

# Why Does Statistics Work? An Essay on the Foundations of Statistical Inference

Duncan K. Foley<sup>1</sup>

January 27, 2005

<sup>1</sup>Preliminary and incomplete. All rights reserved.

# Contents

<b>1</b>	<b>Interpretation of probability</b>	<b>1</b>
1.1	Models and interpretations . . . . .	1
1.2	Some history . . . . .	2
1.3	Toward a unified interpretation of probability . . . . .	4
1.4	Mathematics . . . . .	6
<b>2</b>	<b>Systems of probabilities</b>	<b>7</b>
2.1	Finite event spaces . . . . .	7
2.1.1	Primitive events and probabilities . . . . .	7
2.1.2	Compound events and probabilities . . . . .	7
2.1.3	Joint and conditional probabilities . . . . .	8
2.1.4	Exercises . . . . .	9
2.2	Infinite event spaces . . . . .	9
<b>3</b>	<b>Probability distributions</b>	<b>12</b>
3.1	Representing probabilities . . . . .	12
3.1.1	Probability mass functions . . . . .	12
3.1.2	Probability density functions . . . . .	12
3.1.3	Probability distributions . . . . .	14
3.1.4	Normalizing probabilities . . . . .	14
3.2	Joint probability distributions . . . . .	15
3.2.1	Conditional probability distributions . . . . .	15
3.2.2	Marginal probability distributions . . . . .	16
3.3	Simplifying probability distributions . . . . .	18
3.4	Convolutions of probability distributions . . . . .	18
3.5	Problems . . . . .	19
<b>4</b>	<b>The economic model of probability</b>	<b>20</b>
4.1	The Laplacian concept of probability . . . . .	20
4.1.1	The scope of probability judgments . . . . .	20
4.2	The economic offer set . . . . .	21
4.3	No Dutch book . . . . .	23
4.4	Differentiability at the origin . . . . .	25
4.4.1	Markets and probabilities . . . . .	25

4.5	Probability judgments . . . . .	28
4.6	Consistency and reasonableness . . . . .	28
4.7	Probability and evolution . . . . .	29
4.8	Differentiability . . . . .	30
4.9	Dutch book . . . . .	30
4.10	A tradeoff . . . . .	32
4.11	Who has probabilities? . . . . .	32
4.11.1	Where consistency counts . . . . .	32
4.12	Probability theory is an ideal abstraction . . . . .	33
<b>5</b>	<b>The laws of probability</b>	<b>35</b>
5.1	The linearity of expectations . . . . .	35
5.2	Logical operations and the laws of probability . . . . .	36
5.3	Coherence and probability . . . . .	37
<b>6</b>	<b>Conditional probability</b>	<b>39</b>
6.1	Calling the bet off . . . . .	39
6.2	Two moments of learning . . . . .	40
<b>7</b>	<b>What can we bet on?</b>	<b>43</b>
7.1	Operationalism . . . . .	43
7.2	Unresolvable bets . . . . .	44
<b>8</b>	<b>The frequency model of probability</b>	<b>47</b>
8.1	Probabilities as frequencies . . . . .	48
8.1.1	The problem of finite sample size . . . . .	48
8.1.2	Probabilities as tendencies . . . . .	49
8.2	Are frequencies “objective”? . . . . .	49
8.3	Probabilities that cannot be frequencies . . . . .	50
8.4	Transcending the frequency model . . . . .	51
<b>9</b>	<b>Statistical inference</b>	<b>52</b>
9.1	The problem of statistical inference . . . . .	52
9.2	The structure of statistical inference . . . . .	53
9.3	Approaches to statistical inference . . . . .	54
9.3.1	Operational Laplacianism . . . . .	54
9.3.2	Parametric Laplacianism . . . . .	55
9.3.3	Parametric classicalism . . . . .	56
<b>10</b>	<b>Operational Bernoulli Inference</b>	<b>57</b>
10.1	Repeated Bernoulli trials . . . . .	57
10.2	Fully informative statistics and exchangeability . . . . .	57
10.3	Posterior probabilities . . . . .	58
10.3.1	Exercises . . . . .	59
10.4	Bernoulli Posterior . . . . .	59
10.4.1	Bernoulli Posterior when $m = 1$ . . . . .	60

10.4.2 Exercises . . . . .	60
10.4.3 $m \rightarrow \infty$ . . . . .	60
10.4.4 Exercises . . . . .	61
10.4.5 The Bernoulli Posterior and the hypergeometric probability distribution . . . . .	62
10.4.6 Exercises . . . . .	62
10.5 The Normal Approximation to the Bernoulli Posterior . . . . .	62
10.5.1 Exercises . . . . .	68
10.6 Common statistical practice . . . . .	69
10.6.1 The Normal Approximation to the Relative Entropy . . . . .	69
10.7 Informational economy in the Bernoulli model . . . . .	70
10.8 Summary . . . . .	71
<b>11 Parametric Bernoulli inference</b>	<b>72</b>
11.1 Parametric Laplacian approach . . . . .	72
11.1.1 Exercises . . . . .	74
11.1.2 Parametric and operational Laplacianism . . . . .	74
11.2 The classical parametric approach . . . . .	74
11.2.1 Criticisms of the classical approach . . . . .	76
11.2.2 Classical sampling from an operational Laplacian point of view . . . . .	77
11.3 Fully informative and sufficient statistics . . . . .	78
11.4 Conclusion . . . . .	78
<b>12 Rationalizing statistical practice</b>	<b>80</b>
12.1 Reporting statistical results . . . . .	80
12.1.1 The “true” model . . . . .	81
<b>13 Comparing Bernoulli trials</b>	<b>84</b>
13.1 Comparative experiments . . . . .	84
13.2 Statistical analysis of experimental data . . . . .	84
13.2.1 Should we take into account an interaction? . . . . .	86
13.3 Signal and noise in experimental Bernoulli situations . . . . .	87
13.4 <i>Timon of Athens</i> . . . . .	87
<b>14 The multinomial model</b>	<b>91</b>
14.1 Generalizing the Bernoulli model . . . . .	91
14.1.1 Exercises . . . . .	92
14.2 Multinomial Posterior for $m = 1$ . . . . .	92
14.3 Relative entropy . . . . .	93
14.3.1 Exercises . . . . .	94
14.4 Multinomial Posterior . . . . .	94
14.4.1 Exercises . . . . .	95

<b>15 A general model of statistical inference</b>	<b>96</b>
15.1 Bins and pixels . . . . .	96
15.2 Implementing the multinomial approach . . . . .	97
15.2.1 Examples of the multinomial model . . . . .	98
15.3 Evaluation of the multinomial approach . . . . .	101
15.4 Ways to achieve smoothness . . . . .	101
15.5 Ways to achieve simplification . . . . .	102
15.6 Simplification, smoothness and prior beliefs . . . . .	102
<b>16 Smoothness versus Fit</b>	<b>104</b>
16.1 When do we want smooth posterior distributions? . . . . .	104
16.2 Global smoothness . . . . .	104
<b>17 One-dimensional linear model</b>	<b>106</b>
17.1 Linear model with one-dimensional data . . . . .	106
17.2 Real variable observations:	
fully informative statistics . . . . .	107
17.2.1 Exercise . . . . .	108
17.3 Describing samples in generalized polar coordinates . . . . .	108
17.4 Exercise . . . . .	108
17.5 Jeffreys' prior: $d\mu \frac{ds}{s}$ . . . . .	108
17.6 Posterior probabilities . . . . .	109
17.7 Transformations of the scale of fluctuation . . . . .	110
17.8 The shape of the SP . . . . .	110
17.9 Two limiting cases . . . . .	110
17.9.1 $m = 1$ . . . . .	112
17.9.2 $m \rightarrow \infty$ . . . . .	112
17.9.3 The marginal posterior distribution of $\mu[\mathbf{y}]$ . . . . .	113
17.9.4 The marginal posterior distribution of $s[\mathbf{y}]$ . . . . .	113
17.9.5 Normal Distribution . . . . .	114
17.10 The posterior probability of $\mathbf{y}$ given $\mathbf{x}$ . . . . .	115
17.11 The parametric Lapacian approach . . . . .	115
17.12 The Gaussian hypothesis . . . . .	116
<b>18 Multidimensional linear model</b>	<b>118</b>
18.1 Generalizing the one-dimensional linear model . . . . .	118
18.2 Multivariate posterior . . . . .	118
18.2.1 Multivariate real data . . . . .	118
18.2.2 Generalizing Jeffreys' prior to the multivariate case . . . . .	119
18.2.3 The limit of the VP as $m \rightarrow \infty$ . . . . .	120
18.2.4 The VP when $m = 1$ . . . . .	120
18.2.5 The shape of the VP . . . . .	121
18.2.6 Eliminating variables by marginalizing . . . . .	121
18.3 Regression lines . . . . .	121
18.3.1 The posterior distribution of regression coefficients . . . . .	124
18.3.2 The marginal distribution of $S[Y]$ . . . . .	124

18.3.3	The asymptotic conditional posterior distribution of the regression coefficients . . . . .	125
18.3.4	Regression coefficients in the parametric Laplacian approach	126
18.3.5	The VP for $m = 1$ is asymptotically normal as $n \rightarrow \infty$ . . . . .	126
18.4	Operational Laplacianism in the multivariate model . . . . .	127
18.4.1	Smoothness and information compression in the multivariate model . . . . .	127
18.5	Fragility of linear regression . . . . .	129
<b>19</b>	<b>Mathematical Appendix</b>	<b>130</b>
19.1	General remarks . . . . .	130
19.2	Normalization and expectations of the Relative Entropy Distribution . . . . .	130
19.3	The sample covariance matrix . . . . .	131
19.3.1	The information in a sample . . . . .	131
19.3.2	The covariance matrix is determined by $r(r - 1)/2$ rotations and $r$ stretches . . . . .	132
19.3.3	The covariance matrix is determined by a lower triangular matrix . . . . .	132
19.4	Two key matrix facts . . . . .	132
19.5	A helpful integral . . . . .	133
19.6	Change of variables in integrals . . . . .	133
19.7	Samples with given $\mu$ and $S$ . . . . .	133
19.8	SP and SSP . . . . .	135
19.8.1	The marginal distributions of the SP . . . . .	136
19.8.2	Properties of the SSP . . . . .	136
19.9	Properties of the VP . . . . .	137
19.9.1	The normalizing constant for the VP . . . . .	137
19.9.2	The marginal VP for $S_Y$ . . . . .	140
19.9.3	The VP is a matrix- $t$ distribution . . . . .	140
19.10	Invariance and the generalization of Jeffreys' prior . . . . .	141
19.10.1	The limit of the VP as $m \rightarrow \infty$ . . . . .	141
19.10.2	The limit of the marginal distribution of $S_Y$ as $m \rightarrow \infty$ . . . . .	142
19.10.3	The conditional distribution of the regression coefficients	143
19.10.4	Marginalizing the VP . . . . .	145
	<b>References</b>	<b>146</b>

# Chapter 1

## Interpretations of probability and statistics

### 1.1 Models and interpretations

While there is little disagreement about the formal definition and properties of probabilities, there is considerable disagreement about the *interpretation* of the probability model, and the derivation of the formal properties of probability from an underlying model. This disagreement is important, because it influences how people use probability analysis, and what they think it means and can prove.

The *classical* or *frequentist* interpretation sees probabilities as objective properties of the external world, which manifest themselves in the relative frequency of occurrence of different outcomes of random processes. In this view, a change in the informational status of the observer can have no influence on the probabilities governing the phenomenon being observed. The classical approach flatters itself as being “objective”.

The *Bayesian*, or *Laplacian* interpretation sees probabilities as a description of the beliefs of an observer who has some specific information about the observed system. These beliefs may be influenced, or indeed, determined in some situations by the relative frequencies of events the observer has observed, but an observer may have a probability system over an event which she knows will occur only once. A change in the observer’s information will in general change her probability system. Since the Laplacian approach puts the observer at the center of the formation of probabilities, it has a “subjective” element.

But the distinction between “subjective” and “objective” is somewhat deceptive. Hegel reminds us that subjective and objective are different aspects of the same dialectical unity. In the end the classical probabilist cannot avoid making subjective judgements about what observations to count as arising from a given system, and hence contributing to the observed frequencies that characterize that system. The Laplacian probabilist wants to use evidence to convince others of the validity of her probability judgements, and thus is driven to seek

an inter-subjective consensus on probabilities that has an objective aspect.

The most coherent way to understand the logic of probabilistic and statistical arguments is to start from the Laplacian point of view that explicitly puts an observer in possession of explicitly defined information in the center of the theory. It is possible from this starting point to give a satisfactory and transparent account of frequentist methods and results. Basically, the argument is that frequentist methods are Laplacian methods derived from particular assumptions about the informational state of the observer. One chief aim of this book is to make explicit exactly what implicit assumptions about the informational state of the observer support frequentist analysis.

## 1.2 Some history

The formal mathematical study of probability began in the 17th century with the practical question of how much to pay for various bets in games of chance (See Hacking [1975] for a graceful account of these early developments). This problem neatly combines the frequentist and informational approaches. A bet expresses the opinion of the bettor, for example, as to the relative qualities of horses in a race, an opinion which depends in part on the information the bettor has about the situation, for example, the past performances of the horses. Thus both the observed frequency of events (how often a die has come up with a six, or how often a horse has won a race) and the opinion and information of the bettor (who may know that the die is loaded, or that the horse is drugged) play a role in the willingness of the bettor to wager.<sup>1</sup>

The mathematical theory of probability reached a mature state of development over the 18th century, culminating in the synthetic work of Pierre Simon, Baron de la Place. Laplace, citing the then obscure work of Thomas Bayes, argues that from a mathematical point of view the most that can be expected of a bettor is that her bets based on various states of information be consistent with each other. This requirement implies that the bettor starts with a consistent prior assignment of probabilities over all contingencies; the role of new information is simply to restrict the range of contingencies she regards as relevant, that is to move from a prior joint probability distribution over all the possible outcomes to a conditional distribution that depends on the information she actually has.

The strength of Laplace's point of view is its logical clarity. Once we accept the need for a prior assignment of probabilities, the incorporation of new information (for example, further observations of a process) becomes logically trivial (although the mathematical computations involved can be subtle and

---

<sup>1</sup>The willingness of a real human bettor to wager also depends in practical situations involving uncertainty on "secondary satisfactions" such as pleasure in betting, or difficulty in making commitments in the face of uncertain outcomes. In the interests of focusing on the logic of probability theory and statistics I will generally abstract from these problems, without suggesting in any way that they are unimportant practically. See Pope [2001] for a thorough discussion of this issue.

involved). The weakness of Laplace's point of view is that bettors with different prior assignments of probability will in general have different opinions even when confronted with the same observational information. This possibility is troubling to those who are heavily invested in a positivist/empiricist philosophy of science that imagines the growth of human knowledge as merely a process of revelation of an objective structure immanent in the external world. The judgments arising from Laplace's method appear to have a subjective content incompatible with a belief in the objectivity of scientific knowledge. Laplacians take the position that objectivity is better understood as agreement or consensus. The objective scientist is one whose prior assignment of probabilities can be regarded as neutral among the available hypotheses, or more precisely, the objective scientist has a prior assignment of probabilities representing complete lack of prior information about the process under study. All objective scientists will thus share the same prior assignment of probabilities and arrive at the same conclusions. The weak point of this argument is that it has proved difficult to propose prior assignments of probabilities that command universal, or even widespread, scientific assent as representing a complete lack of prior information.

The subjective probability of an outcome can be thought of as the price a bettor will pay or accept for a ticket paying \$1 just in case the outcome obtains. The theory of probability as a theory of betting concerns the processes by which an idealized bettor (who, for example, is immune from the effects of secondary satisfactions such as pleasure or anxiety about betting itself) might formulate such offers and the impact of changes in her information on her willingness to make or accept offers. In the case of games like roulette or dice the bettor confronts a device or process, which is contrived to produce regular frequencies of various outcomes over long series of trials. The bettor's willingness to pay for or sell tickets in these situations is naturally focussed on these frequencies, since if the stakes are small in relation to her wealth, it is plausible to assume that an idealized bettor will pay or accept a price for a bet on an outcome proportional to its frequency. In these situations it is easy to project the concept of probability as expressing the bettor's willingness to buy or sell tickets into the device itself, and to view probability as a propensity or tendency of a device to produce various outcomes.

In the nineteenth and early twentieth centuries George Boole, Karl Pearson, and R. A. Fisher sought to move observed frequencies, rather than the opinion of a bettor or observer to the central place in the theory of probability. Their work was the foundation for the "classical" or "frequentist" theory that posits the existence of a "true process" generating data that exhibit constant frequencies of various outcomes in long series of observations. The problem of the statistician according to this line of reasoning is to "estimate" a mathematical model of this process on the basis of a finite set of observations. This classical statistical approach appears to avoid the problem of agreeing on a prior assignment of probabilities over outcomes that haunts Laplacianism, although it substitutes a suspiciously similar problem, the need to agree on the class of mathematical models that will be considered. (See Porter [1986] and Stigler [1986] for the his-

tory of these developments.) Unfortunately, as we will see below, this approach is riddled with conceptual and logical inconsistencies, which can be traced back to its failure to acknowledge the necessity of a consistent prior assignment of probabilities in order to make coherent probability assessments. On the other hand, we will also see that the actual methods developed by the classical statisticians can be given a coherent interpretation on a rigorously Laplacian basis.

Classical statistics has given rise to a group of very widely used and very influential statistical procedures, including the measurement of the “statistical significance” of the difference of outcomes from controlled experiments, the practice of “accepting” or “rejecting” hypotheses on the basis of statistical analysis of data, and the analysis of correlation in experimental and historical data through “linear regression” methods. These methods I call “common statistical practice.” One of the themes of this book is that these procedures, each of which is flawed and incoherent in its usual presentation, express a logic which is more robust than the classical theory that is commonly used to support them.

In the middle years of the 20th century a group of philosophically minded mathematicians and physicists, including Bruno de Finetti, Leonard Savage, Frank Ramsey, Harold Jeffreys and Edwin T. Jaynes, revived the Laplacian point of view. The work of these scholars and their followers constitute a devastating critique of classical frequentist logic and method, without, however, completely resolving the problem of finding a universally acceptable prior assignment of probabilities to represent the objective, neutral or completely uninformed observer. This problem is a chief preoccupation of the discussion in this book.

Savage and Ramsey put forward a radically strengthened version of Laplace’s theory in the form of a set of behavioral axioms they claimed to represent “rationality.” Laplace’s axioms require a bettor to have a consistent assignment of probabilities over all contingencies, but the Savage/Ramsey theory requires the rational decisionmaker to have not only consistent probabilities over outcomes, but also a consistent system of preferences over lotteries in outcomes. The Savage/Ramsey theory has come to be called “Bayesian”. In an effort to avoid unnecessary confusion, I will distinguish between the “Laplacian” position that requires the bettor to have a consistent prior assignment of probabilities, and the “Bayesian” position that also requires a consistent system of preferences over lotteries in outcomes. A Bayesian is, *a fortiori*, a Laplacian, but a Laplacian need not be a Bayesian according to these definitions.

### 1.3 Toward a unified interpretation of probability

The purpose of this book is to explain the analytical results reached by the theories of probability and statistics from a unified Laplacian philosophical, mathematical, and scientific point of view. It reports the results of my own (and undoubtedly idiosyncratic) attempt to reconstruct probability theory and

statistical practice on a consistent, logical, workable, and, above all, teachable basis. The spirit of the book is not to deal exhaustively with all possible objections to the interpretation proposed, but to set the theories of probability and statistics out in positive terms that are pedagogically effective.

Much of this book repeats and amplifies the ideas of Laplace [1995], Jeffreys [1939], de Finetti [1974] and Rosenkrantz [1989], Janes and G. Larry Bretthorst (ed.) [2003], but I propose some significant innovations and modifications to their positions. The most important innovation is the elimination of the concept of underlying unobserved parameters in statistical models, which I hope to show is an unnecessary encumbrance of the theory. I call this approach “non-parametric Laplacian”. It provides a logically satisfactory and complete development of probability theory and statistical method which I hope will contribute to dispelling the fog of confusion that surrounds the application of statistical technique. I believe that this approach can also alleviate the anxiety that currently entangles many students of probability and statistics, when they are taught conventional approaches which mystify and complicate the logic of statistical analysis unnecessarily. The social and political importance assumed by probabilistic and statistical arguments in the current era argues strongly for a transparent and universally accessible pedagogy in the field.

My point of view is at heart Laplacian, in that I see probability statements as expressing the state of information of some particular observer of a system, not a property of the system itself. But I will show that many widely employed “classical” statistical techniques which are viewed as “objective” can be interpreted as rigorous applications of the Laplacian theory with specific but intuitively persuasive prior probabilities. Thus my first conclusion is that the long debate between classical and Laplacian approaches to probability and statistics is largely irrelevant from a scientific point of view, since they lead in practical situations to the same results. This reformulation also puts common practice statistical techniques on a transparent and logical foundation.

This re-interpretation, however, reveals other, more vexing, problems in the application of statistical methods to scientific problems. These issues arise from a recognition of our inability to use the statistical analysis of observed data to decide between competing scientific theories. This fundamental point has been fudged both in the development of statistical theory and in common applied statistical practice. The traditional exposition of statistical technique, which begins with a “true model” that is presumed to have “generated” the data suggests by a rhetorical sleight-of-hand that statistical analysis of data can recover the “truth”. In fact, statistical analysis can only refine opinions already held on other grounds. I will argue that these other grounds do not lie so much in our observation of the world as in certain deep human perceptual structures, for example, our gestalt preference for smoothness, symmetry, and self-similarity in the patterns through which we observe and explain the world.

This investigation reveals a further curious feature of the traditional rhetoric of statistics. In fact, given the limitations of human life, we always confront a finite number of observations of the systems we study. Furthermore, we know that we and our successors will always have only a finite (though perhaps much

larger) set of observations on which to base our judgments. Nonetheless, classical theories of probability and statistics make constant reference to infinite sets of potential observations, whose properties are viewed as the “true” state of affairs. This device makes the smoothness which we seek in our explanations appear to be a feature of an external reality. In many situations it does not make much difference whether we explicitly acknowledge our own preference for smoothness as the basis of our judgments, or project this preference into an imaginary “objective” reality. But in other situations the traditional procedure can lead to serious confusions, both in obscuring patterns that we could perceive in data, and in suggesting that data can support conclusions that by its very limitations it can not speak to. Much traditional statistical practice makes data analysis the prisoner of unexamined prior assumptions which the investigator adopts not because of her considered appraisal of their relevance to the situation she confronts, but because they are deeply embedded in tools she has been taught to see as “objective” methods of research.

A correct understanding of these points leads to a statistical practice which is both less ambitious and more effective than traditional methods: less ambitious in that it recognizes that data analysis cannot settle theoretical differences; more effective in that it is more open to the revelation of significant patterns in the data.

## 1.4 Mathematics

The development of the mathematical details of the theory is carried out in some depth, and requires the reader to be familiar with certain mathematical tools. The first part of the book, on the theory of probability, requires mathematical tools at the level of a course in analytic geometry, especially the graphical presentation of data and relationships. The second part, on the problem of observations of processes with a finite set of possible outcomes, requires elementary combinatorial ideas and elementary calculus. The third part of the book, on the statistical analysis of real variables and linear regression, requires mathematics at the level of college courses in calculus and linear algebra. Those specific mathematical results that are critical in each section are reviewed in mathematical appendices.

The conceptual development is largely independent of these mathematical details.

## Chapter 2

# Systems of probabilities

### 2.1 Mathematical probability: finite event spaces

This chapter sets out the basic structure of the mathematical theory of probability. The mathematics of probability require only the concepts of elementary arithmetic. A minimal context for these definitions is the idea of an *experiment* (a throw of dice, or spinning of a roulette wheel, or a clinical trial of a drug) that has various possible outcomes. Statistics concerns experiments that can be repeated.

#### 2.1.1 Primitive events and probabilities

Suppose we have a finite set of  $n$  *outcomes*, or *primitive events*,  $\mathcal{E} = \{e_1, \dots, e_n\}$ , the *event space*. From a purely mathematical point of view, a system of *probabilities* over the event space  $\mathcal{E}$  is a corresponding list of  $n$  non-negative numbers  $\{p_1, \dots, p_n\}$  that add up to 1:

$$p_i \geq 0, \forall i = 1, \dots, n \quad (2.1)$$

$$\sum_{i=1}^n p_i = 1 \quad (2.2)$$

We call  $p_i$  as the *probability* of the corresponding event  $e_i$ , and write  $p[e_i] = p_i$ .<sup>1</sup>

#### 2.1.2 Compound events and probabilities

A *compound event*, which we will for brevity call an *event*, is a subset of the event space. We will denote compound events as  $A, B, \dots$ . The probability of

---

<sup>1</sup>In this book I will use square brackets to indicate functional dependence and parentheses to group expressions. Thus  $p[e]$  is the probability of event  $e$ , while  $p(a + b)$  is the product of the number  $p$  with the sum  $a + b$ .

an event  $A$  is just the sum of the probabilities assigned to the primitive events that belong to  $A$ :

$$P[A] \equiv \sum_{e_i \in A} p[e_i] \quad (2.3)$$

The *complement* of an event  $A$ ,  $\bar{A}$ , is the set of primitive events that are not contained in  $A$ . Equation 2.2 implies that:

$$P[A] + P[\bar{A}] = 1 \quad (2.4)$$

The *intersection* of two events  $A$  and  $B$ ,  $A \wedge B$ , is the set of primitive events that belong to both:

$$A \wedge B = \{e_i \mid e_i \in A \text{ and } e_i \in B\}$$

The *union* of compound events  $A$  and  $B$ ,  $A \vee B$ , is the set of primitive events that belong to one or the other:

$$A \vee B = \{e_i \mid e_i \in A \text{ or } e_i \in B, \text{ or both}\}$$

The probabilities of the intersection and union of compound events are related to the probabilities of the compound events by the relation:

$$P[A \vee B] = P[A] + P[B] - P[A \wedge B] \quad (2.5)$$

$P[A] + P[B]$  counts the events in  $A \wedge B$  twice, so the probability of  $A \wedge B$  has to be subtracted in order to arrive at  $P[A \vee B]$ .

A more symmetric way to express this relation is:

$$P[A \vee B] + P[A \wedge B] = P[A] + P[B] \quad (2.6)$$

### 2.1.3 Joint and conditional probabilities

The probability  $P[A \wedge B]$  is called the *joint* probability of  $A$  and  $B$ . If  $P[B] \neq 0$ , the *conditional probability* of  $A$  given  $B$ ,  $P[A \mid B]$  is the ratio:

$$P[A \mid B] \equiv \frac{P[A \wedge B]}{P[B]} \quad (2.7)$$

The conditional probability  $P[A \mid B]$  expresses the probability of the primitive events in  $B$  that are *also* contained in  $A$  as a fraction of the whole probability assigned to the event  $B$ .

When  $P[B] = 0$  the conditional probability is undefined. From the definition of conditional probabilities, it is clear that when they are well defined:

$$P[A \mid B]P[B] = P[A \wedge B] = P[B \mid A]P[A]$$

In the case where  $P[A]$  and  $P[B]$  are both positive, this relation implies *Bayes' Theorem*:

$$P[A | B] = \frac{P[B | A]P[A]}{P[B]} \quad (2.8)$$

This relationship, as we will see, is the foundation of all statistical methods, since it expresses the impact of information (the fact that event  $B$  has occurred, for example) on the probability of  $A$ .

In Bayesian or Laplacian probability theory  $A$  is often called the *hypothesis*, and  $B$  is called the *evidence*. Laplacians refer to  $P[A]$  as the *prior probability* of  $A$ , since it represents the probability judgement of an observer who has not seen the evidence  $B$ , and to  $P[A|B]$  as the *posterior probability* of  $A$ , since it represents the probability judgement of an observer who started with the prior  $P[A]$  and then used the evidence  $B$  to modify her opinion.  $P[B|A]$  is the *likelihood*, representing the probability of the evidence  $B$  on the hypothesis  $A$ .  $P[B]$  is basically a normalizing term, to make the posterior probability a proper probability distribution.

### 2.1.4 Exercises

**Problem 2.1** The event space is  $\mathcal{E} = \{e_1, e_2\}$ . List all the compound events and calculate their probabilities on the assumption that  $p[e_1] = 1/3$ .

**Problem 2.2** The event space is  $\mathcal{E} = \{e_1, e_2, e_3\}$ . The events  $A = \{e_1, e_2\}$ , and  $B = \{e_2, e_3\}$ . Calculate the probabilities  $P[A]$ ,  $P[B]$ ,  $P[A \wedge B]$ ,  $P[A \vee B]$ ,  $P[A | B]$ , and  $P[B | A]$  on the assumption that  $p[e_1] = 1/6$  and  $p[e_3] = 1/3$ .

**Problem 2.3** An epidemic disease afflicts 1% of a country's population. The disease can be treated in its early stages when it, however, produces no directly observable symptoms. The best test for the disease has a false-negative rate of 5%, that is 5% of the people who have the disease test negative, and a false-positive rate of 2%, that is 2% of the people who do not have the disease test positive. What is the probability that a person who tests positive actually has the disease?

## 2.2 Infinite event spaces\*

In practical fact, we always deal with finite sets of events. Even if the variables that define events are theoretically continuous, such as temperature, any real measuring system will have a limited precision, so that we can in principle divide the space of possible observations into a finite (though perhaps very large) number of possible events. The theory of probability set out in the section 2.1 is all that we need in principle to deal with any real measurement situation.

The direct calculation of probabilities over complex event spaces, however, often involves a large amount of computation. The widespread availability of powerful computers makes a wide range of these calculations feasible. But there are practical problem situations where the number of events is so large that even the fastest computers cannot calculate the required probabilities in

a reasonable time. This, in fact, was the typical situation before the advent of computing machinery, when the direct calculation of even relatively small systems of probabilities was infeasible given the computing resources available (basically pencil and paper).

To deal with this computational difficulty, ingenious mathematicians devised shortcuts that allowed them to approximate the direct calculation of probabilities in special cases. These methods often involved approximating the finite event space defined by the actual problem at hand by a fictional infinite event space on which the required calculations could be made more easily.

In the simplest cases, the finite event space was approximated by an infinite but countable primitive event space  $\mathcal{E} = \{e_1, e_2, \dots\}$ . In this case mathematicians defined a system of probabilities as a sequence of nonnegative numbers  $\{p_1, p_2, \dots\}$  with the property that the limit of the partial sums of the sequence converged to 1. Even in this apparently simple generalization of the basic model of probability a number of technical difficulties and potential paradoxes arise. There are questions as to the order in which the infinite sum of probabilities should be evaluated. The probabilities of compound events which are finite subsets of the event space are unambiguously defined, but the definition of probabilities of compound events that are infinite subsets of the event space can give rise to paradoxes.

An even more complicated mathematical system arises when the event space is regarded as a continuum, analogous to the real number system. In this situation it becomes impossible to maintain a consistent definition of primitive events with positive probabilities. Instead, mathematicians adopted the practice of assuming the laws of probability 2.4 and 2.5 directly for systems of events defined as subsets of the event space. The difficulty here is to define a set of events on which these laws of probability can be unambiguously defined. The *Kolmogorov* system is based on the concept of a *sigma field* of events, which is a set of subsets of the event space that contains the complement of each of its members, and the union of any set of its members and the intersection of any finite set of its members. In the Kolmogorov system a system of probabilities is a function whose domain is the sigma field of events and whose range is the nonnegative real numbers, that satisfies the properties 2.4 and 2.5. A very frequently used probability system is the one defined on the real line by the sigma field is the *Borel sets*, the sets generated by the union and intersection of closed intervals. But the reader should be cautious about the applicability of this very special model to real-world problems of probability and statistics, especially those arising in the social sciences. The difficulty is that the real line has a number of peculiar special properties which do not have a natural analog in real social situations. In economics, for example, there is a temptation to identify households or firms in very large economies with points on the real line, and use the Borel sigma field to define measurable allocations of goods to these economic agents, a procedure that can easily lead to paradoxes Khan and Sun [1996, 1999].

The philosophy of this book is to avoid both the technical mathematical intricacies and the potential ambiguities and paradoxes inherent in the use of

infinite event spaces. In those situations where it is impossible to avoid infinite event spaces, I will develop the requisite mathematical tools explicitly in as elementary a fashion as possible.

## Chapter 3

# Probability distributions

In many practical applications events can be defined quantitatively as an integer or a subset of real numbers. For example, the number of people in a household is an integer, and the temperature at noon in New York can be in various 1-degree ranges, such as 50-51 degrees, or 52-53 degrees. Even purely qualitative observations are often coded as numbers: a 1 might represent a "yes" response on a questionnaire and a 0 a "no". In these situations it is often helpful to visualise the system of probabilities as a graph with probabilities on the vertical axis and the numbers representing the events on the horizontal axis.

### 3.1 Representing probabilities

#### 3.1.1 Probability mass functions

When the event space is an integer or a finite set of intervals on the real number line we can visualize a probability system as a function from the real numbers to the interval  $[0,1]$ . Such a function is called a *probability mass function*,  $P[x] : R \rightarrow [0, 1]$ . Figure 3.1 illustrates a probability mass function assigning an equal probability to the two integer events 1 and 0 in a single observation of a process, so that  $P[0] = P[1] = .5$  and  $P[x] = 0$  for  $x \neq 0$  or 1.

Figure 3.2 illustrates a probability mass function assigning various probabilities to the number of 1's observed in 10 observations of an experiment whose outcome is either a 1 or a 0.

#### 3.1.2 Probability density functions

When the number of outcomes becomes very large it is often convenient to approximate a probability mass function by a continuous function, called a *probability density function*. The probability density function  $P[x]$  assigns a probability  $\int_a^b P[x]dx$  to the interval  $(a, b)$ . By the mean value theorem of calculus, this means that the probability density function assigns a probability approximately equal to  $P[x]dx$  to the interval  $(x, x + dx)$ .

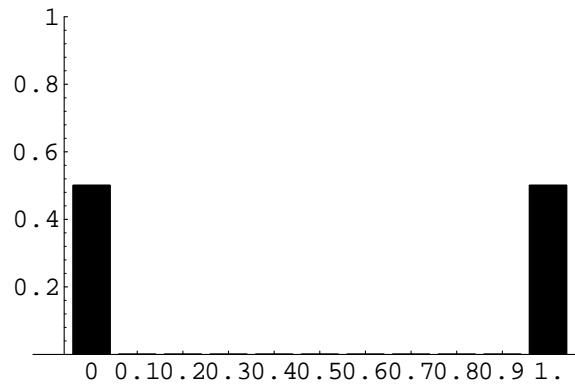


Figure 3.1: A probability mass function that assigns a probability of .5 to the outcome 1 and a probability .5 to the outcome 0.

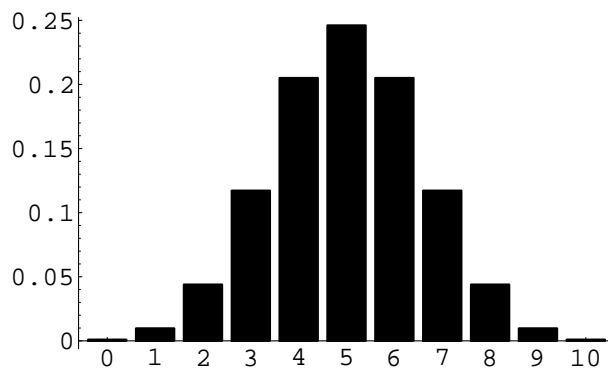


Figure 3.2: A probability mass function assigning various probabilities to the outcomes  $x = 0, 1, \dots, 10$ .

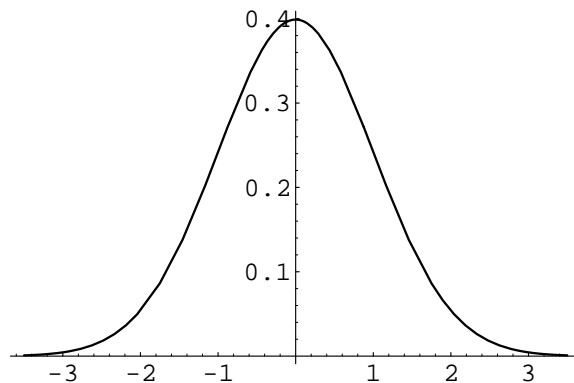


Figure 3.3: The standard normal probability density.

Figure 3.3 illustrates the familiar “standard normal”, or “Gaussian” probability density function

$$P[x] = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$$

(In this book the normal probability distribution will play a role only as a convenient approximation to other probability distributions.)

The normal probability distribution can be relocated by subtracting a constant  $\mu$ , called the *mean*, and rescaled by dividing by a constant  $\sigma$ , called the *standard deviation*. After this transformation the probability density function for  $x$  is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(1/2)\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

This assignment of probability is called the *normal distribution with mean  $\mu$  and standard deviation  $\sigma$* .

### 3.1.3 Probability distributions

An assignment of probability over outcomes, however it might be described, is called a *probability distribution*.

### 3.1.4 Normalizing probabilities

The most important aspect of a probability distribution is the *relative weight* it places on various outcomes. Thus we can describe a probability distribution by any assignment of nonnegative weights to the events, and if we multiply these weights by any nonnegative number we leave the relative weights and therefore the underlying probability distribution unchanged. If we take any assignment

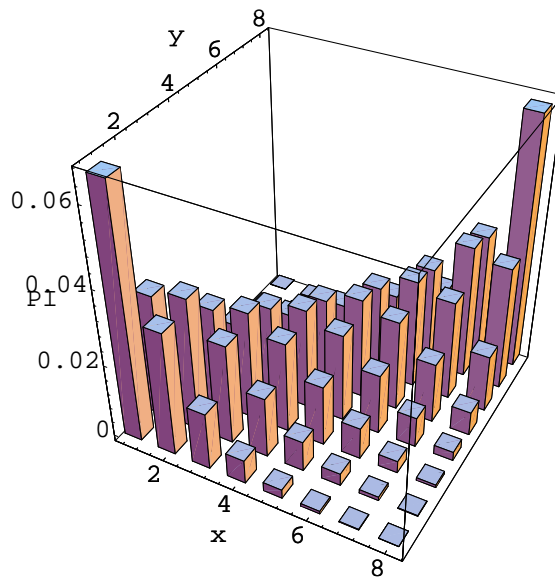


Figure 3.4: A joint probability mass function assigning probabilities to vector observations  $(x, y)$ , where  $x$  and  $y$  can take on values from 1 to 8.

of nonnegative weights to events, we can add the weights up over all the events, arriving at a *normalizing factor* for that assignment of weights. If we then divide all the weights by the normalizing factor the resulting weights will add up to one, and will be a probability mass or density function.

## 3.2 Joint probability distributions

When a process generates data on more than one, say  $r$ , dimensions, the events are an  $r$ -dimensional vector whose elements are the various dimensions of the outcome. In this case the probability mass function is a function of several variables,  $P[x_1, x_2, \dots, x_r]$  and is called a *joint probability mass function*. Figure 3.4 illustrates a joint probability mass function over two outputs, one with possible outcomes  $x = 1, 2, \dots, 8$ , and the other with outcomes  $y = 1, 2, \dots, 8$ .

### 3.2.1 Conditional probability distributions

In cases where the outcomes are described by more than one variable, so that we have a joint probability distribution, we are often interested in the probability distribution for one of the variables, or a subset of them, holding constant the others at some particular value. Such a distribution is called a *conditional*

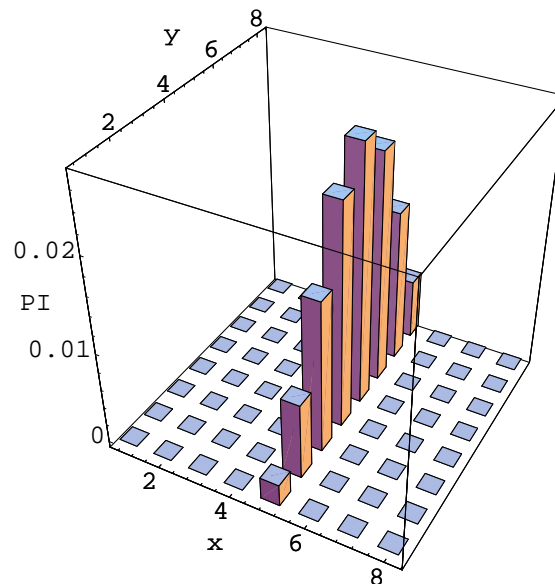


Figure 3.5: The conditional probability distribution for  $y$ , given  $x = 4$ , derived from the joint probability distribution in Figure 3.4.

*probability distribution.* If we have a joint probability  $P[x_1, \dots, x_r]$ , then the conditional distribution for the first  $s$  variables, given specific values to the last  $r-s$  variables is just  $P[x_1, \dots, x_s, x_{s+1}^*, \dots, x_r^*]$ , where  $x_j^*$  represents the specific value of variable  $x_j$  at which we are calculating the conditional probability, appropriately normalized to be a probability distribution.

Figure 3.5 illustrates the conditional probability distribution derived from the joint probability distribution in Figure 3.4, holding  $x$  constant at 4.

### 3.2.2 Marginal probability distributions

We often want to know the probability distribution for one subset of variables for *all* the possible values of the other variables. This *marginal probability distribution* can be calculated by summing or integrating the joint probability distribution for each value of the selected variables over all the possibilities for the other variables. This process is conceptually simple, but often leads to mathematical complexities when the joint probability distribution is described as a formula. In principle, however, a computer can easily calculate marginal probabilities from joint probabilities, simply by adding up terms.

Figure 3.6 illustrates the marginal probabilities for  $y$  corresponding to the joint probability distribution in Figure 3.4.

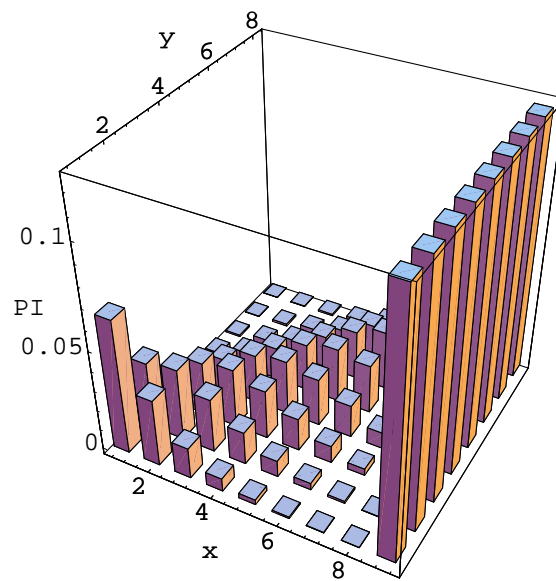


Figure 3.6: The marginal probabilities for  $y$ , together with the joint distribution for  $x$  and  $y$  from Figure 3.4. In this case the marginal probabilities are uniform over the 8 possibilities  $y = 1, \dots, 8$ .

### 3.3 Simplifying probability distributions

Marginalization and conditionalization are the two fundamental strategies for simplifying joint probability distributions so that they express states of information in a useful form. A joint probability distribution often conveys more information than we need to answer a question, and often in a confusing form.

The most common situation is where the joint probability distribution is defined over a large number of variables, but we are concerned with the relation between a smaller number. (For example economic analyses of demand and supply relationships have to be based on data including income, cost factors, taste factors, and so forth, but in the end are aimed at identifying the relation between price and quantity demanded or quantity supplied.) We may want to eliminate variables either because we already know what value they take in situations of interest, or because we have no information about what value they take.

In the first case conditionalization is the appropriate approach. We set the values of the variables we know (or can control) at the appropriate levels, and use the resulting conditional probability distribution for the analysis.

In the second case marginalization is the appropriate approach. We sum or integrate over the variable we have no information on, and use the resulting joint probability distribution over the variables we do care about. Marginalization eliminates the dependence of the probability of outcomes on the marginalized variables by “averaging out” over their possible values, which is the best we can do if we have no information at all about the eliminated variables. If, of course, we actually have relevant information about a variable, then we should not marginalize, but use that information to conditionalize.

Both marginalization and conditionalization reduce the dimension of the joint probability distribution with which we have to deal, which is a tremendous simplification. They do this in different ways. Marginalization eliminates variables by attributing to any configuration of the other variables the sum of the probabilities over all the possible values of the marginalized variable. Conditionalization eliminates a dimension of the joint probability by fixing the value of the corresponding variable at a particular level.

It is possible to use conditionalization and marginalization to represent partial information about a variable. If we know that a variable must lie in some range, we can marginalize over the range and conditionalize out the irrelevant cases.

### 3.4 Convolutions of probability distributions

Suppose we have two probability systems, the first of which,  $x$  has  $n + 1$  possible outcomes, the integers  $0, 1, \dots, n$  with probabilities  $p_0, \dots, p_n$ , and the second,  $y$ , has  $m + 1$  possible outcomes, the integers  $0, 1, \dots, m$  with probabilities  $q_0, \dots, q_m$ . Suppose also that these probabilities are independent, so that the probability of observing  $x = i$  and  $y = j$  is just  $p_i q_j$ . In this type of situation

the sum  $x + y$  has  $n + m + 1$  possible outcomes  $0, 1, \dots, n + m$ , and it is often of interest to calculate the probability that the sum has a particular value,  $k$ .

The event that the  $x + y = k$  can occur in several different ways. For example,  $x = 0$  and  $y = k$  is one possibility, or  $x = 1$  and  $y = k - 1$ , or in general  $x = i \leq k$  and  $y = k - i$ . The probability that  $x + y = k$  is thus

$$r_k = \sum_{i=0}^k p_i q_{k-i}$$

This expression is the *convolution* of the two probability distributions  $p_i$  and  $q_j$ .

The analogous expression for two probability density functions  $f[x]$  and  $g[y]$  defined over the real number line  $(-\infty, \infty)$  is

$$h[z] = \int_{-\infty}^{\infty} f[x]g[z-x]dx$$

### 3.5 Problems

**Problem 3.1** *The event space is generated by two observations, A and B, each of which can take on the value 0 or 1. The joint probability mass function is  $P[1, 1] = .0095$ ,  $P[1, 0] = .0005$ ,  $P[0, 1] = .0198$ ,  $P[0, 0] = .9702$ . Verify that this is indeed a probability distribution, and calculate all the conditional and marginal probabilities.*

**Problem 3.2** *Prove that the convolution of two probability mass functions is a probability mass function (that is, a set of nonnegative numbers that sum to 1).*

## Chapter 4

# The economic model of probability

### 4.1 The Laplacian concept of probability

Bruno de Finetti de Finetti [1974] (in agreement with Leonard J. Savage Savage [1954] and Frank Ramsey Ramsey [1950]) defines the probability of a verifiable event to a particular decisionmaker in terms of the odds at which the decisionmaker would bet on or against the occurrence of the event for small stakes. (This idea is evidently already a considerable abstraction from any real-world circumstance of any real decisionmaker. In particular it abstracts from the ever-present problem of “secondary satisfactions” Pope 2001.) We can denote events by  $A, B, \dots$ , the complement of the event  $A$  by  $\bar{A}$  and the probability of an event to a decisionmaker  $I$  by  $P_I(A)$ . The odds against an event are  $P_I(\bar{A})/P_I(A)$ , so that if  $P_I(A) + P_I(\bar{A}) = 1$ , knowledge of the odds against an event is equivalent to knowledge of the probability of the event.

De Finetti (and, to an even greater extent, Savage and Ramsey) spends some effort to persuade his readers that all decisionmakers have probabilities in this sense over events. I do not take this position; in fact, the formation of a probability in this sense in human beings appears to me to be an unusual and remarkable occurrence. Many people have a strong reluctance, almost a conscientious objection, to accepting and offering bets even at small stakes on events of interest. For the moment, however, let us defer the discussion of when and in what circumstances probabilities of this kind get formed, and look more closely at the implications of de Finetti’s definition.

#### 4.1.1 The scope of probability judgments

The language of betting is convenient expositionally, but should not mislead us into ignoring the breadth of cases to which this discussion applies. Interesting human decisions always involve stakes of one kind or another depending

on unresolved contingencies. Betting is thus a metaphor for the more general category of human action.

The relation of betting to practical human concerns is perhaps clearest in the sphere of finance. A wealthholder investing in a portfolio of assets inevitably faces uncertainty about the returns to be expected, and can be viewed as betting on the contingencies affecting these returns (for example, the future rate of inflation, which affects the value of cash and bonds, or the competence or health of key managers of particular firms, which affect the dividends to be expected from holding stock in those firms). But important aspects of almost every human decision can be illuminated by considering them as bets. The choice to commit emotional and material resources to a romantic relationship, for example, is a kind of bet on the success of the relationship. The choice of an educational program, career, or job similarly involve bets on uncertain outcomes. The choice of medical treatment almost always involves significant uncertainties and costs (often only approximately understood by patients and physicians) and is formally similar to a bet. Public policy decisions, from the development of particular military technologies, through land use and environmental regulation, to the penalties assigned for criminal acts, involve the balancing of costs against uncertain benefits, and are a kind of bet. Whatever we learn about bets in the abstract has immediate ramifications in all these spheres of human life.

## 4.2 The economic offer set

It is helpful in understanding de Finetti's approach to borrow from economics the device of visualizing behavior (like betting) geometrically. Let us interpret ordinary Euclidean space in  $m$  dimensions,  $R^m$ , in the following fashion. The dimensions,  $1, 2, \dots, m-1, m$ , we interpret as money contingent on the occurrence of particular events. An event is a description of some aspect of the world which the decisionmaker (and others) can verify practically, such as "Native Dancer won the race," or "the current at the resistor lies between 200 and 210 ma." For simplicity, from this point on let us confine ourselves to the case  $m = 2$ , so that there are only two events of interest, which we will call  $A$  and  $\bar{A}$ ; the arguments can be generalized in obvious ways to the case  $m > 2$ . In this case  $x_1$  represents "dollars to be paid in case the event  $A$  takes place," and  $x_2$  represents "dollars to be paid if the event  $A$  does not take place". We can think of betting on the event as buying  $x_1$  in exchange for  $x_2$ , that is, paying a certain sum of money for an envelope that contains \$1 just in case the event  $A$  occurs, and is worthless otherwise. Similarly, betting against the event is selling  $x_1$  in exchange for  $x_2$ , selling an envelope that contains \$1 just in case the event  $A$  occurs for a certain sum of money. The ratio  $p = \frac{-x_2}{x_1}$ , the price of the envelope, is the amount of money the decisionmaker pays (or receives) for the envelope. This price is the slope of the line connecting the origin to the point  $\{x_1, x_2\}$ .

A bet on an event can be visualized as the exchange of money in the case the event does not occur,  $x_2$ , for money contingent on the occurrence of the event,  $x_1$ . The point  $\{x_1, x_2\}$  represents the bet; the negative of the slope of the line

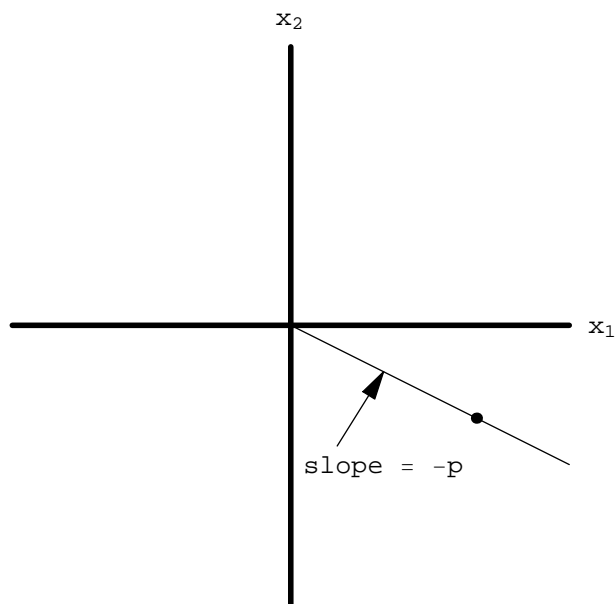


Figure 4.1: The negative of the slope of the line connecting the origin to the point  $\{x_1, x_2\}$  represents the price of  $x_1$  in terms of  $x_2$ .

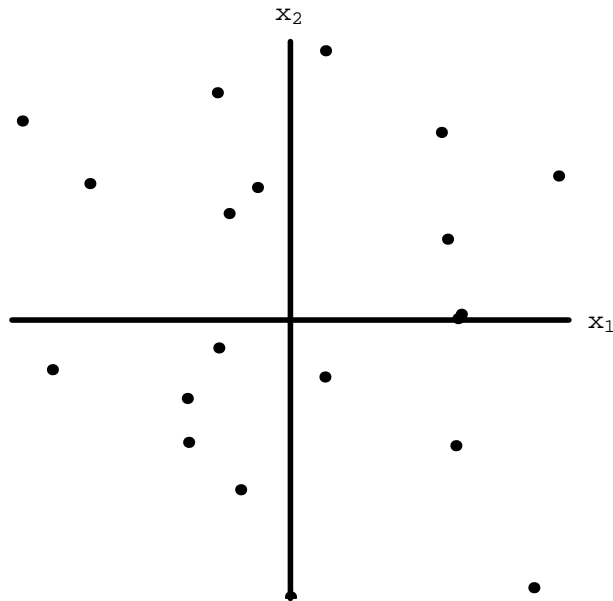


Figure 4.2: An offer set can be thought of as a set of bets a decisionmaker will take.

connecting the bet to the origin is the price,  $p$  paid for the contingent payoff.

The simplest analytical concept to describe behavior in this setting is the *offer set*, a set of bets which the individual will in fact accept. We can visualize the offer set in as in Figure 4.2.

De Finetti's definition of probability assumes that the offer set of an individual has two key properties: a) the offer set must not span the negative orthant ("no Dutch book" in the jargon of betting); b) the offer set boundary must be differentiable at the origin.

### 4.3 No Dutch book

If there are bets in the offer set that span the negative orthant, the decisionmaker will bet against the event at better odds than she will bet for it, so that by calling both bets an adversary could make a certain gain. De Finetti refers to this as the possibility of making "Dutch book" against the decisionmaker. Figure 4.3 illustrates an offer set that is open to a Dutch book.

This offer set indicates a willingness to bet on the event at one price and simultaneously against the event at a lower price. An adversary who calls both bets makes a sure gain, indicated by the point on the line connecting the bets.

It is worth thinking about Dutch book in relation to the broader interpretation of bets as a model of general human decisionmaking. In the case of the

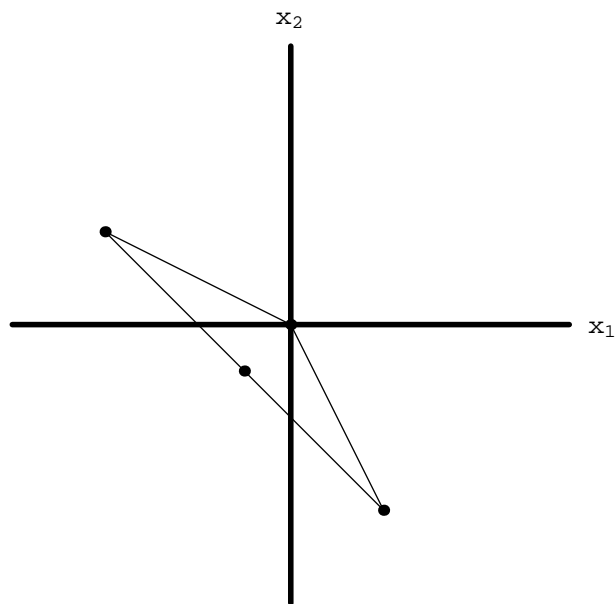


Figure 4.3: If points in the offer set span the negative orthant, the individual is open to transactions that lead to a certain loss.

portfolio investor, an offer set open to Dutch book corresponds to a portfolio which will certainly lose money no matter what the resolution of the uncertainty may be. For example, suppose the investor buys stock in competing companies at prices reflecting her hope that each of their managements will turn out to be above average in competence; since some of the managements must turn out to be below average in competence, the investor is doomed to disappointment. A patient who pursues two incompatible courses of treatment whose effects cancel each other out is also effectively open to a Dutch book. A government that subsidizes polluting industries in the hope of producing jobs and simultaneously pays for the environmental damage they cause is effectively open to Dutch book.

## 4.4 The boundary of the offer set must be differentiable at the origin

The differentiability of the offer set boundary at the origin is required in the development of Laplacian theory to assure the existence of a unique probability for the contingency  $A$ . De Finetti expresses this in arguing that decisionmakers must be prepared to accept infinitely small bets for or against the contingency at the same odds. A convex, continuous preference ordering over the bets will automatically satisfy these axioms, but the axioms could hold without the decisionmaker possessing such an ordering.

When the offer set boundary is not differentiable at the origin, we cannot identify a unique probability to assign to the event. The best we can do is to find an “upper” and “lower” probability representing respectively the lowest price at which the decisionmaker will bet against the event and the highest price at which she will bet on it. The failure of differentiability of the boundary of the offer set at the origin is illustrated in Figure 4.4.

The lowest price at which the decisionmaker will bet against the event is strictly larger than the highest price at which she will bet on it. Her behavior defines a range of probabilities for the event rather than a unique probability.

We can visualize the type of offer set de Finetti has in mind as in Figure 4.5.

De Finetti assumes that the boundary of the offer set is differentiable at the origin, so that there exists a unique tangent plane (or more generally hyperplane) to the offer set at the origin, and that the offer set lies entirely above this unique tangent plane. Notice that the offer set boundary need not be differentiable away from the origin, nor does the offer set need to be convex in order to satisfy de Finetti’s axioms, although convex offer sets with differentiable boundaries will satisfy them.

### 4.4.1 Markets and probabilities

Though de Finetti describes the properties of probabilities in terms of the offer set of an individual bettor, these same properties are inherent in the formation of prices for bets in a market. The no-Dutch book property de Finetti proposes for individual offer sets corresponds in a market setting to the absence of

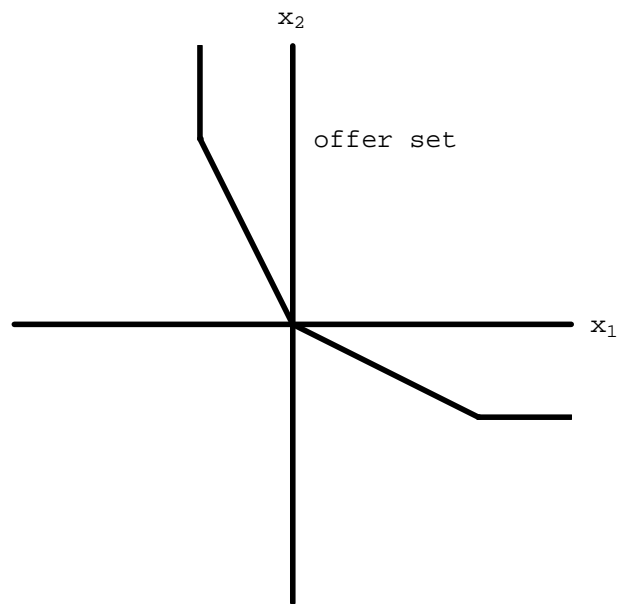


Figure 4.4: If the boundary of the offer set has a kink at the origin, it defines only a range of possible probabilities, not a unique probability for the event.

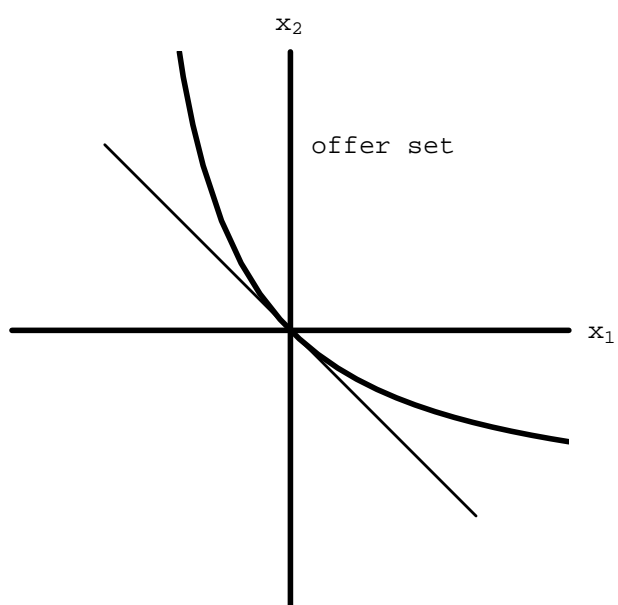


Figure 4.5: To satisfy de Finetti's axioms, the offer set must have a boundary which is differentiable at the origin, and lie entirely above its tangent plane at the origin.

pure arbitrage opportunities at the going market prices. The differentiability of the offer set at the origin, which de Finetti needs to establish the existence of definite unique subjective probabilities, corresponds in the market setting to the existence of definite market prices for contingencies. Of course, the market interpretation is very different from the individual interpretation because market prices do not necessarily reflect the considered judgement of any individual bettor, even if they satisfy de Finetti's requirements for consistent probabilities.

## 4.5 When a decisionmaker has a probability judgment

Let us put off the discussion of the circumstances under which we might expect human beings to satisfy these axioms, and summarize the discussion by saying that we will regard a decisionmaker as having a probability for an event  $A$  if she is willing to bet on and against  $A$  at the same odds for small stakes, and unwilling to take and make bets simultaneously that expose her to a certain loss.

## 4.6 Consistency and reasonableness

It is crucial to recognize that de Finetti's axioms guarantee only the internal consistency of probabilities, not their reasonableness. The axioms prevent someone from paying \$.51 for heads and at the same time \$.51 for tails in a bet on the fall of a coin, but they do not prevent a bettor paying \$.99 for heads as long as she pays no more than \$.01 for tails. The problem of reasonableness is in fact the same as the problem of consensus on a prior assignment of probabilities. The bettor who is prepared to pay \$.99 for heads may know that the toss of the coin is rigged so that heads is almost certain to occur, for example. Other bettors who do not have this information will not agree with her probability, and will sell heads to her.

This scenario raises several questions. First, the bettor's wealth (or the other bettors' estimate of her wealth) may put a limit on how many heads she can buy. If the bets are all cash transactions, the limit will be her actual liquid wealth. But in many situations (in financial markets, for example) transactions are made on credit, so that a bettor can take a position which is long on heads, without having to settle it in cash until after the coin is tossed. There is obviously some limit on the position the market will allow an individual bettor to take, but probability theory itself does not address the determination of this limit.

Second, the presence of a bettor aggressively buying heads may begin to persuade other bettors to be suspicious of the toss of the coin itself. They may infer from the long position someone is taking in heads that she has information they do not have, and this may lead them to change their offer sets, in effect, changing their probabilities. On the other hand, they may think that the long bettor on heads is irrational, or persuaded by incorrect information, or in the grip of

a mistaken theory, and continue to hold their original probabilities. Again, the theory of probability itself cannot address these complex social interactions.

Third, as the long bettor on heads accumulates more and more head bets, her willingness to continue to buy heads may decline. The full Bayesian decision theory of Savage and Ramsey attributes this effect to declining marginal utility of wealth: the bettor with a long position in heads experiences an increasing risk of losing a large amount of money, which begins to weigh more heavily in her actions than the prospect of a symmetrical gain. It may be difficult in practical situations to distinguish the case in which the other bettors are limiting the position of the long bettor by refusing to accept her credit from the case in which the long bettor herself begins to lose enthusiasm for the bet as her position increases.

## 4.7 Probability and evolution

A decisionmaker who has a probability for an event avoids certain kinds of inconsistency and irrationality in her betting. Having a probability is necessary for fully “rational” behavior, understood in the economic sense as acting so as to maximize the expectation of a differentiable utility function, but is a much weaker condition than full rationality.

The most persuasive basis of any theory of human or institutional behavior lies in a demonstration that the behavior is adapted to a particular situation or class of situations in an evolutionary sense. Is there any reason to think that the characteristic of having a probability has survival or reproductive value, and therefore will be selected for by the forces of either social or biological evolution? The most convincing form of such an argument would be a model that would specify the survival and reproductive value of a variety of behaviors, including probabilistically consistent behavior, and then could demonstrate the evolutionary dominance of probabilistic behavior. The failure of recent parallel researches into the evolutionary value of fully rational behavior suggests that it will be difficult to carry out this program.

Appropriate canons of behavior are context-dependent. There are circumstances in which probabilistic consistency has a survival payoff, and in these circumstances we might expect to see probabilistically consistent behavior. But there are other circumstances (many of the ordinary circumstances of life, in fact) where probabilistic consistency does not have decisive survival or reproductive value. In these contexts only much weaker postulates can be supported. If this is true, we should be cautious in assuming probabilistic consistency as a general postulate of behavior, and be careful to specify the reasons why it is a suitable hypothesis in particular contexts.

With this evolutionary viewpoint in mind, let us consider de Finetti’s two axioms in more detail.

## 4.8 The differentiability of the offer set boundary

Human life in fact is finite, and in strict realism, we should always suppose offer sets are finite. It is legitimate, however, to approximate very dense finite offer sets by infinite sets as a methodological procedure. Thus it is not a relevant criticism of De Finetti's definition to insist that the offer set is always finite, and that therefore its boundary cannot possibly be differentiable. It would be enough for the Laplacians if the offer set were very dense near the origin, and if the difference between the worst odds at which the decisionmaker would bet against the contingency and the best odds at which she would bet for the contingency were very small.

But is there any reason to think that human decisionmakers will make small bets at the same odds for and against any event? Or, to put this question in evolutionary terms, is there any general survival or reproductive value in being willing to make small bets on either side of every event? It is hard to see what this value could be. It is true that there is an obvious survival value in being willing to make large bets on certain contingencies. An agent who will not do this effectively condemns herself to inaction in the face of uncertainty, and it is not hard to think of numerous circumstances in which a reluctance to take action (to jump either to the right or the left in the face of the oncoming car, for example) when the stakes are high is disastrous for survival.

The trouble is that the willingness to make bets for large stakes does not necessarily establish a unique and consistent system of probabilities, since the odds at which the agent will take one side of the bet might differ substantially from the odds at which she would take the other. The Bayesians would say such a decisionmaker has an "interval" of probability for the contingency. It isn't clear why the language of probability is any more helpful than the image of the offer set itself in this connection.

## 4.9 Dutch book

The second Laplacian axiom requires that the offer set not span the negative orthant. This is usually phrased by the Bayesians as the condition that probabilities be consistent, and is motivated by the argument that a decisionmaker whose offer set spans the negative orthant will take both sides of a bet at unfavorable odds, guaranteeing herself a loss. In Laplacian jargon it is possible to make "Dutch book" on such an offer set.

Here it is worth reminding ourselves again that betting in this context is a metaphor for human action in general. Decisionmakers whose offer sets permit a Dutch book will act in self-defeating ways in this perspective. But experience and history confirm strongly the propensity of human beings to act in self-defeating ways. This often happens because decisionmakers fail to perceive the connection between their actions that make them self-defeating. In betting language, the self-defeating decisionmaker does not understand, because of the

way the issues are framed, that she is betting for and against the same event at inconsistent odds.

The avoidance of Dutch book does seem to have some survival and reproductive value in an evolutionary sense, since it is not hard to see that exposure to certain loss could be connected in some circumstances to a reduction in the probability of reproduction. An financial trader, for example, who exposes herself to infinitely large certain losses will not be a factor in the market for long. But the “Dutch book” axiom as it is stated seems far too strong to be supported by evolutionary arguments. Suppose, for example, that a decisionmaker exposes herself to bounded, in fact, rather small, certain losses. Is this behavior certain to lower her reproductive probability in all circumstances? If the agent is caught in a certain loss fairly infrequently, say because of a reluctance to take small actions at all, the penalty exacted by the environment may be quite moderate. This penalty might even be offset by the saving of energy that would otherwise be required to police the boundary of the offer set and recognize all the vulnerable areas. This policing might require a considerable amount of costly analysis in many situations (for example, in an environment of extremely complex derivative financial securities), and the effort required might reduce survival value in more important areas more than the gain would be worth.

In reality, of course, we know that lots of individuals are exposed in this way to small certain losses, and that criminals who practice fraud can make a living by identifying and exploiting these individuals. If the Laplacian axiom actually held, there would presumably be no need for laws against frauds that depend on this type of exposure.

But, the Laplacian might say, if you expose yourself to small certain losses, the environment might repeatedly penalize you, leading to a very large loss. This is true, but it is more reasonable to suppose that decisionmakers, when they notice a continual leakage to Dutch books of one kind or another, will take steps to understand the situation better and remove enough points from their offer sets to avoid the problem. This is a hypothesis of rational learning, but it is brought into play by evolutionary arguments only in cases where the environment actually exacts the certain loss over and over again. (This is also the most favorable circumstance for rational learning to take place, since the process in question is repeated many times.) Con men, for example, rarely hit the same sucker twice, and con men are too rare to impoverish every sucker completely.

In particular, evolutionary forces may be very weak in areas where decisionmakers act only once, or a very few times, in their lifetimes. The propensity of people to make disastrous romantic choices, for example, may be penalized severely enough, but if people make only a few such choices in their lives, the learning process may not eliminate the inconsistent behavior very quickly. Furthermore, new participants may not be able to learn very well from their predecessors’ mistakes, so the inconsistent behavior may be a continuing and important feature of the world despite its evolutionary costs.

## 4.10 A tradeoff

From an evolutionary point of view there appears, in fact, to be a tradeoff between De Finetti's two axioms of behavior. One way a decisionmaker might protect herself from Dutch book is to refuse to bet on some events at all, or at least not for small stakes, and to insist on better odds in making bets than in taking them. Laplacian probability will emerge only in the limit when the environmental pressures both strongly reward the willingness of the decisionmaker to make small bets for and against events at the same odds, and strongly penalize the willingness of the decisionmaker to bet on both sides at inconsistent odds. Trading in financial assets, for example, may approximate this environmental situation. The financial trader who is too cautious offers prices that are unattractive compared to her competition. Thus there is strong pressure for her to take both sides at nearly the same price. (Of course, in real financial markets there is always a margin between buying and selling prices, though it may be forced by competition to be quite small.) Similarly, any slip the financial trader makes in evaluating her position that exposes her to certain loss is likely to prove disastrous because of the large sums involved.

## 4.11 Who has probabilities?

What survives from this examination of the Laplacian axioms is the presumption that we should not expect on evolutionary grounds to see decisionmakers who expose themselves to unboundedly large certain losses, or, what amounts to the same thing, to an unbounded repetition of smaller certain losses. But this axiom is much weaker than the no-Dutch book axiom. It is not sufficient to guarantee a consistent probability system for the decisionmaker.

### 4.11.1 Where consistency counts

Our conclusions up to this point might seem very negative for the importance of probability theory and rational decisionmaking. But there are important areas of life where probabilistic consistency has survival value, and where, as a result, the theory of probability must play a central role.

As the example of the financial securities trader suggests, fierce competition in a market setting tends to create environmental conditions that select for probabilistically consistent decisionmaking. Where transactions are frequent and involve lots of money, the stakes at risk are large, and the market effectively searches out and penalizes offer sets that are vulnerable to Dutch books. Furthermore, competition rewards traders who offer to buy at the highest prices and sell at the lowest prices, so that the environment encourages the narrowing of any gap between the odds at which a decisionmaker will bet for and against an event. These forces are not unboundedly effective, of course. Even very successful traders occasionally make mistakes that lose them money on Dutch book, and even the most liquid of markets exhibits a residual gap between buying and

selling prices.

This example suggests that probabilistically consistent behavior emerges adaptively from the pressure of competition. Where the environment is less rigorous, that is, where transactions were less frequent and smaller, and where competitive pressures were less severe, we would expect to see more violations of the axioms of probabilistic consistency, more Dutch book, more refusal to bet at nearly the same odds either for or against events. Probabilistic consistency is an ideal model of behavior, to which we can expect at best an approximation in the real world. The most rigorously consistent trader of financial assets may in other spheres of behavior, like human relations or scientific judgments, where stakes are smaller and transactions less frequent, fall far short of probabilistically consistent behavior.

From this point of view it is no surprise that one of the spheres in which probabilistically consistent behavior appears to play a large role is science. Science is rigorously competitive. Furthermore, the stakes in science are large, whether we consider the subjective stakes of reputation and self-respect, or the objective stakes of research patronage. Effectively the same evolutionary pressures that lead to selection for probabilistically consistent behavior in economic markets operate in science. The practice and success of science, and of scientists, depends on scientists acting consistently in relation to their research. The aim of science is to apply logic to the understanding of reality. A failure to heed the rules of consistency has to defeat this aim. Thus we would expect scientists, in their role as scientists, to approximate the axioms of Laplacian probability theory in their scientific practice. Of course, this presumption need not extend to the behavior of scientists as human beings in other aspects of life, say, in their roles as investors, or lovers, or citizens.

The arguments I have given thus, perhaps paradoxically, support the rigorous deployment of probability theory in the scientific arena, and the social scientific arena, despite our doubts that the hypothesis of probabilistic consistency is a very good ground for the investigation of human social behavior in general. It makes a great deal of sense for scientists to formulate the odds for and against hypotheses in the light of the evidence, whenever they are fortunate enough to be in a position to do so. Successful scientists, in fact, like successful financial traders, are likely to be people who have the faculty of unconsciously but accurately formulating odds in complex situations. It is perhaps particularly hard for social scientists to live with the strictures of probabilistic consistency as scientists without projecting the same behavior on their subjects.

## 4.12 Probability theory is an ideal abstraction

These observations point to the conclusion that Laplacian probability theory is not really a theory of how human beings approach situations involving uncertainty, but an account of how an imaginary ideal decision maker might approach such situations. This ideal decision maker would have an offer set satisfying de Finetti's axioms, and hence, a well-defined system of probabilities for contingen-

cies. The role of probabilistic and statistical arguments, then, is not so much to convince individuals of the correctness of particular arguments, as to function as a benchmark against which real-world decisions as to individual actions and public policy can be tested.

## Chapter 5

# The laws of probability

### 5.1 The linearity of expectations

De Finetti's identification of probability with offer prices for bets on events leads naturally to a simple, general formulation of the general law of probability in terms of the linearity of the expectation (which de Finetti calls *previsions*) of uncertain quantities.

To begin with, let us regard events themselves as quantities, giving the event  $A$  the value 1 if it actually occurs and the value 0 if it does not. In fact, it is intuitively helpful to think of these values as quantities of money, so that the value of an event is \$1 if it occurs and \$0 if it does not. This way of thinking of events links up directly to the image of betting as the buying and selling of envelopes which contain \$1 if the event occurs and \$0 if the event does not occur. With this understanding we can add the values of events.  $A + B$ , for example, is worth \$2 if both  $A$  and  $B$  occur, \$1 if  $A$  occurs and  $B$  does not, \$1 if  $B$  occurs and  $A$  does not, and \$0 if neither  $A$  nor  $B$  occur.

We can then generalize the operator  $P_I$ , which we have so far interpreted in terms of the probability of an event, to become the expectation of an uncertain quantity. Suppose that  $Q$  is a quantity about which we are uncertain. The expectation or prevision  $P_I(Q)$  is the price at which the decisionmaker I would buy or sell an envelope that will contain \$ $Q$ . The prevision of an event is in fact exactly its probability, since  $P_I(A)$  is the price at which the decisionmaker will buy or sell an envelope that will contain \$1 if  $A$  occurs and \$0 if  $A$  does not, which is just the same as  $A$  regarded as an uncertain quantity.

All the considerations we have already mentioned concerning the conditions under which we could assume that a decisionmaker has a probability apply with equal force to previsions. As in the case of probabilities, we will continue the discussion under the assumption that we are dealing with a decisionmaker who has a prevision for a quantity (or, equivalently, a probability for an event).

When we regard events as quantities, the product  $AB$  is exactly equivalent to the logical conjunction " $A \wedge B$ ," since  $AB$  is worth \$1 if both  $A$  and  $B$  occur,

and \$0 if either or both of them fail to occur. “ $A \wedge B$ ” occurs if both  $A$  and  $B$  occur, and fails to occur if either or both of them fail to occur.

The sum of events  $A + B$ , however, is not equivalent to their inclusive disjunction “ $A \vee B$ ,” since  $A + B$  is worth \$2 if both  $A$  and  $B$  occur, whereas “ $A \vee B$ ” is worth only \$1 when both  $A$  and  $B$  occur.

Since de Finetti regards the decisionmaker as equally willing to buy or sell an uncertain quantity, the requirement that the decisionmaker’s offer set allow no Dutch book amounts to the requirement that for any two quantities  $A$  and  $B$  the expectation or prevision be linear:

$$P_I[A + B] = P_I[A] + P_I[B] \quad (5.1)$$

In words, this requirement of linearity of the expectation or prevision amounts to the economically plausible requirement that the decisionmaker should not be willing to sell the sum of the two quantities for less than she will pay for them separately, or, equivalently, should not sell the quantities separately for less than she will pay for their sum. In this form the Dutch book transaction is easy to understand. If we find a decisionmaker who violates 5.1, we can make a certain gain by buying prospects separately from her and selling them back to her for more than we paid, or by buying the sum of the prospects from her and selling them back to her separately and realizing more than we paid for them.

A decisionmaker who has a linear prevision can be said to have coherent views with respect to uncertain quantities. Her previsions of the quantities may be wildly at odds with the judgments of other decisionmakers, but they are not inconsistent among themselves.

The linearity of expectation or prevision leads immediately to the “laws of probability.”

## 5.2 Logical operations and the laws of probability

For example, consider the case where  $B = \bar{A}$ , the negation of an event  $A$ . Then if the value of  $A$  turns out to be \$1, the value of  $\bar{A}$  will have to be \$0, and vice versa, so that in all cases  $A + \bar{A} = 1$ . Understanding this, a decisionmaker will surely pay \$1 for  $A + \bar{A}$ . Then linearity amounts to the probabilistic law that the sum of the probabilities of an event and its negation must be 1:

$$1 = P_I(A + \text{Not } A) = P_I(A) + P_I(\bar{A}) \quad (5.2)$$

We can analyze the probability of the “ $A \vee B$ ,” where  $\vee$  is inclusive disjunction, in similar fashion. The event “ $A \vee B$ ” differs from the sum  $A + B$  because “ $A \vee B$ ” has the value \$1 if either  $A$  or  $B$  or both occur, and the value \$0 if both fail to occur.  $A + B$ , on the other hand, is worth \$2 when both  $A$  and  $B$  are true. How much should a decisionmaker pay for “ $A \vee B$ ?” Clearly something less than  $A + B$ . How much less? “ $A \vee B$ ” is the same as “ $(\bar{A} \wedge \bar{B})$ ,” so that as quantities “ $A \vee B$ ” =  $1 - (1 - A)(1 - B) = A + B - AB$ . Thus linearity implies:

$$\begin{aligned}
 P_I(A \vee B) &= P_I(A) + P_I(B) - P_I(AB) \\
 &= P_I(A) + P_I(B) - P_I(A \wedge B)
 \end{aligned}$$

### 5.3 Coherence and probability

This simple derivation of the laws of probability from the coherence of the prevision underlines the philosophically critical distinction between the consistency and the reasonableness of probability assignments. Probability theory itself can do nothing to guarantee the realism or practical reasonableness of a decisionmaker's system of beliefs. Probability theory can only guarantee the consistency of a set of beliefs. A decisionmaker with coherent beliefs may bet at wildly inappropriate (or, more precisely, unpopular) odds on events, and thus lose money on average at a high rate. But she is not vulnerable to Dutch book, that is, a certain loss on a single group of bets on a single event.

Suppose, for example, that decisionmakers are betting on the fall of tosses of a coin, and let us assume that the "coin-tossing machine," whatever it may be, has produced 500,157 heads in a previous 1 million trials. Consider two decisionmakers. The first is willing to pay \$.90 for heads (or sell heads for \$.90) and to pay \$.10 for tails (or sell tails for \$.10). If we take the 1 million trials as a representative sample of the outcomes, and project them into future outcomes we would predict that this decisionmaker will very likely lose money on average at a high rate, but she is not vulnerable to Dutch book on any single toss of the coin. The second is willing to buy and sell heads for \$.51 and buy and sell tails for \$.51. In a series of bets on one side or the other, we would predict that this decisionmaker will not lose money very rapidly, since she is offering close to what we regard as reasonable odds. On the other hand it is possible for someone to take \$.02 away from her with certainty no matter what the fall of the coin by selling her heads and tails at the same time. The first decisionmaker has coherent but from our point of view unreasonable opinions, while the second has incoherent opinions despite the apparent reasonableness of her judgments. In economic terms we regard the first decisionmaker as ill-informed but not vulnerable to arbitrage, while the second is better-informed, but vulnerable to arbitrage.

The important implication of this philosophical point is that probability theory itself cannot be expected to protect us against bad information and bad judgments, since it is possible to have coherent but unreasonable opinions. To put the point more sharply, there is nothing in probability theory itself that will force the first decisionmaker to revise her opinions towards buying and selling heads for \$.50. Probability theory does require the second decisionmaker to revise her offer set so as to eliminate the arbitrage possibilities. She can satisfy probability theory by changing her offer prices for heads or tails, and it is not

necessary for her to bring these prices into line with the \$.50 that is strongly suggested by the record of the million previous tosses.

If the million previous tosses have an influence on a decisionmaker's offer set, it is because she also has some other beliefs that, together with the requirement of coherence, shape her prices for heads and tails. These other beliefs typically have something to do with the notion that the future is not going to be very different from the present, and that the outcome of another million tosses is quite unlikely to be 900,000 heads and 100,000 tails given that the first million divided equally. But probability theory and coherence do not by themselves compel decisionmakers to beliefs of this kind, reasonable as they are.

## Chapter 6

# Conditional bets and conditional probability

### 6.1 Calling the bet off

Statistical analysis leans heavily on the idea of *conditional probabilities* as a way of expressing the probabilistic dependence of one uncertain quantity on another. The definition of prevision in terms of offer sets and bets is, fortunately, easily extendable to conditions.

We have modeled a bet on an event  $A$  as paying a certain amount of money for an envelope which will contain \$1 if  $A$  occurs and \$0 if  $A$  does not occur. Now consider a bet conditional on a second event  $B$ . You buy an envelope which will have \$1 in it if  $A$  occurs and  $B$  occurs, and \$0 in it if  $B$  occurs and  $A$  does not occur, and you get your money back if  $B$  does not occur no matter what happens with regard to  $A$ . If  $B$  fails to occur the bet is off.

For example, suppose a Presidential election is approaching and you are betting on which candidate will win. Bets on a candidate might be made conditional on the survival of the candidate to election day. You would pay, say \$.55 for an envelope that will have \$1 in it if Jones survives and wins the election, \$0 if Jones survives to election day but loses the election, and you will get your \$.55 back if Jones dies before election day.

The notation for an event  $A$  conditional on another event  $B$  is  $A | B$ . How much should a decisionmaker pay for  $A | B$ ?  $A | B$  is worth more than  $AB$ , since when you buy  $A | B$  you get your money back if  $B$  does not occur, while with  $AB$  you get nothing if  $B$  does not occur. Thus we have:

$$P_I(A | B) = P_I(AB) + P_I(A | B)(1 - P_I(B)), \quad \text{or}$$

$$P_I(A | B)P_I(B) = P_I(AB) = P_I(A \wedge B), \quad \text{or}$$

$$P_I(A | B) = \frac{P_I(A \wedge B)}{P_I(B)}$$

Interchanging the roles of  $A$  and  $B$ , we get *Bayes' Theorem*:

$$\frac{P_I(A | B)}{P_I(A)} = \frac{P_I(B | A)}{P_I(B)}, \quad \text{or}$$

$$P_I(A | B) = P_I(A) \frac{P_I(B | A)}{P_I(B)} \quad (6.1)$$

In economic language Bayes' Theorem tells us that the amount we should pay for  $A$  conditional on  $B$  (our *posterior probability* for  $A$  conditional on  $B$ ) is equal to the amount we would have paid for  $A$  without conditions (our *prior probability* for  $A$ ) multiplied by the ratio of the amount we would pay for  $B$  conditional on  $A$  to the amount we will pay for  $B$  unconditionally (the *likelihood*).

Bayes' Theorem, then, is another law of probability based on coherence. It cannot tell us what our various probabilities (or previsions) ought to be, but only the relations between them that must hold if we are to avoid vulnerability to arbitrage.

## 6.2 Two moments of learning

When we understand the logic of probability correctly as a matter of consistency, it becomes clear that we learn from statistical analyses of data only in a highly restricted sense. Our opinions are already fixed in several dimensions. First of all, we have an observational protocol for the system or phenomenon at issue which specifies which aspects of the system we are going to measure and link together quantitatively. Very often important scientific discoveries turn on a radical change in these observational protocols. For example, Ptolemaic astronomy posed the problem of prediction and observation in terms of the angular position of heavenly objects in relation to the earth. The absolute distances of the objects (and hence their actual size) could play no real role in the Ptolemaic theory. The introduction of absolute distance as an observation concept marked the revolution in thinking associated with Kepler and Newton much more deeply than heliocentricity by itself. (In fact it is possible to reformulate the Ptolemaic system perfectly consistently on a heliocentric basis.)

In addition to the observational protocol, we have a prior joint probability over the conceivable observations of the system, which does not change as a result of observation. What observations do (as we will see in more detail in the discussion of the most commonly used statistical methods below) is to restrict our attention to certain subsets of this joint probability space on which the conditional probabilities may be quite different from the unconditional joint or marginal probabilities of our prior. In strict logic all that probabilistic reasoning can do for us is to police the consistency of our probabilistic beliefs in order to avoid being caught in a Dutch book.

In this logical schema no evidence can overthrow a consistent prior as long as the prior gives a non-zero probability to the evidence. Traditional classical statistical methodology argues for "rejecting" a hypothesis if the probability of

making the observations available on that hypothesis is small. As Bayesians have often eloquently pointed out, this procedure makes no sense. Typically any particular observation has a very low absolute probability. In a sequence of 1000 coin tosses any particular outcome, taken as the actual observed sequence of heads and tails, for example, has on the basis of the hypothesis that the coin is equally likely to come up heads or tails a probability proportional to  $2^{-1000}$ , which is a very small number. There is nothing inconsistent about a decision-maker who, observing 500 heads in a sequence of 1000 tosses, insists on betting that another 1000 tosses will yield around 900 heads. Such a decisionmaker may have reasons for her opinion, (such as the knowledge of several billion tosses in which the proportion of heads is .9, or the suspicion that the reported 1000 tosses were manipulated in some fashion). We will see below that traditional statistical methods can be interpreted quite satisfactorily without the language of “rejection” of hypotheses, and that the scientists who employ these methods can be seen as more sensible than their use of classical statistical language would suggest.

How, then, do we account for the fact that observations do play an important role in changing scientists’ minds about hypotheses? A broader process of imagination and judgment is at work here than can be encompassed by the mechanical formulations of probability theory by itself. To begin with, observations sometimes are logically incompatible with priors, in the sense that the prior gives a zero probability to the observation. This situation, however, is easily remediable through minor amendment of the prior. For example, a scientist may practically amend her prior by adding a low probability of “anomalous” or inconsistent observations. This amendment could be rationalized by the knowledge that observations are subject to experimental error, fraud, confusion, and so forth.

But “saving” theories in this elementary fashion doesn’t move science forward very much. Movement forward depends on a moment of scientific investigation quite different from the statistical analysis of data within a given framework, namely the invention and discovery of radically new frameworks in which to view data. The statistical analysis of data requires scrupulous construction of detailed and sometimes lengthy chains of reasoning connecting observations with a well-defined prior joint probability. The invention and discovery of new frameworks rather employs the human intuitive gifts of analogy and formal transformation through symmetries and generalization.

The scientist who develops a new framework in which to evaluate data may derive the new framework in part or in whole from an existing framework developed to analyze a different phenomenon. In this case the proposal of analogies plays a central role. The electromagnetic forces shaping the atom, for example, might be recognized as mathematically analogous to the gravitational forces governing planetary motion. The tendency for economic competition to eliminate low-profit firms can be seen as analogous to the tendency for reproductive competition to eliminate low-fitness mutants in biology (or, as apparently was the historical case, vice versa). Existing frameworks may also be subjected to formal manipulations, such as the interchange of symmetries, or extension by

generalization. In economics the commodity space might be extended to include the contingent commodities that we used as the basis of the definition of probability and prevision above. The introduction of entropy concepts into economic market analysis may transform the prices from uniform budget lines faced by all agents into a family of iso-probability loci in a Gibbsian distribution.

The leap from one framework to another through formal manipulation of existing frameworks or the discovery of analogies is at an entirely different level from the incorporation of new observations into existing frameworks through statistical analysis. Frequently the new framework suggests a different protocol of observation, requiring and permitting the measurement of a larger number of parameters of the system. The replacement of one framework by another is rarely the result of the existing framework's inability to explain data collected according to its own protocols, and more often the result of the inability of the existing framework even to address certain dimensions of the data collected according to the protocols of new frameworks. Ptolemaic astronomy cannot conceptually come to grips with the problem of the actual distribution of heavenly objects in 3-dimensional space, which is the natural form of astronomical observations in the Kepler-Newton framework. Walrasian economics has no category corresponding to the entropy or disorder of the economic allocation, and cannot as result address observations of this dimension of market outcomes.

Thomas Kuhn, Imre Lakatos [Kuhn, 1962, Lakatos, 1978] and other philosophers and historians of science have grappled with the complex historical dialectic through which anomalous observations in one theoretical framework pass from being ignored, or compounded in observational error through becoming the central problems of explanation in a science, to textbook status as the decisive explanatory trophies of new paradigms. In this dialectic statistical analysis plays an important but ancillary role. Statistical data analysis within a given framework confirms the relevance of observations to the issues posed, without settling those issues. Observations which have little or no effect on conditional probability distributions within a framework (say, scattered or anecdotal evidence) cannot be candidates to play the role of theory-shaping anomalies. Once their candidacy is established statistically, however, the problem of theory reformulation enters the realm of intuition, analogy and formal manipulation.

# Chapter 7

## What can we bet on?

### 7.1 Operationalism

The simplest betting scenario goes like this. It is the morning of a day when there is a horse race scheduled for the afternoon. Native Dancer is entered in the race. We are uncertain as to whether Native Dancer will win the race or not, and we bet with each other on this event. The race takes place, and we all find out whether or not Native Dancer won, and pay off the bets accordingly.

This basic pattern has the form: uncertainty  $\rightarrow$  betting  $\rightarrow$  resolution of uncertainty  $\rightarrow$  payoffs.

In the scientific sphere the interpretation of probabilistic consistency in terms of bets runs into some puzzling issues. Very often scientists discuss hypotheses that by their nature can never be resolved by direct human observation, for example, the origin of the Universe, or its end. Similar issues arise when we discuss historical events about which only limited evidence survives, such as whether the destruction of a particular ancient city was the result of invasion, insurrection, or earthquake. We can see the archeological evidence, the charred timbers and tumbled walls, but we can never travel back in time to observe what actually happened. Scientific investigation is often phrased in terms of inherently unobservable parameters of models, like the mass of Saturn, or the charge of the electron. In some cases these parameters are well-defined conceptually only in terms of impossible observations. The “true” charge of the electron might be thought of as the average of its experimental value in an infinite sequence of correctly conducted experiments. But human life is finite, so that this infinite sequence, though mathematically well-defined, is operationally impossible.

In general, as we will see in more detail below, statistical analysis begins with a system of prior probabilistically consistent beliefs, which constitute a framework in which the evidence of new observations can be incorporated. The Laplacian requirements of probabilistic consistency, together with a system of plausible prior beliefs, often sharply restrict the beliefs we can have about aspects of the system we have not as yet observed.

In the case of the horse race, our prior beliefs and information about Native Dancer's performance in other similar races shape our probabilities for the event that he will win this particular race. Once the race is run, we can observe the outcome. On the other hand, the actual observation of the outcome of the race doesn't seem to have any bearing on the making of bets, only on their settlement. An earthquake that occurs at noon, after the bets are made and before the race is run, which prevents the race from being run, has no impact on the reasoning that led to the bets themselves. The making of bets is an expression of probabilistic beliefs whether or not the actual uncertainties being bet on are actually resolved.

On the other hand, there is clearly no real point in betting on events which are completely unobservable in principle, such as whether or not Native Dancer would have won a race which never actually occurred. Suppose, for example, that there is no earthquake and the race is actually run, but that at noon the bettors themselves are captured by an interstellar spaceship and removed to another star system where they find it difficult to receive news from earth. They know the race was run but there is very little hope that they will find out something bearing on its outcome. In this situation bets would not be strictly pointless, though they might be frustrating in the sense that many of them could never be settled.

## 7.2 Unresolvable bets

Probability theory can be applied formally in many situations to yield similar restrictions on our beliefs about events that we inherently can never observe directly. In these cases a bet would be beside the point, because it could never be settled. Nonetheless, scientific discourse proceeds as if these bets could be made.

In this respect science seems to be more like the bettors who are removed by the spaceship. Scientists have difficulty in obtaining news about many of the events they bet on, but there is always some chance that news will arrive, and allow either a complete or a partial resolution of the uncertainty. When Laplace calculates the odds on the mass of Saturn, for example, he is not proposing that some direct measurement would allow us to settle the mass of Saturn once and for all. He is supposing that the observations he works with are linked systematically to a large set of potential other observations through the framework of Newtonian mechanics. The "mass of Saturn" can be calculated from these other observations, and Laplace's bet on its value is a concise way of betting on their outcome. Laplace himself may have thought of these observations as being limited to continued terrestrial measurements of the positions of Saturn, its satellites, and other planets. Later, technologies may emerge that allow a different class of measurements from interplanetary spacecraft. Laplace's bet, phrased in terms of the mass of Saturn, can still be parsed in terms of the measurements of the interplanetary spacecraft's orbit.

Thus scientific bets on inherently unobservable events can be unpacked into

a series of bets on observables. As we shall see below, in the case of the most commonly used statistical methods, this unpacking is more immediate than traditional expositions reveal. Even when common statistical practice introduces what appear to be unobservable parameters, they turn out to be functions of potentially observable quantities.

Even though it is possible to use statistical reasoning to make consistent probabilistic statements about inherently unobservable quantities, such as the average of an infinite number of measurements of the charge of the electron, the settling of these bets must turn out in actual scientific fact to be unnecessary.

To sum up. The mathematical structure of probability allows us to formulate probabilistically consistent beliefs not only about potentially observable but even about potentially unobservable quantities. The only scientifically relevant beliefs, however, concern potentially observable evidence. The archeologist can never travel back in time to see whether an earthquake or an invading army started the fire that destroyed the city. She might, however, be able to find traces of the siege works of an invader outside the city walls.

The difficulty here is that we have a strong psychological predisposition to couch our knowledge in terms of imaginary visualizations. One such scenario for the archeologist begins with a peaceful day in a functioning city, and moves on to the shaking of the ground, the panic of the population, the fires breaking out in collapsing buildings. Another scenario involves the besieged city, the closed gates and defended walls, the moment of inattention or fatigue or betrayal that opens a breach, the invading army streaming in, the panic, fires, and so on. Probabilistic consistency often fails to satisfy this predisposition because it overlays *all possible hypotheses, even when they are inconsistent with each other*. A bet on whether or not there are siege works outside the city is not the same as a bet on whether the city was burned by invaders, since there are a variety of other hypotheses that also could lead to the existence of the siege works. For example, the invading army might have struck its tents and sailed away the day before the earthquake destroyed the city. When we bet on an observable we bet on the whole class of imaginary hypotheses that are consistent with that observation, leaving us inevitably uncomfortable with our failure to resolve residual uncertainty.

What then becomes of the “truth?” I think we had best dispense with this psychologically powerful but scientifically slippery concept altogether. Evidence has the indispensably important role of restricting our beliefs over the range of possible hypotheses. As a result it often restricts our beliefs over relevant operational measurements and over the consequences of action. Narrowing the range of uncertainty is of tremendous value to us. It leads naturally to an image of a complete resolution of uncertainty, but this image is a psychological projection of the real process of investigation to an ideal extreme that can never be realized. We escape the space invaders and sit in the stands and watch Native Dancer cross the finish line first, receive the blanket of flowers, and hear the victory announced on the public address system. As a result we pay off the bet. But how do we know for sure that Native Dancer’s trainer did not drug the other horses, or bribe the other jockeys? We are content practically

speaking to accept a certain amount of evidence as tantamount to resolution of the uncertainty, even though there strictly remains some margin of unresolved uncertainty in the situation.

Traditional expositions of statistical theory lean heavily on the psychologically attractive, but methodologically questionable, concept of the “true model.” “Suppose,” they say, “that we have  $n$  data points that are generated from a normal probability distribution with an unknown mean and variance.” We have the image of a black box with a dial marked “mean” and another marked “variance” set to particular values. On the basis of any finite set of data generated by this black box it is impossible to rule out logically any setting of the dials. Does it make sense to bet on the settings of these dials? Not really. The most we could bet on would be the sample mean and variance of finite sets of further observations generated by the same process, even though Laplacian probability does allow us to make formally consistent probability statements about the setting of the dials.

One might object that in principle bets on the setting of the dials could be settled by direct observation of the dials. But how do we know that the black box “truly” generates normally distributed random numbers? This hypothesis is also unsettlable on any finite set of observations.

The image of the “true model” is at the root of much of the confusion and imprecision that dogs conventional statistical analysis. It leads directly to the incoherent interpretation of certain statistics in terms of “hypothesis testing,” the rhetorically satisfying but scientifically disastrous “rejection of hypotheses” and so forth.

We will see below that there is, for the most widely used statistical models, an extremely satisfactory resolution of these problems which interprets the statistics actually calculated in common statistical practice in a way that respects Laplacian theory and avoids the concept of the “true” underlying model.

## Chapter 8

# The frequency model of probability

Human beings have evidently played games of chance for a very long time. It is as common to find chance devices, like throwing bones, the precursors of dice, in archeological sites as it is to find cooking pots.

The usefulness of a chance device, like a roulette wheel, a tossed coin, or a pair of dice, lies in the fact that it is very difficult for an observer to predict the outcome when the device is set in motion. We know now that most chance devices work because they are chaotic systems that display sensitive dependence of their outcomes on initial conditions. In fact the fall of a pair of dice or the rest point of a roulette wheel are completely determinate given the initial momenta of the dice or angular velocity of the wheel. It is possible using modern technology to measure the initial momenta of dice with laser sensors and for very fast computers to calculate the trajectory and the final resting state of the dice before they actually fall. With a great deal of practice it is possible for human beings to learn to control the outcome of these random devices through their skill in controlling the initial conditions. (And, of course, much easier for the thrower simply to substitute weighted dice to influence the frequency of the outcomes.) But in ordinary situations the players of the game assume that it is very difficult for the thrower to control the outcome of the dice, and equally difficult for an observer equipped only with unaided vision to predict the outcome. In any case bets are usually made before the throw of the dice, and the integrity of the game depends on the inability of the the thrower to influence the outcome.

Chaotic systems of this kind exhibit statistical regularity in samples of observations of outcomes. If we throw two regular dice repeatedly, “about”  $1/36$ th of the outcomes will have 2 or 12 spots showing, “about”  $1/6$ th of the outcomes will have 7 spots showing, and so on. This observation suggests the idea that the frequency with which an outcome appears in a large number of observations is a measure of its “probability” and is a characteristic of the “random device”

or “random system”. As we will see in more detail later, these ideas do have a certain validity to them. But the coherent path to these conclusions is much more subtle than it might first appear.

## 8.1 Probabilities as frequencies

If we define the probability of an outcome to be its observed frequency in a large sample of observations from some given random device (all of which terms are problematic, as I shall discuss below), we get a set of numbers  $p_i^n = x_i/n$ , where  $i$  runs from 1 to  $r$ , the number of possible outcomes. In this construction  $x_i$  is the number of times the  $i$ th outcome has been observed in the sample of  $n$  observations. These numbers are evidently non-negative, lie between zero and 1, and add up to 1. They therefore satisfy the abstract axioms for a system of probabilities, and the probabilities derived from them for events that are unions and intersections of sets of primitive outcomes will satisfy the laws of probability.

From many points of view this seems to be the simplest model to motivate the laws of probability. These probabilities seem to be a characteristic of the random device that generates the data, and to have nothing to do with the observer who measures the outcome, and are therefore “objective” facts about the world. Yet a closer look reveals that this model raises, even in this very simple form, a number of knotty problems.

### 8.1.1 The problem of finite sample size

First, we know that if we were to run the device to generate another sample of  $n$  observations, we would be unlikely to come up with exactly the same  $p_i^n$ . In fact, if we did always come up with exactly the same  $p_i^n$ , we would be suspicious of the underlying “randomness” of the generating device. One of the fundamental insights of probabilistic thinking is that in repeated finite samples there will be some variation in the observed frequencies of the possible outcomes. This observation, however, reveals the definition of “probability” as the observed frequency of an outcome in a large sample of observations to be incoherent and ill-formed.

In order to remedy this defect, frequentist probability theory is tempted to define the “probability” of an outcome as its frequency in an *infinite* sample of observations. Since there is no such thing as an infinite sample of observations, the frequency theory substitutes a mathematical concept, the limit of the sequence of sample frequencies as  $n \rightarrow \infty$ , so that the probability of outcome  $i$  is actually  $p_i \equiv \lim_{n \rightarrow \infty} p_i^n$ . By the mathematical laws of continuity of the real numbers, these limiting probabilities will also be non-negative and sum to 1, and thus inherit the properties of an abstract system of probabilities from the sample frequencies.

But this mathematical maneuver raises more questions than it answers. How do we know that a limit exists? In fact, is the mathematical notion of a limit

appropriate to describe the situation at all, since at the best we would never actually have an infinite sample, only a finite (though perhaps very large) number of observations. The mathematical notion of a limit is defined for abstract mathematical objects, sequences of real numbers, not for finite sequences of observed frequencies.

### 8.1.2 Probabilities as tendencies

In order to respond to these questions, the frequency school posits the existence of probabilities as parameters,  $\{\theta_i\}$ , which express the *tendency* of the random device to produce the outcomes. These parameters, of course, are not *directly* observable. The best that we can do is to infer them from the observation of finite samples. A considerable amount of mathematical effort has been expended in specifying the mathematical hypotheses which will ensure that sequences of observed frequencies of outcomes of a random device (now elevated to the mathematical category of a *random variable*) will converge to the parametric probabilistic tendencies  $\{\theta_i\}$ , that is, that  $p_i = \theta_i$ . These results are usually called the *Law of Large Numbers*.

Unfortunately the various Laws of Large Numbers do not really address the fundamental difficulty with the frequency approach, which is that *any* random device with positive  $\{\theta_i\}$  can produce *any* finite sample. A coin with an “inherent” tendency to come up heads 50% of the time *could* produce a run of 100 or 1000 or 10 million heads. In fact, the coin would not be “purely” random if it did *not* produce such runs occasionally in very large samples.

Thus the attempt to conceptualize probabilities as frequencies leads to a logically *backward* procedure of statistical inference. Instead of considering the problem of what a given sample tells us about the likely behavior of the random device (most relevantly, what frequency of outcomes we might expect other samples from the device to exhibit), the frequency approach directs our attention to the question of what abstract mathematical devices might have produced the sample evidence. Since this question has no real answer, the frequency approach is forced to adopt a number of ad hoc procedures such as “confidence intervals” and “levels of statistical significance”. These methods sometimes give sensible answers, but often are disastrously misleading.

## 8.2 Are frequencies “objective”?

The appeal of the frequency interpretation of probabilities is that it locates the uncertainty about outcomes in the external world rather than in the interaction of a human observer and the external world. In psychological language we might say that the frequency interpretation *projects* the anxiety of a responsible and uncertain human decision maker onto the external world. Thus the frequency interpretation appears to relieve the human observer of a considerable burden of responsibility.

But a closer look at the classic examples of random devices invoked to support the frequency interpretation, such as tossed coins, thrown dice, roulette wheels, or urns containing balls of different colors, immediately resurrects the problem of the observer's state of knowledge.

Take the very simple example of the tossed coin. The real motivation to suppose that the coin has an inherent tendency to come up heads 50% of the time is the *absence* of any information that would lead the observer to expect a preponderance of one outcome over the other given the symmetry built in to the coin itself. But there are evidently many potential reasons for breaking this symmetry. If the observer knows that the coin is the property of a magician who has practiced many hours to control the fall of the coin, or of a charlatan who has weighted the coin to make it easier to toss it to fall one way or another, she will have a different opinion as to the "inherent" tendency of the coin to come up heads. If the observer is in a position to make a detailed physical investigation of the roulette wheel's dynamics, she will modify her opinion of its "tendency" to produce a uniform distribution of outcomes accordingly. If the observer can watch the experimenter put the colored balls into the urn, and knows how the urn was manipulated during this preparation, whether it was shaken vigorously or not, for example, she will adapt her ideas as to the likely composition of samples drawn from the urn.

The frequency interpretation tends to confuse the issues of the "objective" character of the observed sample observations with the "objective" character of the underlying parametric tendencies it posits. Within the limits of observational error and human fallibility sample outcomes may indeed be "objective" in the sense that almost all observers will agree on their count of the frequency of the various outcomes in any observed sample. (Even here it is evident that the word "objective" really stands for the existence of an "intersubjective consensus", that is, the ability of a large group of observers to agree as to what has actually happened.) But there is a huge leap from the counting of outcomes in finite samples to knowledge of the underlying parameters expressing the tendency of the random device to produce various outcomes, as we have seen.

Thus the frequency interpretation leads back to the fundamental issue of the *information state* of the observer. It is not always evident how the information state of the observer is properly to be taken into account in understanding uncertain situations, but it is clear that the information state of the observer is an indispensable condition of the problem of probability.

### 8.3 Probabilities that cannot be frequencies

A further difficulty with the frequency interpretation of probability is that there are many situations in which we naturally want to talk about the probability of an outcome or event in which the very concept of a large (not to say infinite) sample of observations simply does not make sense.

To take another example from human gambling behavior, each horse race is

fundamentally a unique event. We do not repeatedly run the “same” race over and over again, in the sense of matching exactly the same horses under exactly the same conditions. Even if someone were to try to contrive such a series of “identical” races, they could not be exact repetitions: the horses would age through the process, the riders would learn from their experiences, and so on. Nonetheless, it is an every-day reality that people express their probabilities of the outcomes of horse races in the odds at which they are willing to bet.

This type of situation is very common in finance and business. Many uncertain financial and business contingencies are essentially unique. They arise once, in very particular circumstances and are then resolved. Whether a newly discovered technology will prove economically profitable, or a newly promoted executive will “grow” in the job or not, or a merger will provide the profitable “synergy” that motivated it, all tend to be questions about one-off situations which are never exactly replicated. Nonetheless, the existence of markets for financial assets and other products forces people to make probability judgements about these uncertainties.

The frequency model of probability faces an awkward set of problems in these situations. Either probabilities in the frequency sense simply don’t exist, which gravely restricts the applicability of probability methods, or we have to posit some kind of imaginary world in which repetitions of these inherently unique situations can be contemplated. This is an unpleasant dilemma. When one looks very closely at the notion of “repeated trials” of an experiment or random device, there are practically *no* situations that really fit the model. Scientists make great efforts in the laboratory or in the design of experiments to “control” for disturbing variations in the conditions under which a trial observation is generated, but they cannot completely succeed, even in the most protected realms of the observation of controlled physical systems. Heraclitus puts this horn of the dilemma in the dialectical dictum that “we never step in the same river twice”. But then probabilities in the frequency sense *never* exist, and probability theory has no field of application at all.

But it is equally unpalatable to posit the existence of imaginary repeated trials of what appear to be one-off events. At the extreme, this line of thinking projects the uncertainty of the human observer into the existence of “multiple worlds” in which all the possibilities of every situation are realized. Whatever the appeal of this conception for the writer of time-travel science fiction, its relevance to commonplace human uncertainty is doubtful.

## 8.4 Transcending the frequency model

While the frequencies of observed outcomes in samples are non-negative numbers that sum to unity, and thus provide an intuitive example of a system of probabilities in the abstract sense, the attempt to build a logical theory of probability on this interpretation runs into insuperable logical and philosophical problems. The most fundamental of these problems is that sample frequencies can by their nature never characterize probabilities of a random device uniquely. The

attempt to circumvent this problem by positing unique parametric probabilities measuring the tendency of the random device to produce particular outcomes cannot overcome the difficulty that any finite observed sample is compatible with any probabilistic tendencies of the random device.

There is a kernel of truth in the frequency notion, which de Finetti's axiom of *exchangeability* of samples makes transparent. In order to appreciate this kernel of truth, however, we must embed the problem of statistical inference in a more general interpretation of probability, one that explicitly takes the information state of the observer into account, and indeed puts the information state of the observer at the center of the problem. When we consider the real problem of what a finite sample of observations might tell us about the likely composition of other samples in this framework, the solution is logically transparent and avoids the paradoxes and limitations of the frequency interpretation.

# References

- George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, New York, 1973.
- William S. Cleveland. *Visualizing data*. Hobart Press, Summit NJ, 1993.
- Bruno de Finetti. *Theory of Probability*. Wiley, New York, 1974.
- Franklin A. Graybill. *Matrices with Applications in Statistics*. Wadsworth, Belmont CA, 1983.
- Ian Hacking. *Emergence of probability : a philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press, Cambridge UK, 1975.
- Jonathan Hope. *The authorship of Shakespeare's plays: A socio-linguistic study*. Cambridge University Press, Cambridge UK, 1994.
- Colin Howson and Peter Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, Chicago, 1993.
- E. T. Jaynes and G. Larry Bretthorst (ed.). *Probability Theory: The Logic of Science*. Oxford University Press, Oxford, 2003.
- Harold Jeffreys. *Theory of Probability*. Clarendon, Oxford, 1939.
- Ali Khan and Yeneng Sun. Nonatomic games on Loeb spaces. *Proceedings of the National Academy of Sciences*, 93:15518–15521, 1996.
- Ali Khan and Yeneng Sun. Non-cooperative games on hyperfinite Loeb spaces. *Journal of Mathematical Economics*, 31:455–492, 1999.
- Thomas S. Kuhn. *Structure of scientific revolutions*. University of Chicago Press, Chicago, 1962.
- Frank Lad. *Operational subjective statistical methods : a mathematical, philosophical and historical introduction*. Wiley, New York, 1996.
- Imre Lakatos. *Philosophical papers*. Cambridge University Press, Cambridge UK, 1978.

- Pierre Simon Laplace. *Philosophical essay on probabilities*. Springer-Verlag, New York, 1995. [1825].
- Robin Pope. Evidence of deliberate violations of dominance due to secondary satisfactions—attractions to chance. *Homo Economicus*, XIV(2):47–76, 2001.
- Theodore M. Porter. *Rise of statistical thinking, 1820-1900*. Princeton, Princeton NJ, 1986.
- Frank Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, pages 156–203. Humanities Press, New York, 1950.
- R. D. Rosenkrantz, editor. *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. Kluwer Academic, Dordrecht, 1989.
- Leonard J. Savage. *Foundations of statistics*. Wiley, New York, 1954.
- D. S. Sivia. *Data analysis : a Bayesian tutorial*. Clarendon Press, Oxford, 1996.
- Stephen M. Stigler. *History of statistics : the measurement of uncertainty before 1900*. Harvard University Press, Cambridge MA, 1986.