:PR-ineq}

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

$$\Pr(|X| \dot{\geq} a) \leq \frac{\mathrm{E}(|X|)}{a} \qquad : a > 0. \tag{2.3}$$

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

Markov's inequality is versatile, because it can be applied to any non-negative function of X. One application is

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

where $\mu := \mathrm{E}(X)$. As this holds for all $k > 0$ and also, trivially, for where $k = 0$,

$$\Pr(|X - \mu| \dot{\geq} a) \leq \min_{k \geq 0} \frac{\mathrm{E}(|X - \mu|^k)}{a^k} \qquad : a > 0, \tag{2.4}$$

the *centred moment bound*. 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 (2.4) is σ^2 / a^2, where $\sigma^2 := \mathrm{Var}(X)$. Another application of Markov's inequality along the same lines gives *Cantelli's inequality*

$$\Pr(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2} \qquad : a \geq 0. \tag{2.5}$$

It is good practice to prove this for oneself!

Another way to apply Markov's inequality for all X is first to transform X using a non-negative increasing function. For example,

$$\Pr(X \geq a) = \Pr(e^{kX} \geq e^{ka}) \leq \frac{\mathrm{E}(e^{kX})}{e^{ka}} = e^{-ka} M_X(k) \qquad : k > 0$$

where M_X is the MGF, eq. (1.7), of X. As this holds for all $k > 0$ and also, trivially, for $k = 0$,

$$\Pr(X \geq a) \leq \min_{t \geq 0} e^{-ta} M_X(t) \tag{2.6}$$

known as *Chernoff's inequality*.

2.4 Foundational issues

{sec:PR-dutch}

This section tackles the profound question: *Why these Laws of Probability and not some others?* The answer to this question must involve some desire on our part to adopt exactly these Laws and no others; that is, a common agreement that probabilities which obey these Laws are sensible, and probabilities which do not obey them are not-sensible. This requires us to provide a practical definition of probability which can be used to distinguish between sensible sets of probabilities and not-sensible ones. And then show that, according to this definition, the sensible sets of probabilities are exactly the ones which obey the Laws of Probability.

Another more pragmatic reason for having another look at probabilities is that it is a mouthful to tell someone that a probability for a proposition P is the expectation of the indicator function P, as stated in Def. 2.2 and justified in Thm 2.1. As will now be shown, probabilities can be much more intuitively described in terms of bets (or, more formally, betting contracts).

This strand of reasoning about probabilities goes back to Ramsey (1931)[5] and Savage (1954). The basic idea that $p := \Pr(Q)$ is an expression of my indifference between having £p with certainty, and owning a bet which pays £0 if Q is FALSE, and £1 if Q is TRUE. Call this the *betting interpretation* of probability.

Under the betting interpretation, I would pay £p to buy one unit of bet on Q, or I would accept £p to sell one unit of bet. In general I would exchange $w \cdot £p$ for an outcome of $w \cdot £\mathbb{1}_Q$, where w is the number of units, with $w > 0$ indicating buying w units of bet (paying $w \cdot £p$ to collect $w \cdot £\mathbb{1}_Q$) and $w < 0$ indicating selling w units of bet (receiving $|w| \cdot £p$ to pay out $|w| \cdot £\mathbb{1}_Q$). All together, I am prepared, notionally if not in practice, to enter into contracts of the form

$$w \cdot (\mathbb{1}_Q - p) \quad \text{for any } w \text{, negative or positive.}$$

This is always accepting that $|w|$ is not outlandishly large. There is a generalisation, which goes back to Ramsey (1931), which swaps £ for a more general preference-based currency, which can be thought of as tickets in a lottery.

At this point, just to be precise, I will call these probabilities 'betting rates'. Once we have proved that they really ought to obey the Laws of Probability, we can call them 'probabilities'.

No one would disgree that the probability of an impossible proposition is 0, and the probability of a certain proposition is 1.

[5] JCR: sort out this reference.

This is implied by the betting interpretation (under conditions to be made clear below). What the betting interpretation does is provide a way for us to attach probabilities to propositions that are neither impossible nor certain. In some situations this is straightforward. The situations of classical probability, for example, where we roll dice or toss coins, or sample randomly from a population. In this case, the betting interpretation should give the same answer as classical probabilities. But the betting interpretation extends to arbitrary propositions. For example, proposition Q might be "sea level in 2100 is at least 0.5 m higher than today". One can bet on this proposition, but not embed it in a classical situation: it represents a one-off event which, come 2100, we will know to be either false or true.

What would be a *not-sensible* situation according to the betting interpretation? It would be one where, in a set of betting rates $p_1, \dots, p_k$ on propositions $A_1, \dots, A_k$, it is possible to find a set of amounts $\boldsymbol{w} := (w_1, \dots, w_k)$ such that I can never win. More precisely, I cannot make money on any outcome, and I will lose money on at least one outcome. In the vernacular, with these betting rates I could be turned into a 'money pump'. People would bet with me for as large a $|\boldsymbol{w}|$ as I could stand, confident that they cannot lose money, and on at least one outcome they will make money. Sets of betting rates where this is possible are termed *incoherent*, otherwise they are *coherent*. It seems fundamentally irrational to have incoherent betting rates; indeed, if it were pointed out to me that my betting rates were incoherent, I would definitely want to change them. So not-sensible = incoherent, and sensible = coherent.

Definition 2.3 (Coherent betting rates)**.** Let $A_1, \dots, A_k$ be a set of propositions, with betting rates $\boldsymbol{p} := (p_1, \dots, p_k)$. These $\boldsymbol{p}$ are coherent exactly when there is no set of amounts $(w_1, \dots, w_k)$ for which the agent holding these $\boldsymbol{p}$ cannot make money on any outcome, and will lose money on at least one outcome.

Now for the exciting result. A set of betting rates is coherent *if and only if* the betting rates obey the Laws of Probability given in Def. 2.1. This result knits together the betting interpretation and the Laws of Probability; it is sometimes termed the *Dutch Book* argument, which is the name I will use below. There is a proof of this result using expectations, which I do not like; see Howson (1997) and Kadane (2011, sec. 1.7). I will provide a better proof based on a standard mathematical result, which is given in Sec. 2.8; that section also contains other material relevant to the following proof.

{thm:DBT}

Theorem 2.4 (Dutch Book Theorem)**.** *Betting rates are coherent if and only if they obey the Laws of Probability (Def. 2.1).*

Proof. Consider any two propositions which are mutually exclusive, and label them P and Q. Let $p := \Pr(P)$, $q := \Pr(Q)$, and

$r := \Pr(P \vee Q)$. Construct the outcome matrix of a set of one-unit bets, where each row is one possible outcome. There are three outcomes in total: P is true and Q is false, P is false and Q is true, or P is false and Q is false. Each column is the pay-off for one unit on one of the three bets, on P, on Q, and on $P \vee Q$. Thus

$$M := \begin{array}{r} \\ P \wedge \neg Q \\ \neg P \wedge Q \\ \neg P \wedge \neg Q \end{array} \begin{array}{c} \begin{array}{ccc} \quad P \quad & \quad Q \quad & P \vee Q \end{array} \\ \begin{pmatrix} 1-p & -q & 1-r \\ -p & 1-q & 1-r \\ -p & -q & -r \end{pmatrix} \end{array}.$$

The outcomes for a set of amounts $\boldsymbol{w} := (w_1, w_2, w_3)$ is $M\boldsymbol{w}$.

Consider Stiemke's Theorem (Thm 2.11), with $A \leftarrow -M$ and $x \leftarrow \boldsymbol{w}$. The first alternative now reads $M\boldsymbol{w} < \mathbf{0}$. This is the definition of incoherence. Therefore coherence is equivalent to the second alternative, which now reads "there exists a $\boldsymbol{y} \gg \mathbf{0}$ for which $\boldsymbol{y}^T M = \mathbf{0}^T$", or, more conveniently, "for which $M^T \boldsymbol{y} = \mathbf{0}$." Now $M^T \boldsymbol{y} = \mathbf{0}$ always has at least one solution, so the set of solutions is non-empty; denote it $\mathcal{S}$. We have to show

1. If $\boldsymbol{y} \in \mathcal{S}$ and $\boldsymbol{y} \gg \mathbf{0}$, then LP; and
2. If LP, then $\boldsymbol{y} \in \mathcal{S}$ and $\boldsymbol{y} \gg \mathbf{0}$,

where 'LP' denotes the Laws of Probability.[6] In this set-up, LP represents $0 < p$, $0 < q$, $p + q = r$, and if P is certain then $p = 1$ (and $q = 0$, $r = 1$). Both p and q are strictly positive if P and Q are not impossible (see below).

Let $s := y_1 + y_2 + y_3$. Multiply out $M^T \boldsymbol{y} = \mathbf{0}$ to derive the three equations

$$\begin{aligned} y_1 - p \cdot s &= 0 \\ y_2 - q \cdot s &= 0 \\ y_1 + y_2 - r \cdot s &= 0, \end{aligned}$$

which must be satisfied by all $\boldsymbol{y} \in \mathcal{S}$. To prove (1.), $\boldsymbol{y} \gg \mathbf{0}$ implies that $s > 0$. We infer immediately that $p > 0$, $q > 0$, and $p + q = r$. Now suppose that $\neg P$ is impossible, so that the second and third rows of M disappear. Equivalently, enforce $y_2 = y_3 = 0$. It follows immediately that $p = 1$ (and $q = 0$, $r = 1$), as required.

Proof of (2.), the converse, using the contrapositive. Suppose that $\boldsymbol{y} \in \mathcal{S}$ but $\boldsymbol{y} = \mathbf{0}$, in which case p, q, r are arbitrary, since they satisfy $0 - \square \cdot 0 = 0$, violating LP. □

Now I can refer to betting rates as 'probabilities'.

The proof of Thm 2.4 is quite clear about the equivalence of 'Q is impossible' and $\Pr(Q) = 0$. 'Impossible' means 'logically impossible', not merely 'almost inconceivable'. Impossible outcomes get removed from M, but almost inconceivable ones do not, because one can still lose money if an almost inconceivable outcome occurs. Thus not-impossible outcomes have positive probabilities under coherence, even though they may be tiny. It is a mistake to think

[6] I.e. show in the first case that coherence implies LP, and in the second case that LP implies coherence.

that tiny probabilities can be set to zero. Interesting propositions can be constructed as disjunctions of billions of mutually exclusive atomic propositions (see below). If all tiny probabilities were set to zero, then we could end up with the probability of the certain event being less than 1, and that would be incoherent. Dennis Lindley (1985) made this into a Principle.

Definition 2.4 (Cromwell's Rule)**.** Reserve $\Pr(Q) = 0$ for cases where Q is logically impossible.

* * *

There are a huge number of additional relations that are implied by the Laws of Probability; this is the topic of Probability Theory. If $A_1, \dots, A_k$ were a rich set of propositions, then it would be almost impossible for me to specify coherent probabilities for all of the propositions that could be constructed from $A_1, \dots A_k$. This is why, in practice, it is better to build probabilities by applying the Laws of Probability, achieving probabilities for complicated propositions by combining simpler ones.

The most primitive strategy for doing this is to break all of the propositions down into a set of mutually exclusive and exhaustive 'atoms', so that every proposition can be expressed as a disjunction of atoms. For a finite set of propositions, this takes the form of expanding out the tautology[7]

$$\mathsf{TRUE} = (A_1 \vee \neg A_1) \wedge \cdots \wedge (A_k \vee \neg A_k) = \bigvee_{j=1}^{2^k} A^{(j)}$$

where each atom $A^{(j)}$ has the form $(\tilde{A}_1 \wedge \cdots \wedge \tilde{A}_k)$, where $\tilde{A}_i$ is either A_i or $\neg A_i$. Many of these atoms will be impossible and have zero probabilities. For example, if A_i implies A_j, then all atoms with A_i and $\neg A_j$ in them will have zero probabilities. The rest must have positive probabilities which sum to 1.

This comment is not as abstract as it seems. In Statistics, when the propositions concern random quantities, the atoms are associated with the elements of the joint realm of $\boldsymbol{X}$ represented by the set Ω. We have

$$\mathsf{TRUE} = \bigvee_{\omega \in \Omega} \big(\boldsymbol{X} \doteq \boldsymbol{x}(\omega)\big).$$

The probabilities on the atoms are represented by the function $p \in \mathcal{P}$, according to Thm 2.2. According to Thm 2.4, the two conditions $p(\omega) \geq 0$ and $\sum_\omega p(\omega) = 1$ are necessary and sufficient for coherence. When theorists write "Let Ω be a set, let $\mathcal{F}$ be a σ-algebra over Ω, and let p be a non-negative, finite, σ-additive measure on $\mathcal{F}$, normalised so that $p(\Omega) = 1$" they are doing exactly this, but using concepts that allow generalisation to non-countable Ω, for which the notion of an atom is more tricky.

[7] Remember the distributive rule that $A \wedge (B \vee C) \Leftrightarrow (A \wedge B) \vee (A \wedge C)$.

2.5 Conditional probabilities

{sec:PR-CP}

The stunning result of the Dutch Book Theorem prompts us to go further, and consider conditional probabilities. We need to find a betting interpretation of the conditional betting rate for 'P given Q', and then verify that betting rates are coherent if and only if

$$\Pr(P, Q) = \Pr(P \mid Q) \cdot \Pr(Q) \tag{2.7}$$

which is accepted as the defining property of '$\Pr(P \mid Q)$'.[8] Note the convention in Probability and Statistics of writing a comma in place of the conjunction '$\wedge$', i.e.

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

Some authors write that $\Pr(P \mid Q)$ is undefined when $\Pr(Q) = 0$; this is a mistake. In this case (2.7) has form $0 = \Pr(P \mid Q) \cdot 0$, and hence $\Pr(P \mid Q)$ is arbitrary, not undefined.

The interpretation that works is a 'called-off bet'. Asserting $r = \Pr(P \mid Q)$ is an expression of my indifference between having r with certainty, and owning a bet which pays

$$\mathbb{1}_Q \cdot \mathbb{1}_P + (1 - \mathbb{1}_Q) \cdot r.$$

In this bet I get $\mathbb{1}_P$ if Q is true, and my money back if Q is false. Thus the bet is 'called off' if Q is false. All together, I am prepared to enter into contracts of the form

$$w \cdot \mathbb{1}_Q(\mathbb{1}_P - r) \quad \text{for any } w \text{, positive or negative.}$$

{thm:CDBT}

Theorem 2.5 (Conditional Dutch Book Theorem)**.** *Let P and Q be any two propositions. Then the conditional betting rate $\Pr(P \mid Q)$ is coherent if and only if $\Pr(P, Q) = \Pr(P \mid Q) \cdot \Pr(Q)$.*

Proof. It's the same proof as Thm 2.4. Let P and Q be arbitrary propositions for which all four outcomes concerning P and Q are possible. Let $p := \Pr(P, Q)$, $q := \Pr(Q)$, and $r := \Pr(P \mid Q)$. The outcome matrix is

$$M := \begin{array}{r} \\ \neg P, \neg Q \\ P, \neg Q \\ \neg P,\ Q \\ P,\ Q \end{array} \begin{array}{c} \begin{array}{ccc} P \wedge Q & Q & P \mid Q \end{array} \\ \begin{pmatrix} -p & -q & 0 \\ -p & -q & 0 \\ -p & 1-q & -r \\ 1-p & 1-q & 1-r \end{pmatrix} \end{array}.$$

Letting $\mathcal{S}$ denote the solutions to $M^T \boldsymbol{y} = \mathbf{0}$, coherence is equivalent to $\boldsymbol{y} \in \mathcal{S}$ and $\boldsymbol{y} \gg \mathbf{0}$. We have to show

1. If $\boldsymbol{y} \in \mathcal{S}$ and $\boldsymbol{y} \gg \mathbf{0}$, then $p = r \cdot q$; and

2. If $p = r \cdot q$, then $\boldsymbol{y} \in \mathcal{S}$ and $\boldsymbol{y} \gg \mathbf{0}$.

We must also check that we do not violate LP.

[8] The origin of this property is explored in **??**.

Let $s := y_1 + y_2 + y_3 + y_4$. Multiply out $M^T \boldsymbol{y} = \boldsymbol{0}$ to give the three equations

$$\begin{aligned} y_4 - p \cdot s &= 0 \\ y_3 + y_4 - q \cdot s &= 0 \\ y_4 - (y_3 + y_4) \cdot r &= 0, \end{aligned}$$

which must be satisifed by all $\boldsymbol{y} \in \mathcal{S}$. To prove (1.), $\boldsymbol{y} \gg \boldsymbol{0}$ implies $s > 0$, and it is straightforward to check that $0 < p < q < 1$ and $p = r \cdot q$. To prove (2.), use the same contrapositive as before. □

There are obvious adjustments of this general result for special cases. For example, if Q implies P then $0 < p = q < r = 1$, and if Q implies $\neg P$ then $0 = p = r < q < 1$.

2.6 *Further thoughts on subjective probabilities*

c:PR-further}

Here are some more general comments, which apply as much to expectations as they do to probabilities.

First, how I or anyone else produces a value $\Pr(Q)$ is mysterious. Through my life I have been exposed to information which may be relevant to the truth of Q; some of this information I have remembered more-or-less intact, other information has done no more than leave a vague impression. I may go and seek out new information. In the end, I reach for a probability that 'seems right' to me, and I test out my probability on myself by asking whether I would be willing to buy or sell a bet at price £p. The Laws of Probability say no more than $\Pr(Q) > 0$ if Q is not impossible, and $\Pr(Q) < 1$ if Q is not logically certain. If I have a second proposition R, and Q and R happen to be mutually exclusive, then the Laws have something further to say. If it turns out that my probabilities are incoherent, the Laws do not tell me how to modify them. This is down to me.

On this basis, the impression that we often agree, approximately, about probabilities deserves some thought. Likewise the related impression that we are often willing to accept someone else's probabilities as our own. In fact this latter impression is not so hard to understand. There are some domains, future weather for example, where some people have hard-earned expertise. A meteorologist knows a lot more about future weather than I do, and it would be sensible of me to accept a meteorologist's probabilities as my own, once I have satisfied myself that her probabilities are coherent.[9] I am not accepting her probabilities because they are 'right', a concept which makes no sense. I am accepting them, and sometimes paying for them, because I believe that my decisions made on the basis of her probabilities about future weather will work out better than decisions made on the basis of my own probabilities.

But what to make of the impression that we often agree, approximately, about probabilities? The simplest explanation is that we humans tend to think alike, and, in many cases where we agree,

[9] This is the practical definition of an expert: 'someone whose probabilities you accept as your own'.

it is because we have been exposed to similar models and similar evidence.

Here is a cute result on this topic—I'm not claiming much more for it than this! Suppose there is a sequence of experimental outcomes, $E_1, E_2, \dots$, all of which are implied by a scientific model M. Represent this as

$$\Pr(E_{\mathcal{A}} \mid M) = 1 \quad \text{for all } \mathcal{A},$$

where $E_{\mathcal{A}}$ denotes the conjunction of any subset $\mathcal{A}$ of the experimental outcomes.[10] Then we have the following remarkable result, termed the *First Induction Theorem* by Good (1975), and originally proved by Wrinch and Jeffreys (1921).

{thm:1IT}

Theorem 2.6 (First Induction Theorem). *Let* $\Pr(E_{\mathcal{A}} \mid M) = 1$ *for all* $\mathcal{A}$. *If* $\Pr(M) > 0$, *then*

$$\lim_{n \to \infty} \Pr(E_n \mid E_1, \dots, E_{n-1}) = 1.$$

Proof. Under the conditions of the theorem,

$$\begin{aligned} \Pr(E_{\mathcal{A}}) &= \Pr(E_{\mathcal{A}} \mid M) \Pr(M) + \Pr(E_{\mathcal{A}} \mid \neg M) \Pr(\neg M) \\ &\geq \Pr(E_{\mathcal{A}} \mid M) \Pr(M) \\ &= \Pr(M) \end{aligned}$$

for al $\mathcal{A}$. Now let $\mathcal{A} \leftarrow \{1, \dots, n\}$ and write the lefthand side as

$$p_n := \Pr(E_1, \dots, E_n) = \Pr(E_1) \prod_{i=2}^{n} \Pr(E_i \mid E_1, \dots, E_{i-1}).$$

$p_1, p_2, \dots$ is a monotone non-increasing sequence bounded below by $\Pr(M)$. Since $\Pr(M) > 0$ it converges to a positive limit, in which case $\Pr(E_n \mid E_1, \dots, E_{n-1})$ converges to 1. □

The remarkable thing about this result is that the displayed equation in Thm 2.6 makes no reference to model M at all. It indicates that anyone who believes that M implies the E's and that M is not logically impossible is bound, sooner or later, on the accumulation of enough evidence, to act as though M is true, in terms of their probabilities for other implications of M. Relaxing the conditions of the result, to allow for 'fuzziness' in the definition of M and in the nature of the evidence, we can still infer that probabilities will tend to be similar, because we will be channeled by exposure to similar evidence into probabilistically similar models for the world.

2.7 *Uncountable realms*

{sec:PR-uncountable}

The case of countable realms was discussed in Sec. 1.6. Here I consider the case of uncountable realms. The crucial result, from Analysis, is as follows.

{thm:count}

Theorem 2.7. *Let* Ω *be a set, not necessarily finite or countable, and define* $\sum_{\omega \in \Omega} a(\omega)$ *to be the supremum of* $\sum_{\omega \in W} a(\omega)$ *over all finite* $W \subset \Omega$. *If* $a(\omega) \geq 0$ *and* $\sum_{\omega \in \Omega} a(\omega) < \infty$, *then at most countably many of the* a*'s are non-zero.*

[10] I.e., $E_{\mathcal{A}} := \wedge_{i \in \mathcal{A}} E_i$ where $\mathcal{A} \subset \mathbb{N}$.

Proof. This proof is sketched in Schechter (1997, sec. 10.40). The definition of the sum over Ω generalises the case where Ω is finite or countable. Accepting the conditions of the theorem, let $\sum_{\omega \in \Omega} a(\omega) = 1$, without loss of generality. Then $\Omega_m := \{\omega \mid a(\omega) > 1/m\}$ must be finite, indeed $|\Omega_m| < m$. But $\{\Omega_m\}$ is an non-decreasing sequence and

$$\sum_{\omega \in \Omega} a(\omega) = \lim_{m \to \infty} \sum_{\omega \in \Omega_m} a(\omega)$$

from which the result follows. □

The Laws of Probability (Def. 2.1) imply that $p(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$. Thm 2.7 therefore asserts that at most a countable subset of Ω has non-zero $p(\omega)$. In accordance with the proof, denote this subset as Ω_∞. This subset is most clearly visualised in terms of the *distribution function* of $\boldsymbol{X} := (X_1, \dots, X_m)$, denoted

$$F_{\boldsymbol{X}}(\boldsymbol{x}) := \Pr(\boldsymbol{X} \mathrel{\dot{\leq}} \boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^m. \tag{2.8}$$

This has the obvious properties of $F_{\boldsymbol{X}}(-\infty \cdot \mathbf{1}) = 0$, $F_{\boldsymbol{X}}(\infty \cdot \mathbf{1}) = 1$, and $F_{\boldsymbol{X}}$ non-decreasing. According to Thm 2.7, additionally $F_{\boldsymbol{X}}$ is constant almost everywhere, except for right-continuous discontinuities at each $\boldsymbol{x} \in \Omega_\infty$.

Perhaps it is not surprising that statisticians make very little effort to construct $F_{\boldsymbol{X}}$ correctly according to Thm 2.7, which would involve specifying a Ω_∞, and $p(\omega)$ for every $\omega \in \Omega_\infty$. Instead, the ubiquitous practice when Ω is uncountable is to use a smooth
{eq:PDF0} approximation,

$$F_{\boldsymbol{X}}(\boldsymbol{x}) \approx \tilde{F}_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{y}) \, \mathbf{d}\boldsymbol{y} \tag{2.9a}$$

for some specified $\mathrm{p}_{\boldsymbol{X}}$ which is piecewise continuous[11] and satisfies

$$\mathrm{p}_{\boldsymbol{X}}(\boldsymbol{y}) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{y}) \, \mathbf{d}\boldsymbol{y} = 1, \tag{2.9b}$$

where $\int \square \, \mathbf{d}\boldsymbol{y}$ denotes the usual m-dimensional Riemann integral. In this approximation $\mathrm{p}_{\boldsymbol{X}}$ is termed the *probability density function (PDF)* of $\boldsymbol{X}$. Unfortunately, the ubiquitous practice in Statistics is to use the same symbol $\mathrm{p}_{\boldsymbol{X}}$ for both PMFs and PDFs, although they do not even have the same units.

Any $\boldsymbol{X}$ whose distribution function can be represented by (2.9) is termed *continuous*. Otherwise, $\boldsymbol{X}$ is *discrete*; any $\boldsymbol{X}$ for which Ω is finite or countable is discrete.

In one sense (2.9) is a horrible approximation. It asserts that $\Pr(\boldsymbol{X} \mathrel{\dot{=}} \boldsymbol{x}) = 0$ for every $\boldsymbol{x} \in \mathbb{R}^m$. And yet, were we to measure $\boldsymbol{X}$, the result would be a value $\boldsymbol{X} = \boldsymbol{x}^{\text{obs}} \in \mathbb{R}^m$, which would seem to be an impossible outcome according to these probabilities. The sophisticated answer is that what we actually measure is a value for $\boldsymbol{X}$ in the tiny region near to $\boldsymbol{x}^{\text{obs}}$, for which (2.9) asserts

$$\Pr(\boldsymbol{x} \mathrel{\dot{<}} \boldsymbol{X} \mathrel{\dot{\leq}} \boldsymbol{x} + \mathbf{d}\boldsymbol{x}) = \mathrm{p}_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathbf{d}\boldsymbol{x}. \tag{2.10}$$

[11] I.e., made up of a finite number of continuous pieces.

This interpretation underpins the following useful principal for doing mathematics with continuous random quantities:

> *Treat all random quantities as discrete, but write the PMF of continuous random quantities as* $\mathrm{p}_{\boldsymbol{X}}(\boldsymbol{x})\,\boldsymbol{dx}$*. Marginalise over continuous random quantities using a Riemann integral.*

In this principal, the symbol $\mathbf{d}x$ is standing in for the precision of the measurement, and therefore it represents a positive real number. So it can be treated as such, for example when cancelling.

There are, of course, much more elegant treatments of continuous random quantities within Measure Theory. They are internally consistent, in a way that the above principal is not. What Measure Theory cannot do, though, is fix the basic problem that no operationally-defined quantity in nature has a continuous distribution.

2.8 Convexity and Stiemke's Theorem

{sec:PR-convex}

This is a self-contained derivation of the Supporting Hyperplane Theorem and Stiemke's Theorem; see Çınlar and Vanderbei (2013) for a brief review of closed sets and convexity, and Rockafellar (1970) for more details. All lower-case symbols are vectors, except for d, and so I have not bothered to distinguish between vectors and scalars using bold, as I did for the previous sections of this chapter; I do, however, write 0 for scalar zero and $\mathbf{0}$ for vector zero.

Let $\mathbb{R}^n$ be n-dimensional Euclidean space, and let $C \subset \mathbb{R}^n$. Recall that C is open exactly when for every $c \in C$ there is an $r > 0$ for which $B_r(c) \subset C$.[12] A set is closed exactly when its complement is open. The closure of C, denoted $\operatorname{cl} C$, is the smallest closed set containing C. The interior of C, denoted $\operatorname{int} C$, is the largest open set contained in C. The boundary of C is $\partial C := \operatorname{cl} C \setminus \operatorname{int} C$. The boundary of a closed set lies in the set.

The set C is convex exactly when the chord between any two elements of C lies entirely within C. Formally, C is convex exactly when $\lambda x + (1-\lambda) y \in C$ whenever $x, y \in C$ and $0 < \lambda < 1$.

Let C be a non-empty closed convex set and choose a $y \notin C$. It is almost obvious that the projection of y onto C, denoted $\bar{y}$, is unique. To find this projection, construct a closed ball $\bar{B}_r(y)$, and gradually increase r from 0 until $\bar{B}_r(y)$ and C touch; they first touch at $\bar{y} \in \partial C \subset S$, and $\bar{y}$ must be unique because both the ball and C are closed and convex. For every point $c \in C$, the line from y to $\bar{y}$ to c forms an non-acute angle. Therefore the hyperplane normal to $y - \bar{y}$ which contains $\bar{y}$ separates y from C. Call this a *tangential hyperplane* at $\bar{y}$. Its expression is $(y-\bar{y})^T(x-\bar{y}) = 0$, or $a^T x = d$, for scalar d. Thus we have:

{thm:sepHT}

Theorem 2.8 (Mini Separating Hyperplane Theorem)**.** *Let $C \subset \mathbb{R}^n$ be a non-empty closed convex set. if $y \notin C$ then there exists $a \in \mathbb{R}^n$ and $d \in \mathbb{R}$ such that $a^T y < d$ and $a^T c \geq d$ for all $c \in C$.*

[12] $B_r(c)$ is the (open) ball of radius r centred at c, defined as

$$B_r(c) := \{x \mid \|c - x\| < r\}.$$

Below, the closed ball is defined as

$$\bar{B}_r(c) := \{x \mid \|c - x\| \leq r\}.$$

The projection mapping from the exterior of C to ∂C is surjective; that is, every point on ∂C is the projection of at least one point in the exterior of C, and has at least one tangential hyperplane. This gives the Supporting Hyperplane Theorem.

{thm:supHT}

Theorem 2.9 (Supporting Hyperplane Theorem, SHT)**.** *Let $C \subset \mathbb{R}^n$ be a non-empty closed convex set and let c_0 be any element in the boundary of C. Then there exists an $a \in \mathbb{R}^n$, $a \neq \mathbf{0}$, such that $a^T c \geq a^T c_0$ for all $c \in C$.*

Before proving the next result, a brief aside on vector inequalities:

1. $x \geq \mathbf{0}$ exactly when $x_i \geq 0$ for all i.

2. $x > \mathbf{0}$ exactly when $x \geq 0$ and $x \neq \mathbf{0}$.

3. $x \gg \mathbf{0}$ exactly when $x_i > 0$ for all i.

Conventions differ concerning the interpretation of $x > \mathbf{0}$, so it is best to check.[13]

Now consider the set

$$C := \{ y \mid y = Ax, x \geq \mathbf{0} \},$$

where A is an $m \times n$ matrix. It is straightforward to check that this set is closed and convex; in fact, it is termed a *convex cone*. For any point $y \in \mathbb{R}^m$, there are only two possibilities; either $y \in C$ or $y \notin C$. In the first case, there exists $x \geq \mathbf{0}$ such that $y = Ax$. In the second case, the Mini Separating Hyperplane Theorem (Thm 2.8) asserts the existence of a and d such that $a^T y < d$ and $a^T c \geq d$ for all $c \in C$. So in this case

$$a^T y < d \leq a^T A x \quad \text{for all } x \geq \mathbf{0}. \tag{†}$$

Since $x = \mathbf{0}$ is possible, $d \leq 0$, and hence $a^T y < 0$. If any component of $a^T A$ was negative, then $a^T A x$ could be made arbitrarily small, and hence, necessarily, $a^T A \geq \mathbf{0}$. This result is Farkas's Lemma. It is generally expressed as follows.[14]

Theorem 2.10 (Farkas's Lemma)**.** *Let A be an $m \times n$ real-valued matrix. Then exactly one of the two alternatives is possible:*

1. *$Ax = b$ has a solution $x \geq \mathbf{0}$; or*

2. *$y^T b < 0$ and $y^T A \geq \mathbf{0}^T$ has a solution $y \in \mathbb{R}^m$.*

There are lots of variants of Farkas's Lemma; the one I used in the proof of Thm 2.4 was Stiemke's Theorem, and another closely-related result is Gordan's Theorem. They can both be proved as follows.

For arbitrary $x \in \mathbb{R}^n$, write

$$x = x^+ - x^-,$$

[13] The main alternative convention is to use $\geq$ and $>$ for element-wise inequalities, and $x \gneq \mathbf{0}$ for the case '$x \geq 0$ and $x \neq \mathbf{0}$'. In other words, to use $\geq$, $\gneq$, and $>$ where I use $\geq$, $>$, and $\gg$.

[14] With $a \leftarrow y$ and $y \leftarrow b$.

where x^+ and x^- were defined in (1.9); in the vector case, the maximums are taken element-wise. Then

$$Ax = A(x^+ - x^-) = \begin{bmatrix} A & -A \end{bmatrix} \begin{bmatrix} x^+ \\ x^- \end{bmatrix}$$

where the vector is now non-negative. Applying Farkas's Lemma then gives the two alternatives:

1. $Ax = b$ for some $x \in \mathbb{R}^n$; or

2. $y^T b < 0$ and $y^T A = \mathbf{0}^T$ for some $y \in \mathbb{R}^m$.

The second part of the second condition is the unique solution to $y^T[A, -A] \geq \mathbf{0}^T$. For Stiemke's Theorem, let b be some value satisfying $b < \mathbf{0}$. If there is no such value for which the first alternative is satisfied, then $y^T b < 0$ for every $b < \mathbf{0}$, and hence $y \gg \mathbf{0}$. By multiplying A by -1 we get the usual expression of Stiemke's Theorem.

{thm:stiemke}

Theorem 2.11 (Stiemke's Theorem). *Let A be an $m \times n$ matrix of reals. Then exactly one of these two alternatives is true:*

1. *$Ax > \mathbf{0}$ for some $x \in \mathbb{R}^n$; or*

2. *$y^T A = \mathbf{0}$ for some $y \gg \mathbf{0}$.*

Gordan's Theorem is the same, except with $b \ll \mathbf{0}$, in which case $Ax \gg \mathbf{0}$ and $y > \mathbf{0}$.