# 1
# *Expectation as primitive*



## *1.0 Some preliminary observations*

Currently, the dominant approach to statistical inference is to start with probability as the primitive notion, implemented, in its most abstract form, in terms of a 'probability triple': an atomic set usually denoted $\Omega$, a 'sigma field' of sets of $\Omega$, often denoted $\mathcal{F}$, and a measure on $\mathcal{F}$, usually denoted $\mathbb{P}$:

$$\text{Probability triple} = \{\Omega, \mathcal{F}, \mathbb{P}\};$$

see, for example, the opening chapters of Grimmett and Stirzaker (2001) or Rosenthal (2006). This starting-point enables probability theory to bed itself in the rich soil of measure theory. Treatments of probability from this starting-point are Kingman and Taylor (1966), Billingsley (1979), or Williams (1991). A comprehensive treatment of statistical theory from this starting-point is Schervish (1995).

But for me, this starting-point interposes too much distance between the practice of reasoning about uncertainty, and formalising that reasoning in a useful calculus of uncertainty. Let us accept, as a fact established beyond doubt, that humans are poor at reasoning about uncertainty. In fact we can take this as a corollary of the more general fact that humans are bad at reasoning, as discussed, for example, in Kahneman (2011). And let us accept also that mathematics can help in this respect, by automating features of reasoning that we recognise as self-evident (see, e.g., Ellenberg, 2014). And finally, let us appreciate that statistical inference is not performed in a vacuum, but in the context of people needing to make choices under uncertainty, as captured by our mission statement, *Helping people to make better choices under uncertainty*. How does this shape a useful calculus of uncertainty?

I find the following narrative helpful, based on the discussion in Smith (2010, ch. 1). Some stakeholders with a common interest get together and hire a risk manager to make decisions on their behalf. She in turn hires a statistician to help her in reasoning under uncertainty, as well as other domain experts (e.g. engineers). She is answerable to an auditor, who is also hired by the stakeholders. Effectively, she needs to be able to convince the auditor that the choices she makes are in the best interests of the stakeholders. In a

nutshell, her choices must be *transparent and defensible*. I think these should be the hallmarks of a good statistical inference, and the yardstick by which the statistician's contribution is judged. While in many applications one or more of these players is absent, and the statistician may be his own client, I think the same standards should prevail. And this leads me to the conclusion that uncertain quantities and our beliefs about them should be the primitives of an uncertainty calculus, if it aspires to being more than a mathematical edifice.

This is quite contrary to the starting-point of probability as the primitive notion, for two reasons. First, if probability is primitive then a random quantity is a derived concept, constructed on top of a specified probability triple.[1] Second, 'probability as primitive' presupposes too much: that our beliefs about uncertain quantities extend all the way to specifying probability distributions. In practice, we humans struggle to quantify our beliefs about interesting and decision-relevant quantities, and we may find that we have only a very limited appreciation of our uncertainty about, say, sea-level rise by 2100. Neither of these objections applies if we take uncertain quantities as primitive, and 'expectation' as the elementary quantification of belief.

[1] It is a specified measurable function of $\Omega$.

This chapter and the two that follow expound this approach, of taking random quantities and their expectations as the starting-point of a calculus of uncertainty: 'expectation as primitive'. This chapter defines expectation (and probability) and identifies some of its properties. Chapter 2 considers the task of statistical inference, and the role of data. Chapter 3 defines conditional expectation (and conditional probability) and explores its relationship with 'unconditional' expectation. All of the standard results of the probability calculus are recovered, although I contend that they acquire meaning through the interpretation of expectation and conditional expectation as extra-mathematical constructs.[2]

[2] As I explain in Sec. 3.1, I prefer 'hypothetical' to 'conditional'.

Although these three chapters represent my own construction of expectation and statistical inference, they are heavily influenced by much deeper thinkers than myself. The most famous and influential proponent of 'expectation as primitive' was Bruno de Finetti (1937, 1972, 1974/75), whose work strongly influenced Lad (1996) and Goldstein and Wooff (2007). In a related strand, Walley (1991) provides a detailed assessment of how we might reason with only a limited set of beliefs; Troffaes and de Cooman (2014) is a recent contribution to the same question. The celebrated book of Whittle (2000) is also relevent. This might be thought of as a probabilist's interpretation of 'expectation as primitive', ignoring vexed questions of interpretation, but full of elegant and useful results.

## *1.1 Random quantities and their realms*

My starting-point is a *random quantity*. A random quantity is a set of instructions which, if followed, will yield a real value; this is

an *operational definition*. Experience suggests that thinking about random quantities is already hard enough, without having to factor in ambiguitities of definition—hence my insistence on operational definitions. Real-valued functions of random quantities are also random quantities.

It is conventional in statistics to represent random quantities using capital letters from the end of the alphabet, such as $X$, $Y$, and $Z$, and, where more quantities are required, using ornaments such as subscripts and primes (e.g. $X_i$, $Y'$). Functions of random quantities are expressed directly. Thus $XY$ represents the random quantity that arises when the instructions $X$ and $Y$ are both performed, and the resulting two values are multiplied together. Representative values of random quantities are denoted with small letters. I will write '$X \to x$' to represent 'instructions $X$ were performed and the value $x$ was the result'.

The *realm* of a random quantity is the set of possible values it might take; this is implicit in the instructions. I denote this with a curly capital letter, such as $\mathcal{X}$ for the realm of $X$, where $\mathcal{X}$ is always a subset of $\mathbb{R}$.[3] I write a collection of random quantities as $\boldsymbol{X} := (X_1, \ldots, X_m)$, and their joint realm as $\boldsymbol{\mathcal{X}}$, where $\boldsymbol{x} := (x_1, \ldots, x_m)$ is an element of $\boldsymbol{\mathcal{X}}$, and

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

[3] I have taken the word 'realm' from Lad (1996); 'range' is also used, although this might better be reserved for the lower and upper limits of the realm.

Figure 1.1 shows some examples of realms for $\boldsymbol{X} \leftarrow (X_1, X_2)$. A random quantity in which the realm contains only a single element is a *constant*, and typically denoted by a small letter from the start of the alphabet, such as $a$, $b$, or $c$.

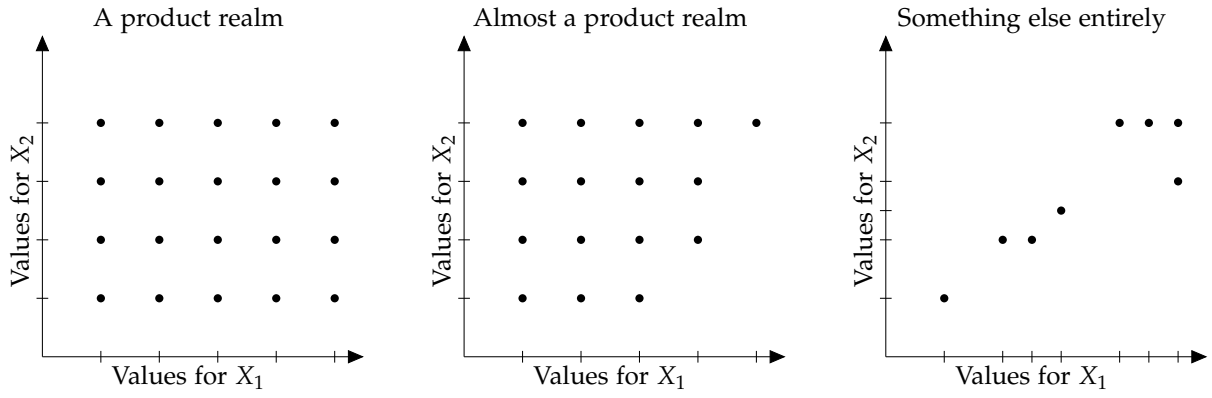


Figure 1.1: Some examples of the realm of $\boldsymbol{X} \leftarrow (X_1, X_2)$. In each case the realm of $X_1$ is the union of the projection of the points onto the horizontal axis. Similarly for $X_2$ (vertical axis).

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

> The author holds firmly to the view that this contingent and cognitive universe of ours is in reality only finite and, therefore, discrete. In this essay we steer clear of the logical quick sands of 'infinity' and

> the 'infinitesimal'. Infinite and continuous models will be used in the sequel, but they are to be looked upon as mere approximations to the finite realities. (Basu, 1975, footnote, p. 4)

For similar sentiments, see, e.g., Hacking (1965, ch. 5), Berger and Wolpert (1984, sec. 3.4), Cox (2006, sec. 1.6), or Aitkin (2010, ch. 1). This is not just statistical parochialism. David Hilbert, one of the great mathematicians and a huge admirer of Cantor's work on non-finite sets, stated

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

and later in the same essay,

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

The complications and paradoxes of the infinite are well-summarised in Vilenkin (1995).[4] I reckon the task of the statistician is hard enough, without having to grapple with an abstraction which has so consistently perplexed and bamboozled.

[4] Wallace (2003) is also worth a look. David Foster Wallace was a tremendous writer of fiction and essays, but this book displays the limitations of his literary style when writing about highly technical matters—also one has to acknowledge that he did not have sufficient mastery of his material.

However, as Kadane (2011, ch. 3) discusses, it is convenient to be able to work with non-finite and unbounded realms, to avoid the need to make an explicit truncation. Likewise, it is convenient to work with infinite sequences rather than long but finite sequences. Finally, for the purposes of statistical modelling we often introduce auxiliary random variables (e.g. statistical parameters) and these are conveniently represented with non-finite and unbounded realms.

So I will presume the following principle:

**Definition 1.1** (Principle of Excluding Pathologies, PEP)**.**
*Extensions to non-finite realms are made for the convenience of the statistician; it is the statistician's responsibility to ensure that such extensions do not introduce pathologies that are not present in the finite realm.*

These notes consider random quantities with finite realms. But I have taken care to ensure that the results also apply, with minor amendments, in the more convenient (but less realistic) case of non-finite and even non-countable realms.

## 1.2 *Introduction to expectation*

Let $X$ be a random quantity—under what conditions might I be said to 'know' $X$? Philosophers have developed a working definition for knowledge: knowledge is 'justified true belief' (Ladyman, 2002, pp. 5–6). So I would know $X$ if I had carried out the instructions specified by $X$ myself, or if they had been carried out by someone I trusted. In other circumstances—for example instructions

that take place in the future—I have belief, but not knowledge.[5] Expectations and the expectations calculus are a way of quantifying and organising these beliefs, so that they hold together sensibly.

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

For concreteness, let $X$ be sea-level rise by 2100, suitably operationalised.[6] This is a random quantity about which no one currently has knowledge, and about which beliefs vary widely from person to person. When I consider my own beliefs about sea-level rise, I find I do not have a single value in mind. Instead, I have values, more or less nebulous, for quantities that I consider to be related to sea-level rise. So I believe, for example, that sea-level rise over the last century is of the order of 10's of centimetres. That the Greenland icesheet and the Western Antarctic icesheet each contain enough ice to raise sea-level globally by between 6 and 7 metres. But that simulations suggest that they will not melt substantially by 2100. But I am cautious about the value of simulations of complex environmental systems. And a lot more things too: about people I know who work in this field, the degree of group-think in the field as a whole, the pressures of doing science in a field related to the effects of climate change, and so on. I do not have well-formed beliefs about sea-level rise, but it turns out that I have lots of ill-formed beliefs about things related to sea-level rise.

[6] Actually, it is an interesting challenge to operationalise this quantity.

And if I wanted to I could easily acquire more beliefs: for example I could ask a glaciologist for her opinion. But once this was given, this would simply represent more related beliefs (my beliefs about her beliefs) to incorporate into my beliefs. And she will be facing exactly the same challenge as me, albeit with a richer set of beliefs about things related to sea-level rise.

I do not think there is any formal way to model the mental processes by which this collection of ill-formed beliefs about things related to sea-level rise get turned into a quantitative expression of my beliefs about sea-level rise. Ultimately, though, I can often come up with some values, even though I cannot describe their provenence. For sea-level rise by 2100, 80 cm from today seems about right to me. I could go further, and provide a range: unlikely to be less than 40 cm, or more than 350 cm, perhaps. These are unashamedly guesses, representing my ill-defined synthesis of my beliefs about things related to sea-level rise.[7] Were you the Mayor of London, you would be well-advised to consult someone who knows more about sea-level rise than I do. But you should not think that she has a better method for turning her beliefs into quantities than I do. Rather, she starts with a richer set of beliefs.

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

These considerations lead me to my first informal definition of an expectation.

**Definition 1.2** (Expectation, informal)**.**

*Let $X$ be a random quantity. My expectation for $X$, denoted $\mathrm{E}(X)$, is a sensible guess for $X$ which is likely to be wrong.*

We will need to define 'sensible' in a way that is generally acceptable, in order for you to understand the conditions under which my

expectation is formed (Sec. 1.3). I am using 'guess' to describe my ill-defined synthesis of my beliefs related to $X$. And I am stressing that it is common knowledge that my guess is likely to be wrong. I think this last point is important, because experts (e.g. glaciologists) may be reluctant to provide wrong guesses, preferring to say nothing at all. So let's get the wrongness out in the open. As the Mayor of London, I would much rather have the wrong guess of a glaciologist than the wrong guess of a statistician.

Now I am able to provide an informal definition of statistical inference. This definition is in the same vein as L.J. Savage's definition of 'statistics': "quantitative thinking about uncertainty as it affects scientific and other investigations" (Savage, 1960, p. 541), although adapted to the use of expectation as primitive, and to the limitations of our beliefs.

**Definition 1.3** (Statistical inference, informal)**.**
*Statistical inference is checking that my current set of expectations is sensible, and extending this set to expectations of other random quantities.*

Chapter 2 discusses statistical inference in detail. One point is worth anticipating here. I may well discover that if $X$ is some random quantity for which I cannot provide an expectation directly, then my $\mathrm{E}(X)$ based on my current set of expectations is not constrained to a single value, but may be an interval of possible values: I would say I was 'undecided' about $\mathrm{E}(X)$. The notable absence of 'undecided' in modern statistical inference is intriguing, and will be investigated further, below (Chapter 4).

There is no mention in Def. 1.3 of data, which is because inference can proceed without data. But, obviously, data are often available. They require no special treatment, though. The datum $Y \to y$ is simply a special type of belief: the knowledge (i.e. justified true belief) that the operation $Y$ was carried out and the vaue $y$ was the result. This is discussed in Sec. 2.4.

## 1.3 *Definition and simple implications*

The axioms given below (in Def. 1.4) are the standard axioms of expectation. In this respect they are the 'what' rather than the 'why'. For the 'why' I refer back to the previous section, and the informal definition of expectation in Def. 1.2. I interpret these axioms as a minimal characterisation of 'sensible'.

**Definition 1.4** (Axioms of expectation)**.**
*Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random quantities with finite realms. Then the expectations of $X$ and $Y$ must satisfy the following properties:*

0. $\mathrm{E}(X) \in \mathbb{R}$ *exists and is unique,* *(existence)*

1. $\mathrm{E}(X) \geq \min \mathcal{X}$, *(lower boundedness)*

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

You can see that this sets the bar on 'sensible' quite low—it is continuing a source of amazement to me that we can do so much with such simple beginnings. The 'existence' axiom does not insist that I know my expectation for every random quantity, but only that I acknowedge that it exists as a (real-valued) number and is unique. I use the word *undecided* to describe expectations that I am not currently able to quantify.

'Lower-boundedness' is an extremely weak condition, given that $\mathcal{X}$ ought to be inferrable from $X$ itself, and have nothing to do with my particular beliefs about things related to $X$. For example, if $X$ is the weight of this orange, then $\min \mathcal{X}$ must be 0 g, to represent the physical impossibility of an orange with negative weight. I might believe that the weight cannot be less than 50 g, but lower boundedness only requires that my $\mathrm{E}(X)$ is non-negative.

'Additivity' is a bit more subtle. I think we would all agree that if $X$ and $Y$ were the weights of two oranges, then anything other than $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ would be not-sensible. But there are more interesting situations. Consider the following example, following Ellenberg (2014, ch. 11).[8] A man has seven children, and is planning to leave his £1m fortune to exactly one, the choice to be decided by the day of the week on which he dies. Let $X_i$ be the amount in £m received by the $i$th child. The most likely outcome for each child is $X_i \to 0$. And yet $X_1 + \cdots + X_7 \to 1$ with certainty. And so to interpret $\mathrm{E}(X_i)$ as 'most likely' will not satisfy the additivity axiom. Most people in this case would take $\mathrm{E}(X_i) \leftarrow 1/7$ for each $i$, using a symmetry argument, and this would satisfy all three axioms. Mind you, $\mathrm{E}(X_1) \leftarrow 1$ and $\mathrm{E}(X_2) \leftarrow \cdots \leftarrow \mathrm{E}(X_7) \leftarrow 0$ would also satisfy all three axioms.

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

The asymmetric expectations in the seven children example illustrates the point that the bar on 'sensible' is quite low. There is a strong case for introducing another word to mean precisely that the axioms are satisfied, so that 'sensible' does not seem misapplied. The standard choice among Bayesian statisticians is *coherent*, following de Finetti (1974/75). From now on I will use 'coherent' to describe a set of expectations satisfying Def. 1.4. In public discourse, when my expectations matter to people other than myself, I would use *defensible* to mean something more than simply coherent, in the same way that a logician would distinguish a valid argument from a sound one.

* * *

The axioms in Def. 1.4 have many implications. There are several reasons for considering these implications explicitly:

1. They give us confidence in the axioms, if they seem consistent with our interpretation of expectation.

2. They prevent us from making egregious specifications for expectations.

3. They provide a quick source of results when we assume that our

beliefs are coherent.

Here I will just pick out a few of the basic implications, which are important enough to have names.

**Theorem 1.1** (Implied by additivity alone)**.**

1. $\mathrm{E}(0) = 0$ *and* $\mathrm{E}(-X) = -\mathrm{E}(X)$,
2. $\mathrm{E}(X_1 + \cdots + X_k) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_k)$. *(finite additivity)*
3. $\mathrm{E}(aX) = a\,\mathrm{E}(X)$. *(linearity)*

*Proof.*

1. Since $0 = 0 + 0$, we have $\mathrm{E}(0) = 2\,\mathrm{E}(0)$ from which the result follows. The second result follows from $0 = X + (-X)$.
2. Follows iteratively from $X_1 + \cdots + X_k = X_1 + (X_2 + \cdots + X_k)$.
3. Here is the proof for rational $a$. If $i$ is a non-negative integer, then $\mathrm{E}(iX) = i\,\mathrm{E}(X)$ by the previous result. And if $j$ is a positive integer, then $\mathrm{E}(X) = \mathrm{E}(jX/j) = j\,\mathrm{E}(X/j)$ from which $\mathrm{E}(X/j) = \mathrm{E}(X)/j$. Hence $\mathrm{E}(aX) = a\,\mathrm{E}(X)$ whenever $a$ is a non-negative rational number. Extend to $a < 0$ using $aX = |a|(-X)$.

The extension of the final part to real numbers is slightly subtle; see de Finetti (1974, footnote on p. 75). □

The *linearity* property is usually taken to subsume finite additivity, giving

$$\mathrm{E}(a_1X_1 + \cdots + a_kX_k) = a_1\,\mathrm{E}(X_1) + \cdots + a_k\,\mathrm{E}(X_k). \qquad \textit{(linearity)}$$

This is the property that must be strengthened in the case where there are a non-finite number of random quantities, or, which comes to the same thing, the realm of a random quantity is non-finite. The stronger *countable additivity* axiom extends finite additivity and (finite) linearity to countably-infinite sequences. This stronger axiom is almost universally accepted, as it ought to be according to the PEP (Def. 1.1).[9]

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

Here are some further implications, using both additivity and lower-boundedness.

**Theorem 1.2.**

1. $\mathrm{E}(a) = a$, *(normalisation)*
2. *If* $X \leq Y$, *then* $\mathrm{E}(X) \leq \mathrm{E}(Y)$, *(montonicity)*
3. $\min \mathcal{X} \leq \mathrm{E}(X) \leq \max \mathcal{X}$ *(convexity)*
4. $|\mathrm{E}(X)| \leq \mathrm{E}(|X|)$. *(triangle inequality)*

*Proof.*

1. $a \geq a$, so $\mathrm{E}(a) \geq a$. And $-a \geq -a$, so $\mathrm{E}(-a) \geq -a$, and then $\mathrm{E}(-a) = -\mathrm{E}(a)$ implies that $\mathrm{E}(a) \leq a$; hence $\mathrm{E}(a) = a$.
2. The minimum of the realm of $Y - X$ is non-negative, hence $\mathrm{E}(Y - X) \geq 0$ which implies that $\mathrm{E}(X) \leq \mathrm{E}(Y)$.
3. Same argument as above, as $X$ is never greater than $\max \mathcal{X}$, and $\mathrm{E}(\max \mathcal{X}) = \max \mathcal{X}$.
4. Same argument as above, as $-|X|$ is never greater than $X$, and $X$ is never greater than $|X|$. Together these imply that $\mathrm{E}(X) \leq \mathrm{E}(|X|)$ and $-\mathrm{E}(X) \leq \mathrm{E}(|X|)$, as required.

□

Finally in this section, we have *Schwarz's inequality*, which is proved using linearity and monotonicity.

**Theorem 1.3** (Schwarz's inequality)**.**

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

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

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

This quadratic in $a$ cannot have two distinct real roots, because that would indicate a negative value for the expectation, violating monotonicity. Then it follows from the standard formula for the roots of a quadratic[10] that

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

or $\{\mathrm{E}(XY)\}^2 \leq \mathrm{E}(X^2)\,\mathrm{E}(Y^2)$, as required.[11] □

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

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

[11] This proof is adapted from Williams (1991, sec. 6.8).

Another similarly useful result is Jensen's inequality, which concerns the expectation of convex functions of random quantities. This result can also be proved at this stage using linearity and monotonicity, but only if we accept the Separating Hyperplane Theorem. Instead, I will defer Jensen's inequality until Sec. 2.1, at which point I will be able to give a self-contained proof.

### *1.3.1* Quantities related to expectation*

Here is a brief summary of other quantities that are defined in terms of expections, and their properties. These properties follow immediately from the axioms and are not proved.

If $X$ is a random quantity with expectation $\mu$, then the *variance* of $X$ is defined as

$$\mathrm{Var}(X) := \mathrm{E}\{(X - \mu)^2\},$$

and often denoted $\sigma^2$; clearly $\sigma^2 \geq 0$ by monotonicity. Expanding out shows that

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2.$$

The square root of $\mathrm{Var}(X)$ is termed the *standard deviation*; I denote it as $\mathrm{Sd}(X)$. It has the same units as $X$, and is often denoted as $\sigma$. $\mathrm{Var}(a + bX) = b^2 \mathrm{Var}(X)$, and $\mathrm{Sd}(a + bX) = b\,\mathrm{Sd}(X)$.

If $X$ and $Y$ are two random quantities with expectations $\mu$ and $\nu$ then the covariance of $X$ and $Y$ is defined as

$$\mathrm{Cov}(X, Y) := \mathrm{E}\{(X - \mu)(Y - \nu)\}.$$

Hence $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ and $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. Expanding out shows that

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mu\nu.$$

$\mathrm{Cov}(a + bX, c + dY) = bd\,\mathrm{Cov}(X, Y)$, $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$, and, by iteration,

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

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

$$\mathrm{Var}(X_1 + \cdots + X_m) = \sum_i \mathrm{Var}(X_i).$$

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

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

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

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

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

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

and is often denoted $\rho$. Schwartz's inequality implies that

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

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

## 1.4 Probability

### 1.4.1 Definitions

If expectation is primitive, then probability is just a special type of expectation. In a nutshell, a probability is the expectation of the indicator function of a random proposition.

You may want to consult the material on first order logic in Sec. 1.A: in particular, the definition of a first order sentence on p. 24. This is the basis for the following definition.

**Definition 1.5** (Random proposition)**.**

*A random proposition is a first order sentence in which one or more constants have been replaced by random quantities.*

In the simplest case, if $x$ and $y$ are constants then $x \doteq y$ is a first order sentence.[12] If $X$ and $Y$ are random quantities, then $X \doteq x$ and $X \doteq Y$ are random propositions. The truth value of a first order sentence is known, but the truth value of a random proposition is uncertain, because it contains random quantities instead of constants.

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

The *indicator function* of a first order sentence $\psi$ is the function $\mathbb{1}_\psi$ for which

$$\mathbb{1}_\psi := \begin{cases} 0 & \psi \text{ is false} \\ 1 & \psi \text{ is true.} \end{cases}$$

In other words, the indicator function turns false into *zero* and true into *one*. Note that the indicator function of a conjunction of sentences is the product of the indicator functions:

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

The indicator function is used to define a probability.

**Definition 1.6** (Probability)**.**

*Let $Q$ be a random proposition. Then* $\Pr(Q) := \mathrm{E}(\mathbb{1}_Q)$.

So, continuing the example for the simplest case given above, $\Pr(X \doteq x) := \mathrm{E}(\mathbb{1}_{X \doteq x})$ and $\Pr(X \doteq Y) := \mathrm{E}(\mathbb{1}_{X \doteq Y})$. These probabilities are expectations of specified functions of the random quantities $X$ and $Y$.

As a more detailed illustration, if $X_i$ is the time taken for horse $i$ to complete the course, then the sentence that horse $i$ wins is

$$q(\boldsymbol{x}) \leftarrow \bigwedge_{j \neq i} (x_i \dot{<} x_j);$$

the sentence that horse $i$ is second is

$$q'(\boldsymbol{x}) \leftarrow \bigvee_{k \neq i} \bigwedge_{j \neq i,k} (x_k \dot{<} x_i) \wedge (x_i \dot{<} x_j);$$

the sentence that horse $i$ is first or second is

$$q''(\boldsymbol{x}) \leftarrow q(\boldsymbol{x}) \vee q'(\boldsymbol{x})$$

and so on. So my probability that horse $i$ wins is $\Pr\{q(\boldsymbol{X})\} := \mathrm{E}\{\mathbb{1}_{q(\boldsymbol{X})}\}$, and my probability that horse $i$ is first or second is $\Pr\{q''(\boldsymbol{X})\}$, and so on.

This definition of probability might seem strange to people used to treating probability as primitive. And so it is worth taking a moment to check that the usual axioms of probability are satisfied. Thus, if $P$ and $Q$ are random propositions:

1. $\Pr(P) \geq 0$, by lower-boundedness.

2. If $P$ is a tautology, then $\mathbb{1}_P = 1$ and $\Pr(P) = 1$ by normalisation.

3. If $P$ and $Q$ are incompatible, i.e. $\mathbb{1}_{P\wedge Q} = 0$, then $\mathbb{1}_{P\vee Q} = \mathbb{1}_P + \mathbb{1}_Q$, and $\Pr(P \vee Q) = \Pr(P) + \Pr(Q)$, by linearity.

Thus all of the usual probability results apply; I will not give them here, with the exception of the following basic but important result.

**Theorem 1.4.** *If $q(\boldsymbol{x})$ and $q'(\boldsymbol{x})$ are first-order sentences and $q(\boldsymbol{x}) \implies q'(\boldsymbol{x})$ then $\Pr(Q) \leq \Pr(Q')$, where $Q := q(\boldsymbol{X})$ and $Q' := q'(\boldsymbol{X})$.*

*Proof.* Follows by monotonicity, because $q(\boldsymbol{x}) \implies q'(\boldsymbol{x})$ implies that $\mathbb{1}_{q(\boldsymbol{x})} \leq \mathbb{1}_{q'(\boldsymbol{x})}$. □

* * *

*Notation.* One very useful notation helps us to express probabilities of conjunctions efficiently. If $\{A_1, \ldots, A_k\}$ is a collection of random propositions, then define

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

In other words, commas between random propositions represent conjunctions.

### 1.4.2 *Interpreting probability*

The definition of the probability of a random proposition $Q$ is 'the expectation of the indicator function of $Q$'. This is a mouthful, and it does not lend itself easily to our intuition, even in cases where we feel we have a grip on what an expectation is, because $\mathbb{1}_{q(\boldsymbol{x})}$ is a very non-linear function of $\boldsymbol{x}$. Therefore it is helpful to have a more intuitive understanding of probability, which, in due course, will help us to extend the notion to conditional probability (in Chapter 3).

The intuition comes from betting, a commonplace activity that everyone is aware of, even if not everyone does it. Suppose I bet $p$ on a random proposition $Q$, in order to win 0 if $Q$ is false and 1 is $Q$ is true. My expected outcome for this bet is

$$\mathrm{E}(\mathbb{1}_Q - p).$$

We say that $p$ is my *fair price* for the bet exactly when my expected outcome is 0. From the definition of $\Pr(Q)$ it follows that

> $\Pr(Q)$ *is my fair price for a bet that pays* 0 *if* $Q$ *is false and* 1 *if* $Q$ *is true.*

In other words, if someone offered me a bet on $Q$ with $p < \Pr(Q)$ then I would think this a good bet, and if someone offered me $p > \Pr(Q)$ I would think this a poor bet. Of course it is not necessary for me actually to bet: this is just a reflection on my part. It is like walking past a shop and seeing a washing-machine in the window with a price displayed, and thinking "that's cheap for a

washing-machine". I do not have to buy the washing-machine just because I think it is cheap.

Confusingly, actual bets tend not to be expressed in terms of $p$. Two conventions are *fractional odds* and *decimal odds*. A bookmaker who offers you fractional odds of '$n$-to-$k$ on' is indicating that if you bet $n$ he will contribute $k$, with the total going to you for a win, and him for a loss. To find $p$ divide by $n+k$ to reduce the value of a win to 1, and then

$$p = \frac{n}{n+k}.$$

When $p < 0.5$ the offer is expressed as '$n$-to-$k$ against' or just '$n$-to-$k$', for which

$$1 - p = \frac{n}{n+k} \quad \text{or} \quad p = \frac{k}{n+k}.$$

Odds of 1-to-1 are termed 'evens'. Fractional odds are used by bookmakers in the UK and Ireland.

With decimal odds, $d \geq 1$ represents the amount you get back for a win, including your bet, if your bet is 1. Again, dividing by $d$ to reduce the value of a win to 1 gives

$$p = \frac{1}{d}.$$

Equating $p$ in the two expressions shows that

$$d = 1 + \frac{n}{k}$$

for fractional odds of $n$-to-$k$.[13] Decimal odds are used on betting exchanges.

[13] I.e. $n$-to-$k$ against, also written as $n/k$ or $n : k$. My way of turning $n$-to-$k$ into $p$ is to take the reciprocal of $1 + n/k$. Thus 6-to-1 becomes $1/7$, or 0.14.

This digression into odds is just to make the point that sometimes someone will offer you a $p$ for some $Q$. You can decide whether or not it is a good bet, and in doing so you are ascertaining a lower or an upper bound on your $\Pr(Q)$. Generally, however, no one is offering a bet on the particular $Q$'s that interest you, and the betting analogy is just a device to help you to quantify your $\Pr(Q)$. Having said that, climate scientists have offered each other bets in public as a demonstration of their beliefs about the probability of climate change.[14] In these cases the negotiation of the bet is precisely the process of operationalising the notion of climate change, and it is interesting that this process can be protracted.

[14] E.g. Dr James Annan, see `http://en.wikipedia.org/wiki/James_Annan`.

### 1.4.3 *Simple inequalities*

There are some simple inequalities linking expectations and probabilities, and these can be useful for providing bounds on probabilities, or for specifying beliefs about a random quantity that includes both probabilities of propositions about $X$ and expectations of functions of $X$. The starting-point for many of these is *Markov's inequality*.

**Theorem 1.5** (Markov's inequality)**.**

*If $X$ is non-negative and $a > 0$ then*

$$\Pr(X \dot{\geq} a) \leq \frac{\mathrm{E}(X)}{a}.$$

*Proof.* Follows from monotonicity and linearity, because

$$a\, \mathbb{1}_{X \dot{\geq} a} \leq X,$$

see Figure 1.2. Taking expectations of both sides and rearranging gives the result. □

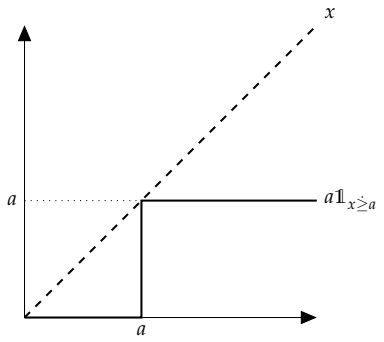

Figure 1.2: Markov's inequality.

One immediate generalisation of Markov's inequality is

$$\Pr(X \dot{\geq} a) \leq \frac{\mathrm{E}\{g(X)\}}{g(a)}$$

whenever $g$ is a non-negative increasing function: this follows because $g(X)$ is non-negative and because $X \dot{\geq} a \iff g(X) \dot{\geq} g(a)$. A useful application of this generalisation is

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

which follows because $|x|^r$ is a non-negative increasing function of $|x|$ for every positive $r$. A special case is *Chebyshev's inequality*. This is usually expressed in terms of $\mu := \mathrm{E}(X)$ and $\sigma^2 := \mathrm{E}\{(X-\mu)^2\}$ (see Sec. 1.3.1). Setting $r \leftarrow 2$ then gives

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

for $a > 0$.

## 1.5 *The Fundamental Theorem of Prevision*

The Fundamental Theorem of Prevision (FTP) is due to Bruno de Finetti (see de Finetti, 1974, sec. 3.10).[15] Its epithet 'fundamental' is well-deserved, because it provides a complete characterisation of the set of expectations that are consistent with the axioms of expectation given in Def. 1.4. It features heavily in Sec. 2.1.

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

The following theorem uses the $(s-1)$-dimensional *unit simplex*, defined as the set

$$\mathbb{S}^{s-1} := \left\{ \boldsymbol{p} \in \mathbb{R}^s : p_j \geq 0 \text{ and } \textstyle\sum_j p_j = 1 \right\}. \tag{1.2}$$

This set occurs repeatly below; Figure 1.3 illustrates $\mathbb{S}^2$. Note the standard mathematical convention that we denote a set by its dimension and not by the space in which it is embedded; thus $\mathbb{S}^2$ is a 2D-triangle embedded in $\mathbb{R}^3$.

Not done yet!

I write the conjunction $X_1 \doteq x_1^{(j)} \wedge \cdots \wedge X_m \doteq x_m^{(j)}$ as $\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}$, but there is a better notation for this given below, in Sec. 1.5.1.

Figure 1.3: The unit simplex $\mathbb{S}^2$.

**Theorem 1.6** (Fundamental Theorem of Prevision, FTP)**.**

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

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

*be their joint realm. Then* E *is a valid expectation if and only if there is a* $\boldsymbol{p} \in \mathbb{S}^{s-1}$ *for which*

$$\mathrm{E}\{g(\boldsymbol{X})\} = \sum_{j=1}^{s} g(\boldsymbol{x}^{(j)}) \cdot p_j \tag{†}$$

*for all* $g : \mathbb{R}^m \to \mathbb{R}$*. In this case,* $p_j = \Pr(\boldsymbol{X} \doteq \boldsymbol{x}^{(j)})$*.*

*Proof.*

(⇐). This is just a matter of checking that (†) satisfies the axioms in Def. 1.4. The zeroth axiom is obviously satisfied. Lower-boundedness follows from

$$\begin{aligned} \mathrm{E}\{g(\boldsymbol{X})\} &= \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot p_j \\ &\geq \min_{\boldsymbol{p} \in \mathbb{S}^{s-1}} \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot p_j = \min_j g(\boldsymbol{x}^{(j)}), \end{aligned}$$

as required. Additivity follows immediately from the linearity of (†). Let $g(\boldsymbol{x}) \leftarrow \mathbb{1}_{\boldsymbol{x} \doteq \boldsymbol{x}^{(i)}}$. Then

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

as required.

(⇒). Note that

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

Hence

$$\begin{aligned} \mathrm{E}\{g(\boldsymbol{X})\} &= \mathrm{E}\left\{ g(\boldsymbol{X}) \textstyle\sum_j \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} \\ &= \mathrm{E}\left\{ \textstyle\sum_j g(\boldsymbol{X}) \cdot \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} \\ &= \mathrm{E}\left\{ \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot \mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}} \right\} \\ &= \textstyle\sum_j g(\boldsymbol{x}^{(j)}) \cdot \mathrm{E}\{\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\} \qquad \text{by linearity.} \end{aligned}$$

The result then follows on setting $p_j \leftarrow \mathrm{E}\{\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}^{(j)}}\}$, as $p_j \geq 0$ by lower-boundedness, and $\sum_j p_j = 1$ by linearity and normalisation, from (‡). Hence $\boldsymbol{p} \in \mathbb{S}^{s-1}$.

□

Eq. (†) is familiar as the definition of an expectation in the case where probability is taken as primitive. In contrast, the FTP states that it is an inevitable consequence of the axioms of expectation that probabilities $\boldsymbol{p} \in \mathbb{S}^{s-1}$ must exist, satisfying (†).

### 1.5.1 Notation and marginalisation

Expressions such as $\Pr(\boldsymbol{X} \doteq \boldsymbol{x})$ occur frequently in statistics, because the FTP is such an important result. A standard notation for such expressions is to write

$$\mathrm{p}(\boldsymbol{x}) := \Pr\left(\boldsymbol{X} \doteq \boldsymbol{x}\right) := \Pr\left(\bigwedge_i X_i \doteq x_i\right)$$

where the random quantities in $\mathrm{p}(\cdot)$ are identified by the argument, using the capital/small letter convention. Ornaments are used when it is necessary to distinguish two different values for the same random quantities, so that $\mathrm{p}(\boldsymbol{x}') := \Pr(\boldsymbol{X} \doteq \boldsymbol{x}')$.

Combining the comma notation and the p notation,

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) := \Pr\left\{\left(\bigwedge_i X_i \doteq x_i\right) \wedge \left(\bigwedge_j Y_j \doteq y_j\right)\right\}$$

and by comparing the left- and right-hand sides of this expression it is clear that these two notations are earning their keep.

One helpful convention with the 'p' notation is to define $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y})$ over the whole of the product set $\mathcal{X} \times \mathcal{Y}$, even though the realm of $(\boldsymbol{X}, \boldsymbol{Y})$ may be a subset of this product set. In other words, set

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) \leftarrow 0$$

if $(\boldsymbol{x}, \boldsymbol{y})$ is in $\mathcal{X} \times \mathcal{Y}$ but not in the realm of $(\boldsymbol{X}, \boldsymbol{Y})$. This simplifies marginalisation operations, as defined in Sec. 1.5.1. It is common to go further and take $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) \leftarrow 0$ for every possible value of $(\boldsymbol{x}, \boldsymbol{y})$ that is not in the realm of $(\boldsymbol{X}, \boldsymbol{Y})$. This it not necessary in what follows, but does no harm.

The first outing for this new notation, and also the opportunity to introduce some other conventions, is in *marginalisation*, which is 'collapsing' a probability assessment onto a subset of random quantities. This is a crucial operation in statistical modelling.

**Theorem 1.7** (Marginalisation)**.** *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two collections of random quantities. Then*

$$\mathrm{p}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}).$$

---

Removing $\boldsymbol{Y}$ in this way is termed *marginalising out* $\boldsymbol{Y}$. Whenever the domain of the index of a sum is left undefined it can be taken to be the realm of the index's parent random quantities: in the case above, the '$\boldsymbol{y}$' under '$\sum$' is to be read as $\boldsymbol{y} \in \mathcal{Y}$.

*Proof.* Set $g(\boldsymbol{x}, \boldsymbol{y}) \leftarrow \mathbb{1}_{\boldsymbol{x} \doteq \boldsymbol{x}'}$ and then

$$\begin{aligned}
\mathrm{p}(\boldsymbol{x}') &= \mathrm{E}(\mathbb{1}_{\boldsymbol{X} \doteq \boldsymbol{x}'}) && \text{by definition} \\
&= \sum_{(\boldsymbol{x}, \boldsymbol{y})} \mathbb{1}_{\boldsymbol{x} \doteq \boldsymbol{x}'} \cdot \Pr\{(\boldsymbol{X}, \boldsymbol{Y}) \doteq (\boldsymbol{x}, \boldsymbol{y})\} && \text{by the FTP, Thm 1.6} \\
&= \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} \mathbb{1}_{\boldsymbol{x} \doteq \boldsymbol{x}'}\ \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) && \text{'p' convention} \\
&= \sum_{\boldsymbol{y}} \Big\{ \sum_{\boldsymbol{x}} \mathbb{1}_{\boldsymbol{x} \doteq \boldsymbol{x}'}\ \mathrm{p}(\boldsymbol{x}, \boldsymbol{y}) \Big\} \\
&= \sum_{\boldsymbol{y}} \mathrm{p}(\boldsymbol{x}', \boldsymbol{y})
\end{aligned}$$

as required. □

A comment about functional relations, of which Thm 1.7 is the first we have seen in this chapter. A functional relation defines a relation for every point in the product set of the free arguments that appear on one or both sides of the relation. So Thm 1.7 is an equality which holds for every point in $\mathcal{X}$ ($x$ being a free argument, and $y$ being a bound argument). Later on, there will be functional relations with $x$ and $y$ on both sides of the relation, and these hold for every point in $\mathcal{X} \times \mathcal{Y}$. Wherever this is not true, the relation will be qualified by an additional condition (see Sec. 3.5.1 for an example).

## *1.A Concepts from first order logic*

Here is a fairly precise statement about commonly-used mathematical terms in first order logic; this account is a *précis* of several sources, including Keisler (2007, ch. 15). First order logic for real numbers is used to define a random proposition and a probability (Sec. 1.4), and an unfamiliar notation is used (e.g. '$\doteq$') to disambiguate a commonly-used notation in statistics.

The language of first order logic comprises functions and variables, predicates, connectives, quantifiers, and punctuation (parentheses and commas). Functions are $n$-ary, indicating that they take $n$ arguments, where $n \geq 0$. Functions that are 0-ary are called *constants*. Variables range over the set of all constants. The meanings of functions (including constants) and predicates depends on the interpretation of the language, but variables, connectives, quantifiers and punctuation have a fixed (conventional) meanings. In these notes, functions, constants, and variables will be real-valued, and predicates will be binary relations.

A *term* is a finite sequence of symbols defined inductively according to:

1. Every constant and every variable is a term;
2. If $t_1, \ldots, t_n$ are terms and $f$ is an $n$-ary function with $n \geq 1$, then $f(t_1, \ldots, t_n)$ is a term.

*Binary relations* have the form $s\, R\, t$, where $s$ and $t$ are terms. The binary relations comprise

$$\doteq,\ \dot{\neq},\ \dot{<},\ \dot{\leq},\ \dot{\geq},\ \text{and}\ \dot{>}.$$

The dot over each symbol indicates that these are predicates, and so mean something different from their usual 'undotted' usage. This is explained further after the description of a first order sentence on p. 24. *Connectives* comprise

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

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

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

There is some redundancy here, since all of these connectives and quantifiers can be constructed from the smaller set $\{\neg, \vee, \exists\}$, but it is much clearer to keep them all.

A *formula* is a finite sequence of symbols defined inductively according to:

1. If $R$ is a relation and $s$ and $t$ are terms then $s\,R\,t$ is a formula.
2. If $\psi$ and $\phi$ are formulae, then

$$\neg\psi,\ \psi \wedge \phi,\ \psi \vee \phi,\ \psi \implies \phi, \text{ and } \psi \iff \phi$$

   are formulae.

3. If $\psi(v)$ is a formula and $v$ is a variable, then

$$\forall v \psi(v) \quad \text{and} \quad \exists v \psi(v)$$

   are formulae.

In a formula, a variable can be either a *free variable* or a *bound variable*. It is free if it is not quantified, otherwise it is bound. For example, in the formula $\forall v(v\,R\,w)$ the variable $v$ is bound and the variable $w$ is free. A formula with no free variables is a *first order sentence*: these are the formulae with well-defined truth values. Thus if $a$ and $b$ are constants then $a \dot{\leq} b$ is a sentence. If $f$ and $g$ are 1-ary functions, then

$$\forall v(f(v) \doteq g(v))$$

is a sentence, which is true if $f$ and $g$ are the same function, and false if they are not. If $\psi(v)$ is a formula with a free variable $v$ and $c$ is a constant, then $\psi(c)$ is a sentence. For example, $(v \dot{\leq} 3)$ is a formula with a free variable $v$, and $(2 \dot{\leq} 3)$ is a sentence.

The truth of a sentence is defined inductively according to:

1. If $R$ is a binary relation then the sentence $a\,R\,b$ is true exactly when the constants $a$ and $b$ are defined and $(a, b) \in R$.
2. If $\psi$ and $\phi$ are sentences and $C$ is a connective then the truth of $\psi\,C\,\phi$ is determined according to the usual truth tables.
3. The sentence $\forall v \psi(v)$ is true exactly when $\psi(c)$ is true for all constants $c$.
4. The sentence $\exists v \psi(v)$ is true exactly when $\psi(c)$ is true for some constant $c$.

It should be clear now why it is important to distinguish the predicate '$\doteq$' from the more usual '$=$'. The first-order sentence '$\psi \doteq \phi$' evaluates to false or true, depending on the values of $\psi$ and $\phi$, but the expression '$\psi = \phi$' is an assertion that the objects $\psi$ and $\phi$ are equal to each other. In first-order logic, predicates are written $P(x, y, z)$. But when the predicates are binary predicates it is much clearer to write $P(x, y)$ as $x\,P\,y$, known as 'infix' notation. Unfortunately for us, this clashes with the more usual uses of

symbols such as '$=$' and '$\leq$', which is why the infix predicates are ornamented with dots.

The logician Gottlob Frege distinguished between the assertion $\psi = \phi$ and the first-order sentence $\psi \doteq \phi$. He would have treated '$=$' as a binary predicate and '$\psi = \phi$' as a first-order sentence. He introduced the 'turnstile' symbol $\vdash$ such that '$\vdash (\psi = \phi)$' read '$\psi = \phi$ is true'; Paris (1994) provides a modern example of the turnstile symbol in use. In these notes I do not need it precisely because I distinguish using a dot between the first-order sentence '$\psi \doteq \phi$' and the assertion '$\psi = \phi$'.