The Probability of Mr Bayes

A constructive re-evaluation of Mr Bayes' essay and of the opinions concerning it expressed by various authorities.

Frank Barker
and
Robin Evans

The Department of
Electrical and Electronic Engineering
University of Melbourne
Parkville 3052
Australia

13 July 2001

Version zh25a.doc at 18:36 on 13 July, 2001

CHANGES

Version, *h22*:

Refines the differentiation between errors and individuals -
see esp the concluding paragraphs of Chapter 10 .
Details of changes can be inspected at
http://www.ee.mu.oz.au/research/reports/Bayes/update_record.pdf

APOLOGIES

In versions prior to *h19*, there were errors
in *Eqn 1* in the Précis. In versions prior to *h25*,
there were typographic errors in Ch 13, *"The Base Rate"* and
the derivation of the Gaussian Ratio given in Appendix E was wrong.

RECORD OF ISSUES

zh8 22 Sept 1999
zh10 20 Oct 1999
zh11 30 Oct 1999
zh12 30 Nov 1999
zh13 6 Dec 1999
zh16 19 Dec 1999
zh18 5 Jan 2000
zh19 29 Jan 2000
zh20 6 Feb 2000
zh21 17 Feb 2000
zh22 24 Mar 2000
zh23 3 August 2000
zh24 6 October 2000
zh25 29 October 2000
zh25a 13 July 2001

For
Carolyn and Margaret

"A wise man proportions his belief to the evidence".
David Hume,
'An Enquiry Concerning Human Understanding'
SECT.X

Version zh25a.doc at 18:36 on 13 July, 2001

Contents

Précis

1. Motivation

1.1 The concept of Probability is important, in science, in engineering, medicine, navigation and in making everyday decisions, where, despite large uncertainties, one has a moral obligation to act on a rational basis. In many situations, people need plain estimates of probability in order that they may make decisions by clear reasoning concerning probable benefits and penalties. Attenuated substitutes - likelihoods and confidence limits - provide little or no direct support for rational decisions. People, especially professional people, need a clear theory for their use of probability.

1.2 Bayes' *Essay towards solving a Problem in the Doctrine of Chances*[1] was revolutionary and foundational in addressing the root problem of rational justification for action under uncertainty. The problem is fundamental to the interpretation of uncertain quantitative evidence: it is embedded - but often unseen - in every measurement. Even in a simple counting of objects we can make mistakes.

1.3 Yet, Bayes' problem has been widely ignored, largely because eminent authorities - Boole, Keynes, Fisher, Fine - appeared to have shown that there could be no general solution and that Bayes' solution, based on the postulate of a uniform distribution of prior probabilities, involves self-contradictions.

2. The main thrust

2.1 Against this background, we embarked upon a constructive re-appraisal of Bayes' essay. By re-analysing the essay in relation to the metric procedures actually used by practical people - builders, doctors, engineers, navigators - we found that Bayes' solution is essentially sound and that the main objections raised against it are fallacious. It thus becomes possible to make clear and objectively valid statements of probabilities concerning metric hypotheses, against defined evidence and assumptions. The difficulties which have long undermined confidence in Bayes' theory, stem from a misleading view of prior probabilities, accidentally suggested by Bayes himself and then taken for granted by almost all subsequent authors, including his most fierce critics. Serious problems are also found in the widespread failure - marked among lawyers and administrators - first, to distinguish between individuals and populations and, secondly, to under-

[1] Bayes (1763), p 376

Version zh25a.doc at 18:36 on 13 July, 2001

stand that probabilities in the real world can be assessed only in relation to the evidence and assumptions on which they are based, and by means of which infinite regressions of uncertainty are avoided.

3. Analysis

3.1 We define a 'Bayes trial' as one where we wish to find the probability of occurrence P_m of an event M in a given type of test. We have no prior information about the value of P_m. We therefore perform a set of n tests and count the number of tests, m, in which the event occurs. From the values m and n we are required to determine the probability that P_m lies between values x_1 and x_2.

3.2 We follow Bayes in measuring the probability of an event as 'the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening'. We find that this requires us to compute probability relative to defined evidence and assumptions.

3.3 Bayes showed that, with two events E_1 and E_2, if it is discovered that E_2 has happened, the probability that E_1 has also happened is given by:-

$$\mathcal{P}(E_1 | E_2) = \frac{\mathcal{P}(E_1) \times \mathcal{P}(E_2 | E_1)}{\mathcal{P}(E_2)} \qquad \textit{Eqn 1}$$

This is often known as "Bayes' theorem."

3.4 However, Bayes and others did not see the need to denote the extrinsic data and assumptions, k, on which the estimates of probability are based. Nor did they see the need to differentiate, in considering a Bayes trial, between $\mathcal{P}_H$ - the probability of an hypothesis concerning a value that is fixed throughout a trial, and $\mathcal{P}_R$ - the probability of an event which varies randomly within such a trial. Doing so gives the more precise form of the theorem:-

$$\mathcal{P}_H(E_1 | k, E_2) = \frac{\mathcal{P}_R(E_1 | k) \times \mathcal{P}_R(E_2 | k, E_1)}{\mathcal{P}_R(E_2 | k)} \qquad \textit{Eqn 2}$$

3.5 Applying the theorem to the determination of P_m, Bayes envisaged an experiment in which a ball is thrown randomly onto a uniformly flat and level table to decide the value of P_m. He postulated that, when we have no prior information about P_m, we should likewise assume a uniform prior probability. This allows us to compute the probability of the hypothesis that P_m lies between values x_1 and x_2 - given all the explicit and implicit assumptions.

3.6 Bayes' postulate of the uniform prior has been attacked on two main grounds. First, that it implies knowledge, which, by definition, we do not possess. Second, when we are concerned with the value of a dimensional parameter, such as a density, a uniform prior in terms of mass per unit of volume becomes non-uniform, and therefore self-contradictory, in terms of volume per unit of mass.

3.7 However, when builders measure a dimensional attribute of an object, *e.g.* the length of a beam, they do not need to invoke any prior probability over the magnitude if they are using a ruler where the random errors are additive, can be assumed to be independent of the magnitude being measured, and have a distribution known by calibration.

3.8 The issue of prior probability in such cases is thus switched from the true value of the attribute to the calibrated measuring process. It is then irrelevant whether we are measuring amps, ohms, volts or watts. The prior probabilities now relate to the calibration of the meter.

3.9 Hence, by arithmetic subtraction we obtain a distribution of probability over the 'true' value relative to the standard by which the meter has been calibrated - provided the meter is working as calibrated.

3.10 When Bayes' theorem is applied to this type of measuring operation, we find that the events to which the probabilities relate are:-

E_1 is the event that the error has a given magnitude. The probability of this is given by the calibration.

E_2 is the event that the measuring process is working according to the calibration.

3.11 Where necessary, we conduct a Bayes' trial to compile a histogram of error frequencies and thence obtain calibrated error-probabilities. The median - beloved by Fisher - appears to be self-calibrating. In trajectory estimation, *e.g.* by a Kalman filter, an assumption of normally distributed errors also simplifies the calibration.

3.12 In a Bayes trial, however, the uncertainties have binomial distributions which depend on the underlying true probabilities. In this situation, the calibrated ruler is not available and to compute the distribution of the uncertainties we have to use Bayes' theorem in a way which forces us now to assume priors - in Bayes' original sense - for the true values of the error probabilities.

3.14 However, for a Bayes trial to converge on the true value of P_m, the only acceptable priors are:- (i) previous observations of the specific individual case, or (ii) priors which exert no influence on the posterior probabilities, such that the posterior distribution is determined entirely by the observations.

3.15 If a distribution of an attribute within a population is used as a prior for an attribute in an individual case, it will, in general prevent the posterior distribution from converging on the true value. An exception occurs when the population is uniformly distributed - as is the case in Bayes' experiment.

3.16 In general, however, where there have been no previous observations on an individual case, the only valid prior is that which represents zero prior information. In a Bayes trial, the 'information zero' prior is the uniform distribution.

4. Conclusion

4.1 The problems traditionally associated with the assumption of a prior distribution in a metric process are not fundamental and are avoidable. In a Bayes trial, the assumption of a uniform prior is objective, fundamental and essential if the result is to converge absolutely on the true value.

4.2 The issues raised by Bayes' essay are fundamental to the acquisition of quantitative evidence in that:-

(1) A Bayes' trial is the most basic metric procedure, involving simply the counting of events with no dimensional parameters.

(2) Every metric process involves uncertainty. To handle that uncertainty in a dimensional situation we have to :-

(a) Calibrate the meter by Bayes trials.

(b) Use Bayes' theorem to combine calibration and measurements.

4.3 Finally, we discuss some implications of these findings for the analysis of causation and for a number of legal and ethical situations. We deal specifically with the so-called 'base rate problem' where accepted wisdom is in a state of serious confusion, stemming from a failure to distinguish between the $\mathcal{P}_H$ and $\mathcal{P}_R$ senses of probability. We point out some consequent problems for rational, moral behaviour and the need for a broad harmonisation of human values which are implied by the use of Bayes' theory as an aid to rational decisions.

Synopsis

Preface
Initial motivation - locations - report of an exploration, not an homogeneous view - central aim - a reliable view of truth - degree of confidence with which we can rationally believe assertions derived from observations.

Acknowledgements
Organisations and friends

Notation
Aim is to make arguments accessible to widest possible readership without sacrifice of rigour where it matters.

Definitions
Bayes' question - Bayes' theorem - Bayes' solution - delta function - Objective, Subjective - Relative, Absolute.

Language
Apology for long sentences - 'which' and 'that' - treatment of the noun 'data'.

Chapter 1 Introduction
Brief historical background - fundamental nature of Bayes' problem - style of Bayes' essay - reasons for its importance - practical applications - misrepresentations - rationale of our exposition - overview.

Chapter 2 The Essay
Covering letter by Richard Price - Bayes' problem - Bayes' definitions - Prop.1: additivity of independent probabilities - Prop.2: expectation a function of probabilities of loss and gain - Prop.3: probability of subsequent events - Prop.4: concerning a complex trial involving two subsequent events - Prop.5: the probability of an hypothesis concerning a correlated event - Prop.6: the multiplication rule for independent events - Prop.7: the binomial formula - Bayes' experiment - Prop.8: the geometry of prior probability in the experiment - Prop.9: the geometry of inverse inference in the experiment - the Scholium concerning the 'proper rule' to be used in cases concerning an 'unknown event'.

Chapter 3 The Elements of Bayes' Probability
Introduction to the analysis of Bayes' argument - detailed analysis of the definitions, propositions as far as Proposition 7.

Derivation of an expression showing the distribution over a parameter which is defined as the ratio of 'Gaussian expressions'.

Preface

This book is the history of an exploration of Thomas Bayes' *Essay towards solving a Problem in the Doctrine of Chances.* The exploration goes back over thirty years; the book itself was conceived some ten years ago in the blissful summer shade of an English cricket ground, the first intent being merely a short note to explain an un-resolved flaw in the theory underlying the application of Bayes' theorem to radar systems. The next day, a visit to the kindly librarian of the Royal Society provided us with a copy of the original. Reading it, we found ever more reasons to doubt all that we had previously gathered about what it actually said and meant. Now, after many a hundred further hours of glazed eyes pointing outwardly at cricket, while inwardly retracing, time after time, the arguments of Boole, Keynes, Fisher, Fine, and many others, we offer the story of our journey and what we found. Often too, we have fought out our doubts while stamping the sand of Melbourne's Port Phillip Bay where, with no social inhibitions to constrain us and the waves to drown our shouting, we have given full voice to the frustration and anger that spring so easily from finding confusion where one had believed there to be good order. We offer, therefore, not an homogeneous view as seen from a concluding position, but, as exhorted by Sir Peter Medawar and Sir Herman Bondi, a report which follows the sequence of our exploration and the unfolding of our perceptions. By this means, we may perhaps convey something of the challenge and excitement which has kept us at this quest for so many years, and in which we invite others now to join.

Truth, however, is slippery stuff and the central aim of this book is to give us a better grip on certain of its aspects. Not on *the* truth, for another axis on which our argument turns, is the perception that, when we endeavour to convey truth in words, we engage in a process of forming and projecting an image. Perforce, therefore, we use lenses, filters and mirrors, each of which in some aspects sharpens and, in other aspects, blurs the image we project. Truth, absolute truth, there may well be in a transcendental space of infinite dimensions. Our ability to project that truth using our own earthly, murky minds as lenses is limited. The resulting image is only to be understood as a convolution of the absolute with the confusion and prejudice of those minds. And often, lenses, mirrors, prisms will invert an image: 'up' is shown as 'down'; 'right' as 'wrong'. Thomas Bayes and countless others engaged themselves in this quest of seeking and projecting the truth, and whatever harsh words we may later have to say about the images they produced, let us remember that they were all engaged in one great quest.

Yet there is a sense in which we seek to project an image which is, if not absolutely true, then is at least solid and trustworthy. For this book is a philosophical horse from an engineering stable; and, from engineers and navigators, as from their doctors, people expect reliability. That the expected performance is not always achieved, is caused, all too often, by a basic failure to understand the nature of probability. Had the *Titanic* been employed on the sea route between India and Britain, she would never have struck an iceberg and could well have become an icon of, apparently, safe design. In some modern fields of engineering, medicine and navigation, however, the concepts of probability and uncertainty are fundamental. This is especially marked in the fields of remote sensing, such as radar and medical scanning, and in communications, where we are concerned not merely with the relatively simple concept of direct probability - as when we draw a card at random from a pack of known constitution - but also with the more difficult and contentious problem of inferring the constitution of the pack from the cards we have drawn. In medicine, there are further, analogous problems: the direct problem of estimating the probability that exposure to infection will give rise to disease and, conversely, the problem of inferring the probability of a specific disease from the evidence of the symptoms. Formally, then:-

Our aim is to clarify some of the central principles involved when, under conditions of uncertainty, we observe certain phenomena and, on that basis, together with various assumptions, we attempt to determine the degrees of confidence with which we can rationally believe assertions, derived from the observations by defined procedures, concerning states or conditions relating to the phenomena we have observed.

Acknowledgements

Throughout this work, we have had great support from many friends and institutions. Foremost among the institutions are the Australian Research Council, the Australian Centre for Sensor, Signal and Information Processing, the University of Melbourne, the University of Newcastle, New South Wales, and the former Siemens Plessey Company, now part of BAE. In all honesty, and despite the fact that it is decidedly un-fashionable to do so, we must also mention the enormous benefits we have enjoyed from the use of certain commercial products, notably Microsoft's 'Word' and The Mathwork's 'Matlab' - for which, we offer our thanks. We are also grateful to the Royal Society for providing us with a copy of Bayes' essay and for permission to reproduce the text.

To many friends, we are grateful. In Australia, Professor Brian Anderson and his wife Dianne have helped and encouraged us from the start, as has, for much of that time, Professor Henry d'Assumpcao. At Newcastle, Professor Graham Goodwin provided unstinting support and, at Melbourne we have had enjoyed the active interest of Professor Iven Mareels, who has provided support and encouragement in some pretty gloomy moments. We must also acknowledge the fundamentally challenging and motivating rôles of friends in the Royal Australian Air Force - Group Captains Ken Dee, Fred Lindsey, Gavin Thoms; and, later, in the Royal Air Force, Air Commodore Peter Eustace, all of whom seemed never to doubt for one moment our ability to deal with some remarkably mind-numbing situations. Among our friends and colleagues in 'the radar business' we are grateful to Peter Bates, John Hakes, George Hockham, Jack Hood, John Howard, Graham Kemp, Trudi Morgan and Alan Morley. Closer to home, Simon Gibbs has provided invaluable interest and encouragement, as have Walter, Jo and Pip Willcox, who have helped with many practical aspects, not the least of which was the re-typing of Bayes' essay. We thank Melissa Norfolk for her help in making this work available on the Web. We would also thank those 'beloved physicians', whose care and skill has enabled us to continue this work when, willing though the spirit, the flesh was in quite some peril. Finally, our wives and children, who, despite their unfailing scepticism, have unwaveringly supported us with their love.

Notation

Having considered various approaches to notation, we have aimed for clarity and simplicity wherever it seemed achievable. Hence, for example, multiplication is, where sensible, denoted explicitly by the symbol ' x ', but where this becomes too cumbersome the ' x ' is omitted and the algebraic notation in which, for example ' $a.b$ ' denotes ' a x b ' is used. To those who have had the benefit of a classical education in mathematics, this may seem trivial, or perhaps even condescending, yet we find that, in today's world, and especially in the English-speaking world, young people are increasingly denied the opportunities to learn classical mathematics which were readily available to previous generations. But the grim fact that we cannot assume familiarity with even moderately sophisticated notation, is not, to our minds, a reason to impede access to Bayes' thinking: the more so, because much of what Bayes has to say is couched in the English of his day rather than in the symbolic language of modern mathematics. There are however certain places where we have to use slightly sophisticated notation, and in such places we ask for the tolerance of those to whom the notation may have little or no meaning. We urge them to read on, for they may well find that, with time - years, perhaps - they have unconsciously absorbed a great deal of what may, at first sight, have seemed a tangle of meaningless symbols.

One small point concerns a conflict between the demands of exactitude and of clarity when discussing the 'probability at a point' and the 'probability density' on a continuous distribution. To deal with this problem, we presume that, when using plain English, it is acceptable to write of the probability at a point on a continuous distribution as if it were a perfectly ordinary quantity. In mathematical notation, however, we use the form $\mathcal{P}(a \approx x \mid k)$ rather than $\mathcal{P}(a = x \mid k)$ to remind ourselves that, if one goes much further into the matter, we encounter some rather sophisticated mathematical issues. Those issues are not of major importance in this context and, were we to indulge them, they could needlessly obstruct the access of many readers to the elements of the argument which really matter.

Version zh25a.doc at 18:36 on 13 July, 2001

So, the symbols we use are:-

C(.)	*a calibration table or function defining a distribution of error probabilities*
C(e)	*the calibrated probability of occurrence of an error* $\approx e$.
dF(x)	*the slope of an 'xy' curve at a point on the curve defined by a specific value of 'x'. If the curve is denoted as F(x), this is technically known as the differential co-efficient of F(x).*
E	*denotes the happening of a defined event in a defined trial, where it may not be known for certain in advance of the trial whether the event will or will not happen.*
exp (x)	*the number 'e' raised to the power 'x', where 'e' is the base of the natural logarithms. It is a transcendental number approximately equal to 2.7183*
F'(x)	*an alternative way of denoting the differential dF(x)*
F''(x)	*the second differential of a function F(x); it may be envisioned as the rate at which the slope of F(x) changes as we move along the x-axis.*
g(.)	*a shorthand to denote a function of several variables, such as ' g(x,m,n)' in which the symbols abbreviated to ' (.) ' can be seen from the context*
g'(.)	*the differential of g(.) : see F'(x) and dF(x)*
g''(.)	*the second differential of a function g(.) - see F''(x) above.*
$\mathcal{G}(\hat{a},\sigma_a, x)$	*a 'normal' or 'Gaussian' distribution of the probability a variable will have a value 'x', given that the mean of the distribution is* $\hat{a}$ *and the standard deviation, (i.e. the square root of the variance) is* σ_a
$\sim\mathcal{N}(a,\hat{\sigma})$	*similar to* $\mathcal{G}(\hat{a},\sigma_a, x)$, *specifies that a variable has a 'normal' or 'Gaussian' distribution about a mean 'a' with standard deviation* $\hat{\sigma}$
$I_Z(x)$	*an information-zero distribution of probabilities over the possible values of 'x'*

ln (x)	*the natural logarithm of 'x'*
$n \rightarrow \infty$	*' as the number 'n' tends to infinity'*
${}^{n}C_{m}$	*the number of different subsets which can be formed by selecting 'm' objects from a set of 'n' distinct objects. All subsets are different if each has at least one member which is not a member of any other subset.*
$\mathcal{P}(E)$	*the probability, expressed without further conditions or qualifications, that the event 'E' will happen in a defined trial. The same symbols may also denote the probability that an event 'E' happened in a trial when we know, or assume, only that such a trial has actually taken place and we have no further information which would affect our assessment of the probability of the outcome.*
$\mathcal{P}(\sim E)$	*the converse of $\mathcal{P}(E)$ i.e. the un-conditional probability that a defined event 'E' will not, or did not happen in a given trial. (But see also below for the use of the symbol '~' to denote a distribution in terms such as '$\varepsilon_i \sim \mathcal{N}(0,\sigma_i)$').*
$\mathcal{P}(E_2 \mid \dots)$	*the conditional probability that an event E_2 will occur in a given trial if it is known or assumed that an event denoted by the dots has occurred also in a given trial. The meaning of the dots is, in each case, determined by the particular context.*
$\mathcal{P}(E_2 \mid E_1)$	*the conditional probability that an event E_2 will occur in a given trial if it is known or assumed that an event E_1 has occurred also in a given trial.*
$\mathcal{P}(E_1 \wedge E_2)$	*the probability that an uncertain event E_1 happened, (or will happen), in a given trial **and** that another uncertain event E_2 also happened, (or will happen), in a given trial. It is important to note exactly how E_1 and E_2 are defined in each case.*
$\mathcal{P}(E_1 \vee E_2)$	*the probability that an uncertain event E_1 happened, in a given trial **or** that another uncertain event E_2 also happened, (or will happen), in a given trial.*

$\mathcal{P}_H(h \mid x)$	*the probability that an hypothesis 'h', concerning the value of a fixed parameter is true, if it is true that a random event 'x' has occurred.*
$\mathcal{P}_{H^*}(h \mid x)$	*the probability of a 'pseudo hypothesis' which actually relates to a random event*
$\mathcal{P}_R(W \mid k)$	*the probability that a measuring system is working as calibrated*
$\mathcal{P}_R(x \mid h)$	*the probability that a random event 'x' will occur if hypothesis 'h' is true.*
$P_u(x)$	*a uniform distribution of the 'a priori' probability density with respect to different values of x.*
$\mathcal{P}_F(X_{n+1} \approx x_g \mid X_n)$	*a special case of* $\mathcal{P}_R(X_i \approx x_g \mid X_j)$ *where* X_{n+1} *is a value derived from* X_n *via a forecasting function.*
$P_0(x)$	*An arbitrary 'a priori' distribution function, i.e. any such function that we care to specify.*
P_d	*in radar, etc., the probablistic frequency of detecting a defined signal in a noisy environment.*
P_m	*a probablistic frequency governing the occurrence of a random event*
$\$V$	*a monetary value of 'V' dollars*

$\forall(i,j)$	*for all values of 'i' and 'j'*
$\not<$	*not less than, e.g.* $x \not< y$ *means 'x is not less than y'*
$\not>$	*not greater than*
$\sum_{i=1}^{N} x_i$	*denotes the summation of a set of numbers, each of which is identified by assigning a specific, consecutive, integer value to the subscript 'i', starting with the lower value and ending with the upper value.*
$\prod_{i=1}^{n} x_i$	*denotes the product of a set of numbers, each of which is identified by assigning a specific, consecutive, integer value to the subscript 'i', starting with the lower value and ending with the upper value.*

$\int_{x_1}^{x_2} F(x)$	*Technically known as a 'definite integral', these symbols can be understood to denote the area between the x-axis and an 'xy' curve, denoted F(x), between points on the x-axis defined by x_1 and x_2 . Thus, if the 'y' axis denotes the density of the probability for a given value on the 'x' axis, the area under the curve gives the total probability of an event occurring somewhere between the points, x_1 and x_2.*
$[\ \mathcal{F}(x)\ ]_{x_1}^{x_2}$	*symbolises the value obtained by performing the subtraction $\mathcal{F}(x_1) - \mathcal{F}(x_2)$*
$\sim$	*' is distributed as ' e.g. $\varepsilon_i \sim \mathcal{N}(0,\sigma_i)$ denotes that ε_i is normally distributed about a zero mean, with variance σ_i^2*

Definitions

We quite often use the following expressions without further explanation. Where they occur, it may be assumed they have the meanings defined below:-

Bayes' question *:* 'Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named'.

Bayes' theorem *:* See the Précis above and chapter 3, equations *(3-51 thru 3-51f)*

Bayes' solution *:* In the *Scholium*, Bayes states:- *'..... therefore I shall take for granted that the rule given concerning the event M in Prop.9 is also the rule to be used in relation to any event concerning the probability of which nothing at all is known antecedently to any trials made or observed concerning it.'*. This invokes the assumption of a uniform prior distribution over the possible values of the unknown probability and gives as the solution to *Bayes' question:-*

Version zh25a.doc at 18:36 on 13 July, 2001

$$\mathcal{P}_H\{(x_1 < P_m < x_2) \mid (m, n, P_u(x))\}$$

$$= \frac{\int_{x_1}^{x_2} x^m (1 - x)^{(n - m)}\, dx}{\int_{0}^{1} x^m (1 - x)^{(n - m)}\, dx}$$

Further details are given in chapter 5.

delta function : This term is apparently not known to the compilers of mathematical dictionaries but is widely used in the signal processing world to denote a probability distribution in which the whole probability is concentrated at a unique value, such that the probability = *1* that an item selected at random from the relevant population will have that unique value.

Objective, Subjective : The meanings with which we use these words are, we believe, those of ordinary discourse in the English language. We have checked this by asking a number of people how they use these words. While most of those asked have given meanings which are pretty well identical with our own, the responses have not been entirely uniform and our meanings certainly differ from those often adopted in philosophical and statistical discourse. The dictionaries are ambivalent. Opting therefore to go with common usage, we use *'objective'* to mean an assessment of probability which is derivable from defined external evidence and assumptions, by a defined process of deduction which is generally accepted as valid and is such that any number of persons or automata could be taught or programmed to reach always the same assessment, given the same evidence and assumptions. We use *'subjective'* to mean a process of assessing probability which is subject to influences which are private to the agent making the assessment so that there are no reasons to believe that any two such agents will ever reach an equal assessment given the same external evidence and assumptions.

Random event : An event which could occur in any or every trial of a set, but we have no way of knowing whether it will or will not occur in any given trial. The occurrence of a random event within a set may, or may not, be governed by an underlying fixed probability having a value which can be arbitrarily close to one, or to zero, or have any value between those limits. An event can also be, but unknown to us, 'deterministic' in a defined type of trial, in which case the underlying probability in such a trial is either one or zero. We therefore view randomness and determinism as being assessed strictly relative to the available information.

Relative, Absolute : We use *'relative'* to mean probabilities which are assessed or asserted on the basis of - in relation to - evidence and assumptions. We use *'absolute'* to mean probabilities which are asserted independently of and without reference to any evidence or assumptions. We therefore presume that the only rational and meaningful assertions of absolute probabilities are in hypothetical statements of the form: *'If a defined event in a defined situation has a probability of occurrence 'x', then .....'*.. While such statements may be absolutely valid, they can be dangerous, as we show in *Chapter 11*.

Language

In considering language and style, we are painfully aware that many of our sentences are long and complex but, all too often, we found that, in trying to use shorter sentences, we were unable to achieve precision.

The second issue is the choice of a relative pronoun, where we tend to use 'which' rather than 'that'. This goes against certain current views on the politically correct use of these words - albeit such views stretch back, it seems, at least to Fowler[1]. It is our view, however, that the subject we address requires the finest possible nuances of language to achieve the precision which is its central purpose. Therefore, with a few exceptions, we use 'which'.

The final point concerns the word 'data'. In Latin, this is a neuter plural noun and would normally, in Latin, attract the plural of the verb. We are, however, writing English and there is no law which requires a word which is adopted into another language to preserve all the attributes of the original. This is especially true when the meaning in the adopting language is rather different from the meaning in the original language. We therefore generally follow academic convention by treating 'data' as a plural form, but, noting that, in classical Greek, a neuter plural noun is often used with a singular form of the verb, we allow 'data', occasionally, to be singular in English also.

[1] Fowler (1930). We are grateful to Professor Donald Knuth who outlined the history of this phenomenon in an after-dinner address to the 1999 gathering of UK TEX society.

Chapter 1

Introduction

The Philosophical Transactions of the Royal Society for the year 1763 present a paper entitled *An Essay towards solving a Problem in the Doctrine of Chances*[1]. The essay had been written by the Rev. Thomas Bayes and was found among his papers after his death. It was sent for publication by his friend, Richard Price. The problem, to which the title of the essay refers, is this:-

> *Given the number of times in which an unknown event has happened and failed : Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named .*

The essay was published at a relatively early point in the modern style of scientific enquiry, and it is arguable that the problem is the most basic in science, for it involves no rulers, no clocks, no dimensional parameters of any kind, but simply the counting of events. It is therefore, in our view, a matter for deep concern that Bayes' problem and the deeper questions to which it leads have been largely ignored by the scientific community: R.A.Fisher being a notable exception who fully perceived the serious nature of the issues[2]. Often, however, the issues have been regarded as trivial snags in the theory of statistics with few, if any, really serious implications. Wide sections of the professional communities in engineering, econometrics and sociology have, perhaps unconsciously, 'looked the other way', and refrained from pursuing issues which required Bayes' question to be answered.

Yet, whether the issues are truly trivial, and can be ignored, we shall find cause to doubt. For this book follows the trail of an investigation, which starts in the middle of the eighteenth century, and concludes - or, rather, pauses to take stock - in today's world, where it is increasingly unacceptable and impossible to ignore the issues raised by Bayes' essay. Matters as wide-ranging as air traffic control, data mining, ethical and risk management problems in clinical medicine and in health care policy, all require Bayes' question to be answered. The issues raised by the essay are fundamental in *Monte Carlo* simulation which is increasingly used with computer models of complex phenomena in, for example, electrical engineering,

---

[1] Bayes (1763), p 376

[2] His views are discussed in detail in Chapter 7.

biology and numerical analysis. Therefore, rather than simply accept today's conventional view, we go back to Bayes' essay, written at a time when mathematical symbolism was still largely unformed. Where today we might simply say, 'let *'x'* denote such-and-such', Bayes had to perform much of the algebra by using whole phrases as the variables. Quite some patience is required to follow closely his reasoning, and this must be, at least in part, a reason why, after nearly *250* years, his analysis has still not received the close attention from the wider philosophical and scientific community which we believe it merits.

The importance of the essay for real-life issues arises largely because Bayes' approach to Probability is based adamantly in the sort of real-life experience which is directly accessible to the vast majority of humankind. He does not define 'probability in itself' but defines instead how probability is to be measured. For his definition, he writes of probability in terms of trials and events. He defines the probability of an event as *'the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening'*. That is, if we have a ticket which will be worth *$1000* if a certain event happens and the value which 'ought' to be placed on that ticket - the value of the 'expectation' - is *$200*, then the ratio *$200 / $1000*, *i.e. 0.2*, is the probability of the event. As we shall see, this definition implies that probability shall be computed relative to defined evidence and assumptions. We shall also see how important it is to use a notation which explicitly designates the evidence and assumptions on which the computations are based. Under Bayes' definition, a trial concerning an event can be repetitive or unique: the definition fits both. This turns out to be a point of crucial significance when we come to examine the relevance of, for example, population statistics to an individual case. Bayes' definition also leads directly to concepts such as 'probable gain', 'probable value' and 'probable cost' which can be of enormous help in achieving rational decisions in uncertain situations. That is not to claim that such decisions are correct or optimal; but that they are rational in the sense that we can explain exactly how they are reached and allow others to test our evidence, assumptions and reasoning.

It is for such reasons that we are so concerned with the exposition of a philosophical essay which was not even published by its author in his own lifetime, but left in his papers to be discovered and published after his death. After that, it was largely ignored for about 100 years, then abused for a further 100 years, and is only nowadays starting to receive close and sympathetic attention. The concept of Probability is of enormous importance, in

pure science, in engineering, medicine, navigation and in making everyday decisions, where often, despite large uncertainties, one has a moral obligation, not only to act, but to act on a rational basis.

It is important also to understand the basic place of the issues raised by Bayes' essay once it is accepted that all measurements are subject to uncertainty. In earlier times, the issue of uncertainty was, understandably, overlooked in the excitement of discovering simple and elegant mathematical models of natural phenomena. Yet once it is accepted that uncertainty is inevitably present in measurement and observation, there is a correspondingly inexorable requirement to face the issues of probability which Bayes presents in stark and, arguably, irreducible simplicity.

Bayes' approach is, however, quite unlike that of, say, Kolmogorov[1], whose approach is purely directed at the axiomatisation of the theory of random phenomena and has no declared connection with situations in which we use evidence, inference and experience in dealing with uncertainty and making decisions. In many situations, people need to know plain probabilities - in order that they may proceed to rational decisions by clear reasoning about probable benefits and costs. Too often, however, they have to accept attenuated substitutes - likelihoods and confidence limits - which provide little or no clear basis for the making of rational decisions. In defence systems, medicine and commerce, the result is a quagmire where people need a firm theoretical basis. In a radar surveillance system, for example, a controller needs information in plain terms such as:- *"On the standard assumptions, there is a probability in excess of 99% that the height of this aircraft is above 30,000ft."* In respected texts concerning radar signal processing and tracking, however, authors such as van Trees and Bar-Shalom[2] choose their words with cautious reserve; as does Pollock, whose work stems largely from econometrics[3]. In archaeology, the report of a very precise dating for a pre-historic timber circle appears to use quite marked circumlocutions, goes so far as to mention 'probability' in the annotation of a diagram, but nowhere does it plainly ascribe a probability to the remarkable result[4]. The newspa-

---

[1] Kolmogorov(1933), English translation (1950). But see v.Plato(1994), Ch.7, for an account of Kolmogorov's acute interest in the physical applications of probability and the probable reasons why he chose an abstract axiomatic basis for his theory.

[2] van Trees (1968), Bar-Shalom (1988)

[3] Pollock (1999).

[4] Bayliss (1999).

pers, however, were not quite so reticent, the *'The Times'* reporting that 'an exact date' had been established[1].

But, as we shall see, Bayes' approach requires us to make assumptions and one in particular, the postulate of the 'uniform prior' has repeatedly been given as a reason for denying the validity of Bayes' conclusion. When one reads, however, the fulminations against Bayes, and against the assignment of probabilities to hypotheses, by authors such as Boole, Fisher, Keynes and Fine, it is scarcely surprising that people should walk with caution over the cracks in this academic volcano. Yet we are forced to suspect that many who have fulminated against Bayes may have been rather hypocritical: for who can doubt that the whole of science is based upon assumptions to which, in the last resort, we accede only as a matter of personal choice? Seen in that light, Bayes' assumption is, we suggest, trivial compared with many others which are routinely taken for granted. Yet we then have to ask why the postulate of the uniform prior became such a stumbling block? Can so many people have been so obstructive for so long without good cause? As we shall see, the answers to these questions begin, rather subtly in what Bayes himself wrote. Later, and more blatantly, we find illicit use, by others, of approaches which are superficially similar to that of Bayes, but in problems which Bayes did not address. Fortunately, we are able to show that Bayes' approach can be applied, coherently, to a wide range of cases which are of great practical importance and without entailing the apparent contradictions which have for so long obstructed its use.

Although Bayes' essay attracted little attention for many years after its publication, it seems that, at some time in the nineteenth century, the method of Bayes became associated with certain other approaches to the general problem of 'Inverse Probability' and with various fallacies and contradictions. These seem to have stemmed largely from the work of Laplace, who casually associated the name of Bayes with his own approach. This caused Bayes' name to become the focus of controversy and doubts which continue to the present time and in which too little attention has been paid to what Bayes himself actually wrote. Fortunately, more recent authors[2] have pointed out that even eminent authorities may have seriously misunderstood Bayes' argument. In our view, however, those authorities often neglected and misunderstood the precise problem which Bayes set out to address, and, by focussing on other issues, they distracted attention from a question which is of considerable importance in many areas of human endeavour. This is

---

[1] *'The Times'*, 2 Dec 1999, p15.

[2] Especially Stigler (1986a).

especially the case where we set out to observe the number of occasions on which an event occurs during a series of trials, and thence to infer the underlying probability of occurrence. (Some of the many assumptions which are taken for granted in this simple formulation, we examine in later chapters: for the moment, we take the formulation simply as it stands).

But it would be unfair to lay all the blame for the misrepresentation of Bayes' original work at the feet of later commentators, for the broadening of the issues began with the covering letter written by Richard Price, which was printed as an introduction to the essay on its first publication. It is for such reasons that in Chapter 2, we reproduce, with the kind agreement of The Royal Society, the introductory letter by Richard Price, followed by the Essay as far as page 394 in the original publication. We would therefore encourage the reader to work gently through Chapter 2, at least once, in order to get the flavour of Bayes' writing and to be more able than otherwise to examine critically our discussion of his argument.

In Chapter 3 we begin our own exposition of Bayes' argument, taking the essay as far as Proposition 9, (*i.e.* up to but stopping short of the *Scholium)*. We paraphrase the argument where this seems necessary to make the reasoning clear for the reader of our own times. For, as Stigler[1] points out, the essay is difficult: it stymied Bayes' own contemporaries, and, for a modern reader to understand the essay in its original form may require many hours of concentrated attention: the print is antiquated in style, important words are used with meanings which we today find strained, the reasoning makes extensive use of whole phrases, where we would today use mathematical symbols, and it is in a style which was old-fashioned even when it was written.

In Chapter 4 we offer an exposition of Bayes' experiment, in which balls are thrown randomly onto a plain and level table. Chapter 5 is an exposition of the *Scholium* which Bayes placed within the essay and in which he discusses the need to assume the uniform prior distribution of probability. In Chapter 6 we expound the concepts of probability and expectation as used by Bayes. To that point, therefore, our approach is essentially one of uncritical exposition and we hold for later chapters the criticisms and reservations which have been expressed concerning the arguments which Bayes deploys to solve the problem he has set.

In Chapter 7, we discuss the views of various *Critics and Defenders*, starting with Richard Price. We note their failure to read and address what

[1] Stigler (1982).

Bayes actually wrote in the essay, and how later critics preferred to address things which other authors, such as Laplace, but not Bayes, may have written. We then analyse in some detail, criticisms levelled against Bayes by R.A.Fisher, together with the daunting task of dealing with the claims which Fisher adduced for the 'fiducial argument'. This largely concludes our critical analysis of Bayes' essay and the classical views taken of it.

In Chapter 8, *'A Critical Case'*, we enter a more constructive phase by showing how we began the excavations which seemed necessary, if we were to find a viable basis for the use of Bayes' approach in real life. This takes the form of an example, where ethical value and probability come to an acute focus in deciding whether or not a certain treatment should be administered to a critically sick person. Having found that Bayes' approach provides, in the critical case, the only solution which is morally and rationally acceptable, we move, in Chapter 9, *'The Ruler'*, to examine some further foundations, this time in relation to the way in which professional people - navigators, doctors and engineers - actually use a measuring device, such as a ruler, in real life. This is particularly important for the light it throws on the problem which has besieged the use of Bayes' theory in questions of parametric measurement. That is, a prior probability which is uniformly distributed with respect to, say, the electrical resistance of a wire, becomes markedly non-uniform if we should choose instead to consider the current which flows in that same wire, given a constant voltage. The outcome of this excavation is a simple resolution of the difficulty and a view of what is meant by 'prior probability' rather different from that which has been assumed by many previous investigators, including, if not necessarily starting with, good Thomas Bayes himself.

In Chapter 10, *'The Individual'*, we explore the impact of this different view of a prior probability on the traditional views of the relationship in probability between an individual and a population. The result is somewhat disturbing for the traditional view. This line of exploration is taken further in Chapter 11, *'The Valid Prior'* to give a clarified view of the conditions under which prior knowledge can be used as a prior probability in the application of Bayes' theory. The outcome is a perspective which draws together the findings of the previous chapters.

In Chapter 12, *'The Trajectory'*, we apply Bayes' theory, seen in this new perspective, to the integration and refinement of repeated observations. This is a fundamental issue in navigation, surveying and in many electronic devices. There are widespread applications in radar and in the many scanning devices which have stemmed from radar[1], as there are also in telecommunications and in process control. We build on the results of the previous

[1] Buderi (1996) (*'The Invention that Changed the World'*.)

chapters to show how Bayes' theory leads rigorously, objectively, and without arbitrary or subjective assumptions, to a widely-used family of recursive integrators, of which the Kalman filter is one example.

Chapter 13, *'The Base Rate'* deals with an area where it is hard to exaggerate the importance of the issues, such are the widespread confusions between individuals and populations in matters so varied as medical diagnosis or the reliability of legal testimony. The roots of 'base rate thinking' seem to stem from certain cognitive psychologists who have developed a remarkably monocular view of how Bayes' theorem applies to the real world. It is an area where there has been an extremely serious failure to perceive the difference between probability in the sense of a frequency within a population and in the other sense of a degree of reason to believe a proposition concerning an individual, based on evidence concerning the individual in question.

In Chapter 14, *'The Probable Cause'*, we consider some wider issues of common interest in situations where Bayes' theory is, at least superficially, relevant to the diagnosis of causes. We note, however, that the diagnosis of causation can present some formidably difficult conceptual problems and that cut-and-dried conclusions are rarely as easy to reach as common sense and common parlance might seem to expect. In Chapter 15 we summarise what we have found and we present our conclusions, some of which are indeed disturbing when we consider the uses made of 'probability' in the reasoning of certain sections of our society.

As we move from chapter to chapter, the views we present are not homogeneous, as seen from the concluding position, but take the reader along the paths we traversed as we pursued the exploration and the perceptions slowly unfolded[1]. Throughout the exploration, however, the unifying aim is to understand the diagnosis - in probability - of situations which are inexorably uncertain, and to decide upon actions which are rational, given the evidence, and given the view of 'the rational' which is generally endorsed by the society in which we happen to live. We accept that our argument, in many places, sacrifices the formal rigour which is often demanded in pure mathematics, but is totally opaque to many intelligent, well-educated readers. Thus, because our concern is with probability as an instrument of practical reason, we have aimed for clarity and precision in the practical sense of prose writers, hoping to be understood [2].

---

[1] As exhorted by Medawar (1963) and Bondi (1967). See also Turkle (1995) p 59.

[2] *cf.* Keynes (1921) p vi

Chapter 2

The Essay

The following is copied from
The Philosophical Transactions of the Royal Society
Volume 53
1763

p370

LII. *An Essay towards solving a problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F.R.S. communicated by Mr.* Price, *in a Letter to* John Canton, *A.M. F.R.S.*

Dear Sir,

Read Dec. 23, 1763

I Now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illustrious Society, and was much esteemed by many in it as a very able mathematician. In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, *(page 371)* it has happened a certain number of times, and failed a certain other number of times. He adds, that he soon perceived that it would not be very difficult to do this, provided some rule could be found according to which we ought to estimate the chance that the probability for the happening of an event perfectly unknown, should lie between any two named degrees of probability, antecedently to any experiments made about it; and that it appeared to him that the rule must be to suppose the chance the same that it should lie between any two equidifferent degrees; which, if it were allowed, all the rest might be easily calculated in the common method of proceeding in the doctrine of chances. Accordingly, I find among his papers a very ingenious solution of this problem in this way. But he afterwards considered, that the postulate on which he had

argued might not perhaps be looked upon by all as reasonable; and therefore he chose to lay down in another form the proposition in which he thought the solution of the problem is contained, and in a scholium to subjoin the reasons why he thought so, rather than to take into his mathematical reasoning any thing that might admit dispute. This, you will observe, is the method which he has pursued in this essay.

Every judicious person will be sensible that the problem now mentioned is by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter. Common sense is indeed sufficient to shew us that, from the observation of what has in former instances been the consequence of a certain cause *(page 372)* or action, one may make a judgement what is likely to be the consequence of it another time, and that the larger number of experiments we have to support a conclusion, so much the more reason we have to take it for granted. But it is certain that we cannot determine, at least not to any nicety, in what degree repeated experiments confirm a conclusion without the particular discussion of the beforementioned problem; which, therefore, is a necessary to be considered by any one who would give a clear account of the strength of analogical or inductive reasoning; concerning, which at present, we seem to know little more than that it does sometimes in fact convince us, and at other times not; and that, as it is the means of [a]cquainting us with many truths, of which otherwise we must have been ignorant; so it is, in all probability, the source of many errors, which perhaps might in some measure be avoided, if the force that this sort of reasoning ought to have with us were more distinctly and clearly understood.

These observations prove that the problem enquired after in this essay is no less important than it is curious. It may be safely added, I fancy, that it is also a problem that has never before been solved. Mr. De Moivre, indeed, the great improver of this part of mathematics, has in his Laws of chance[1], after Bernoulli, and to a greater degree of exactness, given rules to find the probability there is, that if a very great number of trials be made concerning any event, *(page 373)* the proportion of the number of times it will happen, to the number of times it will fail in those trials, should differ less than by small assigned limits from the proportion of the probability of its happening to the probability of its failing in one single trial. But I know of no person

[1] See Mr. De Moivre's *Doctrine of Chances*, p. 243, &c. He has omitted the demonstrations of his rules, but these have been since supplied by Mr. Simpson at the conclusion of his treatise on *The Nature and Laws of Chance.*

who has shewn how to deduce the solution of the converse problem to this; namely, 'the number of times an unknown event has happened and failed being given, to find the chance that the probability of its happening should lie somewhere between any two named degrees of probability'. What Mr. De Moivre has done therefore cannot be thought sufficient to make the consideration of this point unnecessary: especially, as the rules he has given are not pretended to be rigorously exact, except on supposition that the number of trials made are infinite; from whence it is not obvious how large the number of trials must be in order to make them exact enough to be depended on in practice.

Mr. De Moivre calls the problem he has thus solved, the hardest that can be proposed on the subject of chance. His solution he has applied to a very important purpose, and thereby shewn that those are much mistaken who have insinuated that the Doctrine of Chances in mathematics is of trivial consequence, and cannot have a place in any serious enquiry[1]. The purpose I mean is, to shew what reason we have for believing that there are in the constitution of things fixt laws according to which events happen, and that, therefore, the frame of the world must be *(page 374)* the effect of the wisdom and power of an intelligent cause; and thus to confirm the argument taken from final causes for the existence of the Deity. It will be easy to see that the converse problem solved in this essay is more directly applicable to this purpose; for it shews us, with distinctness and precision, in every case of any particular order or recurrency of events, what reason there is to think that such recurrency or order is derived from stable causes or regulations in nature, and not from any of the irregularities of chance.

The two last rules in this essay are given without the deductions of them. I have chosen to do this because these deductions, taking up a good deal of room, would swell the essay too much; and also because these rules, though of considerable use, do not answer the purpose for which they are given as perfectly as could be wished. They are however ready to be produced, if a communication of them should be thought proper. I have in some places writ short notes, and to the whole I have added an application of the rules in the essay to some particular cases, in order to convey a clearer idea of the nature of the problem and to shew how far a solution of it has been carried.

I am sensible that your time is so much taken up that I cannot reasonably expect that you should minutely examine every part of what I now send you. Some of the calculations, particularly in the appendix, no one can make without a good deal of labour. I have taken so much care about them, that I

[1] See his Doctrine of Chances, p. 252, &c.

believe there can be no material error in any of them; but should there be any such errors, I am the only person who ought to be considered as answerable for them.

(page 375)

Mr. Bayes has thought fit to begin his work with a brief demonstration of the general laws of chance. His reason for doing this, as he says in his introduction, was not merely that his reader might not have the trouble of searching elsewhere for the principles on which he has argued, but because he did not know whither to refer him for a clear demonstration of them. He has also made an apology for the peculiar definition he has given of the word chance or probability. His design herein was to cut off all dispute about the meaning of the word, which in common language is used in different senses by persons of different opinions, and according as it is applied to past or future facts. But whatever different senses it may have, all (he observes) will allow that an expectation depending on the truth on any past fact, or the happening of any future event, ought to be estimated so much the more valuable as the fact is more likely to be true, or the event more likely to happen. Instead therefore, of the proper sense of the word probability, he has given that which all will allow to be its proper measure in every case where the word is used. But it is time to conclude this letter. Experimental philosophy is indebted to you for several discoveries and improvements; and, therefore, I cannot help thinking that there is a peculiar propriety in directing to you the following essay and appendix. That your enquiries may be rewarded with many further successes, and that you may enjoy every *[sic]* valuable blessing, is the sincere wish of, Sir,

Newington-Green,
Nov. 10 1763.

Your very humble servant,
Richard Price

[376]

PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

SECTION I.

DEFINITION 1. Several events are *inconsistent*, when if one of them happens, none of the rest can.

2. Two events are *contrary* when one, or other of them must; and both together cannot happen.

3. An event is said to *fail*, when it cannot happen; or, which comes to the same thing, when its contrary has happened.

4. An event is said to be determined when it has either happened or failed.

5. The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening.

6. By *chance* I mean the same as probability.

7. Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest.

PROP. 1.

When several events are inconsistent the probability of the happening of one or other of them is the sum of the probabilities of each of them.

Suppose there be three such events, and which ever of them happens I am to receive N, and that the probability of the 1st, 2d, and 3d are respectively $\frac{a}{\mathrm{N}}, \frac{b}{\mathrm{N}}, \frac{c}{\mathrm{N}}$. Then (by the definition of probability) the value of my expectation from the 1st will be a, from the 2d b, and from the 3d c. Wherefore the value of my expectations from all three will be $a + b + c$. But the sum of my expectations from all three is in this case an expectation of receiving N upon the happening of one or other of them. Wherefore (by definition 5) the

probability of one or other of them is $\frac{a+b+c}{N}$ or $\frac{a}{N}+\frac{b}{N}+\frac{c}{N}$. The sum of the probabilities of each of them.

Corollary. If it be certain that one or other of the three events must happen, then $a + b + c = N$. For in this case all the expectations together amounting to a certain expectation of receiving N, their values together must be equal to N. And from hence it is plain that the probability of an event added to the probability of its failure (or of its contrary) is the ratio of equality. For these are two inconsistent events, one of which necessarily happens. Wherefore if the probability of an event is $\frac{P}{N}$ that of it's failure will be $\frac{N-P}{N}$.

PROP. 2.

If a person has an expectation depending on the happening of an event, the probability of the event is to the probability of its failure as his loss if it fails to his gain if it happens.

Suppose a person has an expectation of receiving N, depending on an event the probability of which is $\frac{P}{N}$. *(page 378)* Then (by definition 5) the value of his expectation is P, and therefore if the event fail, he loses that which in value is P; and if it happens he receives N, but his expectation ceases. His gain therefore is N – P. Likewise since the probability of the event is $\frac{P}{N}$, that of its failure (by corollary prop. 1) is $\frac{N-P}{N}$. But $\frac{P}{N}$ is to $\frac{N-P}{N}$ as P is to N – P, i.e. the probability of the event is to the probability of it's failure, as his loss if it fails to his gain if it happens.

PROP. 3.

The probability that two subsequent events will both happen is a ratio compounded of the probability of the 1st, and the probability of the 2d on supposition that the 1st happens.

Suppose that, if both events happen, I am to receive N, that the probability both will happen is $\frac{P}{N}$, that the 1st will is $\frac{a}{N}$ (and consequently that the 1st will not is $\frac{N-a}{N}$) and that the 2d will happen upon supposition that the 1st does is $\frac{b}{N}$. Then (by definition 5) P will be the value of my expectation, which will become *b* if the 1st happens. Consequently if the 1st happens,

Version zh25a.doc at 18:36 on 13 July, 2001

my gain by it is b – P, and if it fails my loss is P. Wherefore, by the foregoing proposition, $\frac{a}{N}$ is to $\frac{N-a}{N}$ i.e. a is to N – a as P is to b – P. Wherefore (componendo inversè) a is to N as P is to b. But the ratio of P to N is compounded of the ratio of P to b, and that of b to N. Wherefore the *(page 379)* same ratio of P to N is compounded of the ratio of a to N and that of b to N, i.e. the probability that the two subsequent events both happen is compounded of the probability of the 1st and the probability of the 2d on supposition the 1st happens.

Corollary. Hence if of two subsequent events the probability of the 1st be $\frac{a}{N}$, and the probability of both together be $\frac{P}{N}$, then the probability of the 2d on supposition the 1st happens is $\frac{P}{a}$.

PROP. 4.

If there be two subsequent events to be determined every day, and each day the probability of the 2d is $\frac{b}{N}$ and the probability of both $\frac{P}{N}$, and I am to receive N if both the events happen the 1st day on which the 2d does; I say, according to these conditions, the probability of my obtaining N is $\frac{P}{b}$. For if not, let the probability of my obtaining N be $\frac{x}{N}$ and let y be to x as N – b to N. Then since $\frac{x}{N}$ is the probability of my obtaining N (by definition 1) x is the value of my expectation. And again, because according to the foregoing conditions the 1st day I have an expectation of obtaining N depending on the happening of both the events together, the probability of which is $\frac{P}{N}$, the value of this expectation is P. Likewise, if this coincident should not happen I have an expectation of being reinstated in my former circumstances, i.e. of receiving that which in value is x depending *(page 380)* on the failure of the 2d event the probability of which (by cor. prop. 1) is $\frac{N-b}{N}$ or $\frac{y}{x}$, because y is to x as N – b to N. Wherefore since x is the thing expected and $\frac{y}{x}$ the probability of obtaining it, the value of this expectation is y. But these two last expectations together are evidently the same with my original expectation, the value of which is x, and therefore P + y = x. But y is to x as N – b is to

N. Wherefore x is to P as N is to b and $\frac{x}{\mathrm{N}}$ (the probability of my obtaining N) is $\frac{\mathrm{P}}{b}$.

Cor. Suppose after the expectation given me in the foregoing proposition, and before it is at all known whether the 1st event has happened or not, I should find that the 2d event has happened; from hence I can only infer that the event is determined on which my expectation depended, and have no reason to esteem the value of my expectation either greater or less than it was before. For if I have reason to think it less, it would be reasonable for me to give something to be reinstated in my former circumstances, and this over and over again as often as I should be informed that the 2d event had happened, which is evidently absurd. And the like absurdity plainly follows if you say I ought to set a greater value on my expectation than before, for then it would be reasonable for me to refuse something if offered me upon condition I would relinquish it, and be reinstated in my former circumstances; and this likewise over and over again as often as (nothing being known concerning the 1st event) it should appear that the 2d had happened. Notwithstanding therefore this discovery that the 2d *(page 381)* event has happened, my expectation ought to be esteemed the same in value as before, i.e. x, and consequently the probability of my obtaining N is (by definition 5) still $\frac{x}{\mathrm{N}}$ or $\frac{\mathrm{P}}{b}$ [1]. But after this discovery the probability of my obtaining N is the probability that the 1st of two subsequent events has happened upon the supposition that the 2d has, whose probabilities were as before specified. But the probability that an event has happened is the same as the probability I have to guess right if I guess it has happened. Wherefore the following proposition is evident.

PROP. 5.

If there be two subsequent events, the probability of the 2d $\frac{b}{\mathrm{N}}$ and the probability of both together $\frac{\mathrm{P}}{\mathrm{N}}$, and it being 1st discovered that the 2d event

[1] What is here said may perhaps be a little illustrated by considering that all that can be lost by the happening of the 2d event is the chance I should have had of being reinstated in my former circumstances, if the event on which my expectation depended had been determined in the manner expressed in the proposition. But this chance is always as much *against* me as it is *for* me. If the 1st event happens, it is *against* me, and equal to the chance for the 2d event's failing. If the 1st event does not happen, it is *for* me, and equal also to the chance for the 2d event's failing. The loss of it, therefore, can be no disadvantage.

has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is $\frac{P}{b}$ [1].

(page 382) PROP. 6.

The probability that several independent events shall all happen is a ratio compounded of the probabilities of each.

For from the nature of independent events, the probability that any one happens is not altered by the happening or failing of any of the rest, and consequently the probability that the 2d event happens on supposition the 1st does is the same with its original probability; but the probability that any two events happen is a ratio compounded of the probability of the 1st event, and the probability of the 2d on supposition the 1st happens by prop. 3. Wherefore the probability that any two independent events both happen is a ratio compounded of the probability of the 1st and the probability of the 2d. And in like manner considering the 1st and 2d event together as one event; the probability that three independent events all happen is a ratio compounded of the probability that the two 1st both happen and the probability of the 3d. And thus you *(page 383)* may proceed if there be ever so many such events; from whence the proposition is manifest.

Cor. 1. If there be several independent events, the probability that the 1st happens the 2d fails, the 3d fails and the 4th happens, &c. is a ratio compounded of the probability of the 1st, and the probability of the failure of the 2d, and the probability of the failure of the 3d, and the probability of the 4th, &c. For the failure of an event may always be considered as the happening of its contrary.

Cor. 2. If there be several independent events, and the probability of each one be *a*, and that of its failure be *b*, the probability that the 1st happens and the 2d fails, and the 3d fails and the 4th happens, &c. will be *a b b a,* &c.

---

[1] What is proved be Mr. Bayes in this and the preceding proposition is the same with the answer to the following question. What is the probability that a certain event, when it happens, will be accompanied with another to be determined at the same time? In this case, as one of the events is given, nothing can be due for the expectation of it; and, consequently, the value of an expectation depending on the happening of both events must be the same with the value of an expectation depending on the happening of one of them. In other words; the probability that, when one of two events happens, the other will, is the same with the probability of this other. Call *x* then the probability of this other, and if $\frac{b}{N}$ be the probability of the given event, and $\frac{p}{N}$ the probability of both, because $\frac{p}{N} = \frac{b}{N} \times x$, $x = \frac{p}{b}$ = the probability mentioned in these propositions.

Version zh25a.doc at 18:36 on 13 July, 2001

For, according to the algebraic way of notation, if *a* denote any ratio and *b* another, *a b b a* denotes the ratio compounded of the ratios *a*, *b*, *b*, *a*. This corollary therefore is only a particular case of the foregoing.

Definition. If in consequence of certain data there arises a probability that a certain event should happen, its happening or failing, in consequence of these data, I call it's happening or failing in the 1st trial. And if the same data be again repeated, the happening or failing of the event in consequence of them I call its happening or failing in the 2d trial; and so on as often as the same data are repeated. And hence it is manifest that the happening or failing of the same event in so many diffe- trials is in reality the happening or failing of so many distinct independent events exactly familiar to each other.

(page 383) PROP. 7.

If the probability of an event be *a*, and that of its failure be *b* in each single trial, the probability of its happening *p* times, and failing *q* times in *p* + *q* trials is E $a^p\ b^q$ if E be the coefficient of the term in which occurs $a^p\ b^q$ when the binomial $\overline{a+b}|^{p+q}$ is expanded.

For the happening or failing of an event in different trials are so many independent events. Wherefore (by cor. 2. prop. 6.) the probability that the event happens the 1st trial, fails the 2d and 3d, and happens the 4th, fails the 5th, &c. (thus happening and failing till the number of times it happens be *p* and the number it fails be *q* is *a b b a b* &c. till the number of *a*'s be *p* and the number of *b*'s be *q*, that is; 'tis $a^p\ b^q$. In like manner if you consider the event as happening *p* times and failing *q* times in any other particular order, the probability for it is $a^p\ b^q$; but the number of different orders according to which an event may happen or fail, so as in all to happen *p* times and fail *q*, in *p* + *q* trials is equal to the number of permutations that *a a a a b b b* admit of when the number of *a*'s is *p*, and the number of *b*'s is *q*. And this number is equal to E, the coefficient of the term in which occurs $a^p\ b^q$ when $\overline{a+b}|^{p+q}$ is expanded. The event therefore may happen *p* times and fail *q* in *p* + *q* trials E different ways and no more, and its happening and failing these several different ways are so many inconsistent events, the probability for each of which is $a^p\ b^q$, and therefore by *(page 385)* prop. 1. the probability that some way or other it happens *p* times and fails *q* times in *p* + *q* trials is E $a^p\ b^q$.

Version zh25a.doc at 18:36 on 13 July, 2001

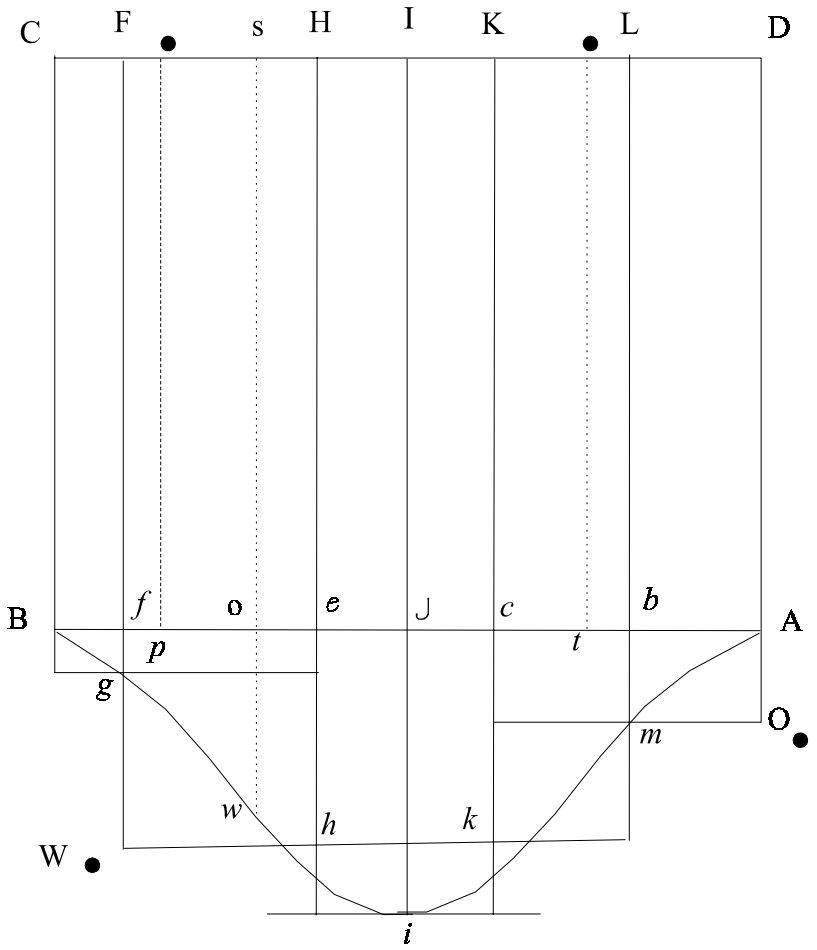
SECTION II.

Postulate. 1. I Suppose the square table or plane A B C D to be so made and levelled, that if either of the balls *o* or W be thrown upon it, there shall be the same probability that it rests upon any one equal part of the plane as another, and that it must necessarily rest somewhere upon it.

2. I suppose that the ball W shall be 1st thrown, and through the point where it rests a line *o s* shall be drawn parallel to A D, and meeting C D and A B in *s* and *o*; and that afterwards the ball O shall be thrown $p + q$ or n times, and that its resting between A D and *o s* after a single throw be called the happening of the event M in a single trial. These things supposed,

Lem. 1. The probability that the point *o* will fall between any two points in the line A B is the ratio of the distance between the two points to the whole line A B.

Let any two points be named, as *f* and *b* in the line A B, and through them parallel to A D draw *f F*, *b* L meeting C D in F and L. Then if the rectangles C *f*, F *b*, L A are *(page 386)* commensurable to each other, they may each be divided into the same equal parts, which being done, and the ball W thrown, the probability it will rest somewhere upon any number of these equal parts will be the sum of the probabilities it has to rest upon each one of them, because its resting upon any different parts of the plane A C are so many inconsistent events; and this sum, because the probability it should rest upon any one equal part as another is the same, is the probability it should rest upon any one equal part multiplied by the number of parts. Consequently, the probability there is that the ball W should rest somewhere upon *(page 387)* F *b* is the probability it has to rest upon one equal part multiplied by the number of equal parts in F *b*; and the probability it rests somewhere upon C *f* or L A, i.e. that it dont rest upon F *b* (because it must rest somewhere upon A C) is the probability it rests upon one equal part multiplied by the number of equal parts in C *f*, L A taken together. Wherefore, the probability it rests upon F *b* is to the probability it dont as the number of equal parts in F *b* is to the number of equal parts in C *f*, L A together, or as F *b* to C *f*, L A together, or as *f b* to B *f* A *b* together. Wherefore the probability it rest upon F *b* is to the probability it dont as *f b* to B *f*, A *b* together. And (*componendo inverse*) the probability it rests upon F *b* is to the probability it rests upon F *b* added to the probability it dont, as *f b* to A B, or as the ratio of *f b* to A B to the ratio of A B to A B. But the probability of any event added to the probability of its failure is the ratio of equality; wherefore, the probability it rest upon F *b* is to the ratio of equality as the ratio of *f b* to A B to the ratio of A B to A B, or the ratio of equality; and therefore the probability it rest upon F *b* is the ratio of *f b* to A B. But *ex hypothesi* according as the ball W falls



upon F *b* or not the point *o* will lie between *f* and *b* or not and therefore the probability the point *o* will lie between *f* and *b* is the ratio of *f b* to A B.

Again; if the rectangles C *f*, F *b*, L A are not commensurable, yet the last mentioned probability can be neither greater nor less than the ratio of *f b* to A B; for, if it be less, let it be the ratio of *f c* to A B, and upon the line *f b* take the points *p* and *t*, so that *p t* shall be greater than *f c*, and the three lines B *p*, *p t*, *t* A commensurable (which it is evident may be always done by dividing A B into equal parts less than half *c b*, and taking *p* and *t* the nearest points of division to *f* and *c* that lie upon *f b*). Then because B *p*, *p t*, *t* A are commensurable, so are the rectangles C *p*, D *t*, and that upon *p t* compleating the square A B. Wherefore, by what has been said, the probability that the point *o* will lie between *p* and *t* is the ratio of *p t* to A B. But if it lies between *p* and *t* it must lie between *f* and *b*. Wherefore, the probability it should lie between *f* and *b* cannot be less than the ratio of *p t* to A B, and therefore must be greater than the ratio of *f c* to A B (since *p t* is greater than *f c*). And after the same manner you may prove that the forementioned

probability cannot be greater than the ratio of *f b* to A B, it must therefore be the same.

Lem. 2. The ball W having been thrown, and the line *o s* drawn, the probability of the event M in a single trial is the ratio of A *o* to A B.

For, in the same manner as the foregoing lemma, the probability that the ball *o* being thrown shall *(page 388)* rest somewhere upon D *o* or between A D and *s o* is the ratio of A *o* to A B. But the resting of the ball *o* between A D and *s o* after a single throw is the happening of the event M in a single trial. Wherefore the lemma is manifest.

PROP. 8.

If upon B A you erect the figure B *g h i k m* A whose property is this, that (the base B A being divided into any two parts, as A B, and B *b* and at the point of division *b* a perpendicular being erected and terminated by the figure in *m*; and *y*, *x*, *r* representing respectively the ratio of *b m*, A *b*, and B *b* to A B, and E being the coefficient of the term in which occurs $a^p\ b^q$ when the binomial $\overline{a+b}|^{p+q}$ is expanded) $y = \mathrm{E}\ x^p\ r^q$. I say that before the ball W is thrown, the probability the point *o* should fall between *f* and *b*, any two points named in the line A B and withall that the event M should happen *p* times and fail *q* in *p* + *q* trials, is the ratio of *f g h i k m b*, the part of the figure B *g h i k m* A intercepted between the perpendiculars *f g*, *b m* raised upon the line A B, to C A the square upon A B.

DEMONSTRATION

For if not; 1st let it be the ratio of D a figure greater than *f g h i k m b* to C A, and through the points *e d c* draw perpendiculars to *f b* meeting the curve A *m i g* B in *h*, *i, k*; the point *d* being so placed that *d i* shall be the longest of the *(page 389)* perpendiculars terminated by the line *f b*, and the curve A *m i g* B; and the points *e*, *d*, *c* being so many and so placed that the rectangles, *b k*, *c i*, *e i*, *f h* taken together shall differ less from *f g h i k m b* than D does; all which may be easily done by the help of the equation of the curve, and the difference between D and the figure *f g h i k m b* given. Then since *d i* is the longest of the perpendicular ordinates that insist upon *f b*, the rest will gradually decrease as they are farther and farther from it on each side, as appears from the construction of the figure, and consequently *e h* is greater than *g f* or any other ordinate that insists upon *e f*.

Now if A *o* were equal to A *e*, then by lem. 2. the probability of the event M in a single trial would be the ratio of A *e* to A B, and consequently by cor. Prop. 1. the probability of it's failure would be the ratio of B *e* to A B. Wherefore, if *x* and *r* be the two forementioned ratios respectively, by Prop. 7. the probability of the event M happening *p* times and failing *q* in *p* + *q* trials would be $\mathrm{E}\ x^p\ r^q$. But *x* and *r* being respectively the ratios of A *e* to A

Version zh25a.doc at 18:36 on 13 July, 2001

B and B *e* to A B, if *y* is the ratio of *e h* to A B, then, by construction of the figure A *i* B, $y = \mathrm{E}\ x^p r^q$. Wherefore, if A *o* were equal to A *e* the probability of the event M happening *p* times and failing *q* in *p* + *q* trials would be *y*, or the ratio of *e h* to A B. And if A *o* were equal to A *f* or were any mean between A *e* and A *f*, the last mentioned probability for the same reasons would be the ratio of *f g* or some other of the ordinates insisting upon *e f*, to A B. But *e h* is the greatest of all the ordinates that insist upon *e f*. Wherefore, upon supposition the point should lie *(page 390)* any where between *f* and *e*, the probability of the event M happens *p* times and fails *q* in *p* + *q* trials can't be greater than the ratio of *e h* to A B. There then being these two subsequent events, the 1st that the point *o* will lie between *e* and *f*, the 2d that the event M will happen *p* times and fail *q* in *p* + *q* trials, and the probability of the 1st (by lemma 1st) is the ratio of *e f* to A B, and upon supposition the 1st happens, by what has been now proved, the probability of the 2d cannot be greater than the ratio of *e h* to A B, it evidently follows (from Prop. 3.) that the probability both together will happen cannot be greater than the ratio compounded of that of *e f* to A B and that of *e h* to A B, which compound ratio is the ratio of *f h* to C A. Wherefore, the probability that the point *o* will lie between *f* and *e,* and the event M happen *p* times and fail *q*, is not greater than the ratio of *f h* to C A. And in like, manner the probability the point *o* will lie between *e* and *d*, and the event M happen and fail as before, cannot be greater than the ratio of *e i* to C A. And again, the probability the point *o* will lie between *d* and *c*, and the event M happen and fail as before, cannot be greater than the ratio of *c i* to C A. And lastly, the probability that the point *o* will lie between *c* and *b*, and the event M happen and fail as before, cannot be greater than the ratio of *b k* to C A. Add now all these several probabilities together, and their sum, (by Prop. 1.) will be the probability that the point will lie somewhere between *f* and *b*, and the event M happen *p* times and fail *q* in *p* + *q* trials. Add likewise the correspondent ratios together, and their sum will be the ratio of the sum of the antecedents *(page 391)* to their common consequent, i.e. the ratio of *f h*, *e i*, *c i*, *b k* together to C A; which ratio is less than that of D to C A, because D is greater than *f h*, *e i*, *c i*, *b k* together. And therefore, the probability that the point *o* will lie between *f* and *b* and withal that the event M will happen *p* times and fail *q* in *p* + *q* trials, is *less* than the ratio of D to C A; but it was supposed the same which is absurd. And in like manner, by inscribing rectangles within the figure, as *e g*, *d h*, *d k*, *c m*, you may prove that the last mentioned probability is *greater* than the ratio of any figure less than *f g h i k m b* to C A.

Wherefore that probability must be the ratio of *f g h i k m b* to C A.

Cor. Before the ball W is thrown the probability that the point *o* will lie somewhere between A and B, or somewhere upon the line A B, and withal

that the event M will happen *p* times, and fail *q* in *p* + *q* trials is the ratio of the whole figure A *i* B to C A. But it is certain that the point *o* will lie somewhere upon A B. Wherefore, before the ball W is thrown the probability the event M will happen *p* times and fail *q* in *p* + *q* trials is the ratio of A *i* B to C A.

PROP. 9.

If before any thing is discovered concerning the place of the point *o*, it should appear that the event M had happened *p* times and failed *q* in *p* + *q* trials, and from hence I guess that the point *o* lies between any two points in the line A B, as *f* and *b*, and consequently that the probability of the event M in a single trial was somewhere between the ratio of A *b* to A B and that of A *f* to A B: the probability I am in the right is the ratio of that part of the figure A *i* B described as before which is intercepted between perpendiculars erected upon A B at the points *f* and *b* to the whole figure A *i* B.

For, there being these two subsequent events, the first that the point *o* will lie between *f* and *b*, the second that the event M should happen *p* times and fail *q* in *p* + *q* trials and (by cor. Prop. 8.) the original probability of the second is the ratio of A *i* B to C A, and (by prop. 8.) the probability of both is the ratio of *f g h i m b* to C A; wherefore (by prop. 5) it being first discovered that the second has happened, and from hence I guess that the first has happened also, the probability I am in *(page 392)* the right is the ratio of *f g h i m b* to A *i* B, the point which was to be proved.

Cor. The same things supposed, if I guess that the probability of the event M lies somewhere between *o* and the ratio of A *b* to A B, my chance to be in the right is the ratio of A *b m* to A *i* B.

SCHOLIUM.

From the preceding proposition it is plain, that in the case of such an event as I there call M, from the number of times it happens and fails in a certain number of trials, without knowing any thing more concerning it, one may give a guess whereabouts it's probability is, and, by the usual methods computing the magnitudes of the areas there mentioned, see the chance that the guess is right. And that the same rule is the proper one to be used in the case of an event concerning the probability of which *(page 393)* we absolutely know nothing antecedently to any trials made concerning it, seems to appear from the following consideration; viz. that concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another. For, on this account, I may justly reason concerning it as if its probability had been at first unfixed, and then determined in such a manner as to give me no reason to think that, in a certain number of trials, it should rather happen any one possible number of

times than another. But this is exactly the case of the event M. For before the ball W is thrown, which determines it's probability in a single trial, (by cor. prop. 8.) the probability it has to happen p times and fail q in $p + q$ or n trials is the ratio of A *i* B to C A, which ratio is the same when $p + q$ or n is given, whatever number p is; as will appear by computing the magnitude of A *i* B by the method[1] of fluxions. And consequently before the place of the point *o* is discovered or the number of times event M has happened in n trials, I can have no reason to think it should rather happen one possible number of times than another.

In what follows therefore I shall take for granted that the rule given concerning the event M in prop. 9. is also the rule to be used in relation to any event concerning the probability of which nothing *(page 394)* at all is known and antecedently to any trials made or observed concerning it. And such an event I shall call an unknown event.

[1] It will be proved presently in art. 4. By computing in the method here mentioned that A *i* B contracted in the ratio of E to 1 is to C A as 1 to $\overline{n + 1}^{1} \times E$: from whence it plainly follows that, antecedently to this contraction, A *i* B must be to C A in the ratio of 1 to n + 1, which is a constant ratio when n is given, whatever p is.

Version zh25a.doc at 18:36 on 13 July, 2001

Chapter 3

The Elements of Bayes' Probability

In this chapter we take the first section of Bayes' essay, element by element in order to clarify the meaning in terms more easily understandable in our own time. Although all exposition of the writings of others involves subjective judgement as to what things are important, it is the purpose of later chapters to evaluate critically Bayes' argument: our purpose here is to clarify the essay point by point and to preserve the form of Bayes' original reasoning. This approach differs from treatments by previous commentators, for we do not re-cast Bayes' argument into a more modern form. However we have attempted in places to clarify the notation, *e.g.* by using the *'$'* sign to denote (monetary) value.

Although Bayes' style of writing is not entirely homogeneous, the structure of his argument is simple: first, the problem is defined; next a set of terms are defined, after which we are given a series of propositions and corollaries derived progressively from the previous material, to establish certain basic rules for the manipulation of probabilities. The probabilities are generally expressed as ratios of numbers, and it is unfortunate that Bayes' adherence to symbolising probabilities in this way rather than as 'real numbers' adds considerably to the superficial complexity of his argument. On the other hand, Bayes' approach has the great advantage of allowing the argument to be presented in simple terms and plain concepts with which wide ranges of practical people as well as mathematicians and other scholars will be familiar. In Section 2 of the essay, although there are variations in local structure, the general tone is preserved until after the Scholium, where the essay suddenly becomes concentrated upon a rather narrow mathematical problem, which we do not discuss[1].

The tone in the earlier part of the essay is therefore in the tradition of dealing with philosophical and logical issues in plain language, with the aim of being read and understood, rather than with the aim of achieving logical perfection by way of symbols and manipulations which may be very hard to

[1] Bayes' treatment is not particularly successful. It is discussed in detail by Stigler(1986)

relate to real life. The penalty of the plain language approach is, that it may lack, or seem to lack, analytical depth: the penalty of the more formal approach is that it often fails to carry real-life conviction: the arguments may overwhelm, but fail totally to persuade[1]. Somewhat ironically, however, our attempt to clarify the presentation requires us, in several places, to use modern mathematical notation in preference to English prose; but on the whole we leave to others the task of a rigorous symbolic analysis, a nice example of which is to be found in Professor Terence Fine's 'Theories of Probability[2]'.

However, having claimed that Bayes' writing is in the 'plain words' tradition, it may seem strange also to claim that he has often been misunderstood and that, in several places, significant effort is needed to tease the meaning from his writing. It is tempting to say that the reasons are largely due to the antiquity of the document. For many people of our own time, this must be indeed some part of the truth: but it does not explain why the essay should have stymied Bayes' contemporaries nor have led to the extensive mis-interpretation and mis-representation which we believe can be found in the writings of Bayes' critics. To answer that question in any depth is a large task, and the conclusions would be deeply uncertain: a plausible guess is that the subject matter was, and is, difficult: but that would not explain how persons of significant intellect could have failed to understand Bayes' argument. In some cases, critics may have simply taken the word of others as regards the content and import of the essay, but that does not explain the apparent failures of authors who quote *verbatim* from the essay, and would appear to have had the original text to hand, and yet seem to have been rendered oblivious to what the essay actually says, by their own *a priori* views of its contents[3].

Our purpose in this chapter therefore, is to try and make clear, in the plainest possible language for readers of our own time, the points made in the essay up to page 394 of the original text. To this end, we have changed some of the letters used by Bayes as mathematical symbols, in order to avoid some confusing ambiguities. We are also, in places, at pains to make explicit some very basic algebraic manipulations with the aim of allowing the reader to follow the line of reasoning as smoothly as possible. This seems necessary because Bayes uses cunning substitutions which often turn out to be quite simple, once one has searched out the basis, but which might prove vexing

---

[1] cf de Moivre *'if not to force the assent of others by a strict demonstration, (then) at least to the satisfaction of the Enquirer'* de Moivre, (1756, p254)

[2] Fine (1973).

[3] See *e.g.* Fisher (1956), also discussed in Appendix A below.

to readers who do not have time to unravel the connections for themselves[1] and whose main concern is to follow the argument to its conclusion, feeling that each step has been made clear *en route.*

The Definitions

The problem having been presented at the head of the essay, our analysis begins with the Definitions, the first four of which are fairly simple, but may be somewhat easier to grasp with slight paraphrasing of the original:-

(1) Several events are ***inconsistent*** *when, if one of them happens, none of the rest can.*

Although this definition of *inconsistent* may be technically acceptable, it represents a somewhat quaint usage in modern parlance, where we might more commonly expect to find expression such as *incompatible* or *mutually exclusive*.

(2) Two events are ***contrary*** *when, in a given trial, one or other of them must happen but the happening of one excludes the happening of the other.*

(3) An event is said to ***fail*** *when the situation changes from one in which the event could possibly happen to one in which it cannot happen. Such a change occurs when the contrary of an event happens.*

This definition of *fail* may also be thought slightly quaint, in that it could include an occasion on which a trial has to be abandoned, *e.g.* because of bad weather. However, such situations do not affect Bayes' argument and we therefore let his definition stand.

(4) An event is said to be ***determined*** *when it has either happened or failed.*

The fifth definition is not however straightforward:-

(5) The probability of any event *is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening.*

To understand this definition we have to look ahead in the essay to the manner in which it is used and it seems helpful to introduce a more explicit notation such that the value of an object, prize or expectation is denoted in monetary units *e.g. $X*, where *X* is a positive, real number. Thus we find, in Proposition 2, for example, that if we denote the probability of the happen-

[1] *Pace* Dr Hiya Freedman who, on being shown a copy of the essay, commented, *'This is not mathematics, it is philosophy. You can express the whole thing in a few lines of algebra. I do not have time to read philosophy'.* (In conversation, November 1996).

ing of an event *E* as $\mathcal{P}(E)$ and our situation is such that, if the event happens we shall acquire something of value *$N*, then the value at which our expectation ought to be computed is given by the expression:-

$$\$V = \$N \times \mathcal{P}(E) \qquad (3\text{-}1)$$

and algebraically therefore, we can write:-

$$\mathcal{P}(E) = \frac{\$V}{\$N} \qquad (3\text{-}2)$$

The trouble with this definition is that, at first sight, it bears no obvious relation to the everyday ways in which we use the word 'probability' and a full discussion of the matter would take us into deep and contentious waters. It can however be resolved quite simply for the present purpose by treating Bayes' definition not as a definition of 'probability in itself' but as a specification of, or rule for the way in which a known probability ought to be used by a rational person in a particular kind of situation. Or, as Price says in his covering letter, *'Instead therefore, of the proper sense of the word probability, he has given that which all will allow to be its proper measure in every case where the word is used [1]'.*

Perhaps the simplest way to envisage the type of situation in which a known probability can be used, as defined by Bayes, is in terms of betting money on the outcome of a game. If we are admitted to a game in which we have a known chance $\mathcal{P}(E)$ of winning a prize worth *$N,* we can certainly say that, if viewed purely from a financial point of view, taken in isolation, and excluding ulterior motives, the amount $\mathcal{P}(E) \times \$N$ is the rational amount which it would be worth paying to gain admission to such a game. As we feel that several points in the argument can be presented most clearly in betting terms, we shall do so where it seems necessary, and trust that none will take offence at this slight liberty[2].

***(6)** By **chance** I mean the same as probability.*

In Bayes' time, mathematical notation was quite primitive compared with today and this added to the need for clarity in verbal reasoning about mathematical matters. To many people, recursive expressions such as 'the

---

[1] Bayes (1763) p375

[2] For people such as Bayes, de Moivre and Price, it was important to avoid the moral stigma of gambling which, at that time, was *'a national disease among the leisured classes of both sexes'.* See Turberville (1926) p86.

probability that the probability .....'. are hard to grasp, but we seem to have less difficulty with an expression such as 'the chance that the probability .....'., even when we know that 'chance' and 'probability' have the same meaning, and this is how Bayes uses these words.

(7) *Events are* ***independent*** *when the happening of any one of them does neither increase nor diminish the probability of any of the rest.*

In itself this definition should give no difficulty, but it may help with Proposition 3, later, if we give at this point an example of how the happening of one event may affect the probability of another, *e.g.* if we are sampling the bird population of an island, the size of a particular bird in our sample may affect the probability that the colour of the beak of the next bird to be observed will be yellow, simply because birds tend often to associate in flocks of similar kinds, and it may be that we are surrounded at some point by a flock of small birds with yellow beaks[1].

Proposition 1

If individual events within a given set of possible events are inconsistent, (i.e. mutually exclusive), the probability that one or other of them will happen is the sum of the probability of each.

Bayes supports this proposition by applying his definition of probability to an implicit gaming situation, *e.g.* a lottery, in which there is a single prize, worth *$N*, which we shall win if any one of three mutually exclusive events should happen in a given trial, *e.g.* we are given three tickets for a lottery and we denote by E_1 the event that the first ticket wins. We likewise define events E_2 and E_3 for the second and third tickets respectively. We then let the probability of winning with the first ticket be $\mathcal{P}(E_1)$, and of winning with the second and third tickets, $\mathcal{P}(E_2)$ and $\mathcal{P}(E_3)$ respectively. Then, by appropriate choice of numbers which we denote by p_1, p_2, p_3, these probabilities can be expressed as $\mathcal{P}(E_1) = p_1 / N$ *etc.*, such that the value of our expectation, $\$V_1$, depending on the happening of the first event (*i.e.* if the first ticket wins) is:-

$$\$V_1 = \$N \times \mathcal{P}(E_1) = \$N \frac{p_1}{N} = \$p_1 \qquad (3\text{-}3)$$

and likewise $\$V_2 = \$p_2$, and $\$V_3 = \$p_3$. Bayes then states that the value of our expectation in the situation where we stand to win the *$N* if any of these events should happen, is the sum of the individual expectations *i.e.*

[1] Bayes' definition is limited to pairwise independence. For a fuller discussion, see Lindley (1965), vol. I, pp14 - 16.

$\$V_S = \$(p_1+p_2+p_3)$. But, from the way in which Bayes has defined his terms, the probability that one of E_1 or E_2 or E_3 will happen can be computed by taking the ratio of the value of the expectation to the value of the prize, thus giving:-

$$\mathcal{P}(E_1 \text{ or } E_2 \text{ or } E_3) = \$(p_1+p_2+p_3) / \$N$$

$$= p_1/N + p_2/N + p_3/N \qquad (3\text{-}4)$$

which is the sum of the probabilities of each of the events, as stated in the proposition[1].

Corollary to Proposition 1

If it is certain that one or other of the three events must happen, then we are bound to win $N whatever the outcome of the trial. Therefore, the sum of the alternative expectations must be:-

$$\$(V_1+V_2+V_3) = \$N \qquad (3\text{-}5)$$

hence, since we have shown above that $\$V_1 = \$p_1$ etc., it follows that:-

$$p_1 + p_2 + p_3 = N \qquad (3\text{-}6)$$

and it further follows that the probability of an event added to the probability of its failure (or of its contrary) is the ratio of equality (i.e. unity).

Bayes' proof of this latter assertion is by an argument that, because the outcome of a trial must be either the happening of a given event or its contrary, one of these outcomes is bound to happen. Hence, if an event E_1 and its contrary $\sim E_1$ are regarded simply as a pair of inconsistent events E_1 and E_2, of which one or the other must happen, then:-

$$p_1 + p_2 = N \qquad (3\text{-}7)$$

whence

$$p_2 = N - p_1 \qquad (3\text{-}8)$$

and since

$$\mathcal{P}(E_2) = \mathcal{P}(\sim E_1) = p_2 / N \qquad (3\text{-}9)$$

then by (3-8):-

$$\mathcal{P}(\sim E_1) = \frac{N - p_1}{N} \qquad (3\text{-}10)$$

Proposition 2

[1] This proposition is discussed more critically in Chapter 6.

With Proposition 2 we encounter a more weighty style of argument and a result which plays an important part in much that follows:-

If we have an expectation depending on the happening of an event, then the probability that the event will happen $\mathcal{P}(E)$, divided by the probability of its failure $\mathcal{P}(\sim E)$, is equal to the ratio of our loss if it fails to our gain if it happens.

The justification is by an argument which uses the extremely subtle idea of computing the value of a situation:-

Suppose we have an expectation of receiving \$N if an event E happens, and that the probability of its happening, $\mathcal{P}(E)$, can be equated to a ratio p/N. Then, by definition, the value of the expectation prior to the trial is

$$\$V = \$N \times (p/N) = \$p \qquad (3\text{-}11)$$

Therefore, if the event fails to happen, this is equivalent to the loss of an expectation of value \$p. However, if the event does happen, we gain \$N but lose the expectation valued at \$p, hence the net improvement in our situation is equivalent to a gain of only \$(N - p).

Therefore the ratio of our loss if the event fails to occur, to our gain if it happens, is p/(N - p). The proposition is proved by the relationships we have established viz:-

$$\frac{p}{(N-p)} = \frac{p/N}{(N-p)/N} = \frac{\mathcal{P}(E)}{\mathcal{P}(\sim E)} = \frac{Loss}{Gain} \qquad (3\text{-}12)$$

Proposition 3

Here, we paraphrase rather heavily to try and make the meaning clear:-

The probability that two consecutive events will both happen is equal to the probability that the first event will happen, multiplied by the probability that the second event will happen if it is known that the first has happened.

A simple and extreme case which illustrates the probability of one event being conditioned by the occurrence of a different event, is one in which the two events are incompatible, *i.e.* if one happens the other cannot happen. Less extreme cases are often found in the biological sciences, where, for instance, particular physical characteristics, such as colour of eyes and colour of hair may be strongly correlated in the population of a certain

city[1]. To justify Proposition 3, Bayes again uses expectations and requires us to suppose that we are already admitted to a two-part trial in which:-

If both events happen we are to receive $N (3-13)

The probability that both events will happen, which we denote as $\mathcal{P}(E_1 \wedge E_2)$, *is proportional to a number B which is selected so that:-*

$$\mathcal{P}(E_1 \wedge E_2) = B/N \quad (3\text{-}14)$$

The probability that the first event will happen is denoted:-

$$\mathcal{P}(E_1) = p_1/N \quad (3\text{-}15)$$

and hence, by the corollary to Proposition 1, the probability that the first event will not happen is:-

$$\mathcal{P}(\sim E_1) = (N - p_1)/N \quad (3\text{-}16)$$

The probability that the second event will happen, on the assumption that the first event has happened is:-

$$\mathcal{P}(E_2 \mid E_1) = r_2/N \quad (3\text{-}17)$$

where the symbol r_2 *is used rather than* p_2 *to emphasise that the probability is relative to* E_1.

The justification proceeds along the following lines:-

Prior to the trial, the value of our expectation is:-

$$\$V = \$N \times (B/N) = \$B \quad (3\text{-}18)$$

If in the first part of the trial, the first event happens, the expectation changes and becomes:-

$$\$V = \$N \times (r_2/N) = \$r_2 \quad (3\text{-}19)$$

Hence the improvement in our position if the first event happens, is $\$(r_2 - B)$.

If, however, in the first part of the trial the first event fails to happen, the loss is $B and we obtain the relationship:-

$$\frac{Loss}{Gain} = \frac{B}{(r_2 - B)} \quad (3\text{-}20)$$

[1] cf Jeffreys, H. (1939, p27)

whence, invoking the result demonstrated for Proposition 2, i.e. that the $\frac{loss}{gain}$ *ratio stemming from the completion of the first part of the trial is equal to the ratio* $\mathcal{P}(E_1) / \mathcal{P}(\sim E_1)$, *gives:-*

$$\frac{Loss}{Gain} = \frac{B}{(r_2 - B)} = \frac{\mathcal{P}(E_1)}{\mathcal{P}(\sim E_1)}$$

$$= \frac{p_1/N}{(N - p_1)/N} = \frac{p_1}{(N - p_1)} \qquad (3\text{-}21)$$

Cross-multiplying the second term in (3-21) and the last term then gives:-

$$N \times B - p_1 \times B = p_1 \times r_2 - p_1 \times B$$

$$i.e. \quad N \times B = p_1 \times r_2$$

$$hence \quad \frac{B}{r_2} = \frac{p_1}{N} = \mathcal{P}(E_1) \qquad (3\text{-}22)$$

Then, since B/N can be expressed as the product of two ratios:-

$$\frac{B}{N} = \frac{B}{r_2} \times \frac{r_2}{N} \qquad (3\text{-}23)$$

and since the components of this product are, first, by (3-22):-

$$\frac{B}{r_2} = \mathcal{P}(E_1) \qquad (3\text{-}24)$$

and, second, by definition:-

$$\frac{r_2}{N} = \mathcal{P}(E_2 | E_1) \qquad (3\text{-}25)$$

we have

$$B/N = \mathcal{P}(E_1) \times \mathcal{P}(E_2 | E_1) \qquad (3\text{-}26)$$

$$= (p_1/N) \times (r_2/N) \qquad (3\text{-}27)$$

That is, the probability that both events will happen is the product of the probability that the first event will happen, and the probability that the second will happen on the assumption that the first has already happened.

Corollary to Proposition 3

If, of two subsequent events, the probability of the first is p_1/N *and the probability of both together is B/N, then the probability of the second event on the supposition that the first event happens, is* B/p_1

This result is achieved by taking the expression *(3-23)*

$$\frac{B}{N} = \frac{B \times r_2}{r_2 \times N}$$

and substituting p_1/N for B/r_2, as derived in *(3-22)* so that:-

$$\frac{B}{N} = \frac{p_1}{N} \times \frac{r_2}{N} \qquad (3\text{-}28)$$

whence

$$r_2/N = \frac{B/N}{(p_1/N)} \qquad (3\text{-}29)$$

$$= B/p_1 \qquad (3\text{-}30)$$

Proposition 4

This proposition and its justification do not make for easy understanding. The proposition says, essentially:-

If a two-part trial is to be conducted every day,

and the probability of the second event happening in a given trial is p_2/N,

and the probability of both events happening in the same trial is B/N,

and we are to receive $N if both events happen on the first day on which the second event happens,

then the probability that we will win the $N is B/p_2.

An immediate difficulty with this proposition is that where it says *'the probability of the second event',* we might expect to see, following from Proposition 3, *'the probability of the second event when it is known that the first has happened'*. The qualifying words are however absent, and although it may be hard to grasp the meaning if the probability of the second event is read as being 'un-conditional', we find as we unfold the argument that this is indeed the meaning which Bayes intends. We therefore here use the symbol p_2 rather than r_2 to denote this probability. Bayes then proceeds to justify the proposition by, first, supposing that the probability of winning the *$N* is equal to an unknown ratio *x/N,* so that the value of our expectation on being admitted to the game, and before any trial, is *$N(x / N) = $x.* But also, by the definition of the problem, the probability that we will win on the first day is *B/N* and therefore the value of the expectation that we will win on the first day, *i.e.* by both events happening, is *$N(B/N) = $B.* But, if we do not win on the first day, this can be due to the occurrence of any one of three possible, paired outcomes:-

$$[\ \sim E_1 \wedge \sim E_2\ ]\ ,\ or\ [\ E_1 \wedge \sim E_2\ ],\ or\ \ [\ \sim E_1 \wedge E_2\ ]\ ,$$

but we will only have lost the expectation in the third case. In the two former cases where the second event has not occurred, we retain our place in the game, and we are effectively once again in the starting situation: therefore the expectation retains its initial value of *$x*.

Hence we are concerned to value our expectation when we have not won and when we have also not necessarily lost; that is, we are concerned with the probability of occurrence of either of the two cases $[\ \sim E_1 \wedge \sim E_2\ ]$ and $[\ E_1 \wedge \sim E_2\ ]$ under which we retain our place in the game, *i.e.* we are concerned with the probability that E_2 has not occurred, irrespective of the occurrence or non-occurrence of E_1. Hence, because $\mathcal{P}(\ E_2\ ) = p_2/N,$ it follows, by the corollary to Proposition 1 that:-

$$\mathcal{P}(\sim E_2\ ) \quad = \quad (N - p_2)\ /\ N \qquad (3\text{-}31)$$

Bayes now introduces an unknown number *'y'* such that

$$y/x \quad = \quad (N - p_2)/N \qquad (3\text{-}32)$$

and therefore

$$y/x \quad = \quad \mathcal{P}(\sim E_2\ ) \qquad (3\text{-}33)$$

Then, since *$x* is the value of the position that we will retain if E_2 has not happened, and the probability of this is the probability that E_2 has not happened, the effective value of our position when we know only that the outcome of the trial has been $\sim(E_1 \wedge E_2\ ),$ is:-

$$\$x \times \mathcal{P}(\sim E_2\ ) \quad = \quad \$x \times\ (N - p_2\ )/N \qquad (3\text{-}34)$$

and since

$$(N - p_2)/N \quad = \quad y/x \qquad (3\text{-}35)$$

$$\$x \times \mathcal{P}(\sim E_2\ ) \quad = \quad \$x \times\ (y/x) \qquad (3\text{-}36)$$

$$= \quad \$y \qquad (3\text{-}37)$$

Bayes then writes:-

$$\$x \times \mathcal{P}(\sim E_2\ ) \quad = \quad \$x \times\ (N - p_2\ )/N \qquad (3\text{-}38)$$

and since also

$$(N - p_2)/N \quad = \quad y/x \qquad (3\text{-}39)$$

$$\$x \times \mathcal{P}(\sim E_2\ ) \quad = \quad \$x \times\ (y/x) \qquad (3\text{-}40)$$

$$= \quad \$y \qquad (3\text{-}41)$$

But these two last expectations together are evidently the same with my original expectation, the value of which is $x, and therefore:-

$$\$B + \$y = \$x \qquad (3\text{-}42)$$

It is not easy to see immediately what Bayes means at this point, but since the assertion concludes with the equation *$B+$y = $x,* it seems fairly certain he means that, in a situation where both parts of the trial have occurred, the possibilities can be expressed as a pair of simple alternatives: either we have won or we have not won. If we have not won, this does not totally amount, in this type of trial, to having lost. Having 'not won', our position in the game is still worth the value of the expectation that E_2 has not occurred. Bayes seems therefore to imply that since the value of our position on being allowed into the game is *$x,* then the outcome, after the trial has taken place but before the results are known, can be divided into the two possibilities of having won or of having not won. Hence, the *$x* can be divided into the two parts, *$B* and *$y* corresponding with the values of the expectations of having won or of having not won respectively. We therefore have the two equations:-

$$B + y = x \qquad (3\text{-}43)$$

i.e.

$$y = x - B \qquad (3\text{-}44)$$

and, by definition

$$y/x = (N - p_2) / N \qquad (3\text{-}45)$$

from which, by substituting $x - B$ for y in *(3-41)* we have:-

$$\frac{x - B}{x} = \frac{N - p_2}{N} \qquad (3\text{-}46)$$

hence $$\frac{B}{x} = \frac{p_2}{N} \qquad (3\text{-}47)$$

hence $$\frac{x}{N} = \frac{B}{p_2} \qquad (3\text{-}48)$$

and therefore, as x/N is defined as being the unknown probability of winning an amount *$N,* the proposition that this probability is equal to B/p_2 is demonstrated.

The Corollary to Proposition 4

Continuing with the type of trial postulated in Proposition 4, let us suppose that, before we know whether the first event has happened, we find

that the second event has happened; then purely on the basis of this information, we can infer that the first part of the trial has taken place but we do not know its outcome, and therefore we have no reason to value the expectation either greater or less than it was before[1].

The main difficulty with this corollary is the manner in which it is presented, particularly in the concluding phrase '*We therefore have no reason to value the expectation either greater or less than it was before',* for it must be a reader of superlative intelligence indeed who can see at a glance that the conclusion follows as a simple 'therefore' from what has gone before. Indeed, to support this deduction, we need to assume that the value of the expectation depends, not upon the true situation, but rather upon what we know, or believe, about the situation, *i.e.* the expectation has to be valued in relation to the information available to us[2]. It is therefore better, we feel, to pose the question as to whether there are good reasons to value the expectation as more than, or less than, or the same as it was before. Indeed, a question of this form is implicit in the manner in which Bayes continues:-

For if we had reason to think it less, it would be reasonable for us to give something to be reinstated in our former circumstances, and this over and over again as often as we should be informed that the second event had happened, which is evidently absurd.

To perceive this absurdity it helps if we make explicit a further assumption which Bayes seems to take for granted at this point, namely that each two-part trial is independent of all others. (Bayes does, however, deal with this question of independence in a series of identical trials, in some detail later, as we find when we consider the Definition which follows Proposition 6). In the present case, however, when event E_2 has happened, we must have either won or lost, but we do no know which; for the case of having 'not-won' but being still in the game is excluded by the happening of E_2.

Bayes now argues that if the value of our expectation were reduced by receipt of this information about E_2, it would be financially sensible to pay some amount, say *$d,* to restore our expectation to its original value, *e.g.* by terminating the game and starting again. But since our expectation would be valued still at only *$V,* it is immediately obvious that such a move would be financially irrational, especially if we had already paid the rational limit of *$V* to gain entry to the game. But even if we had paid nothing to gain

[1] The temporal order of the trials is not however of fundamental importance.

[2] This point is of fundamental importance; see Ch 11 below.

entry, yet following each trial in which we were told that event E_2 has happened, we were to pay *$d* to restore our expectation, then clearly there could come a point at which we would have paid more than the value of the prize itself. This is indeed absurd, and there would be no limit to the amount that we might lose if we were to continue in such a manner.

Conversely, the like absurdity plainly follows if you say we ought to set a greater value on our expectation than before, for then it would be reasonable to refuse something if offered upon condition that we would relinquish the improvement, and be reinstated in our former circumstances; and this likewise over and over again as often as (nothing being known concerning the first event) it should appear that the second event had happened.

That is, there might be no limit to the amount that we might refuse, for the sake of retaining a chance of winning a prize which has a strictly fixed value of *$N*. Therefore, because the expectation can be neither greater nor less than it was before we knew that E_2 had happened, it must have retained exactly its previous value. The argument then concludes, effectively thus:-

Hence if we know only that the second event has happened, the value of our expectation is:-

$$\$V = \frac{N \times x}{N} \qquad (3\text{-}49)$$

and hence by (3-44)

$$= \frac{\$N \times B}{p_2} \qquad (3\text{-}50)$$

whence by Definition 5, the probability of winning is still B/p_2 and, by the above argument, this is the probability that the first event has happened when the second event is known to have happened.

Suddenly, therefore, the nature of the probability being evaluated has undergone a fundamental change: for we are now abruptly concerned with the probability of an hypothesis concerning the occurrence of an event. The next step is the crux of the essay:-

> *But the probability that an event has happened is the same as the probability that I am right if I guess that it has happened. Wherefore the following proposition is evident:-*

Proposition 5

If there are two consecutive events, the probability of the second event being p_2/N, and the probability that both events will occur together being B/N, if it is discovered that the second event has happened, and I then guess that the first event has also happened, the probability that I am right is B/p_2.

That is:-

$$\mathcal{P}(E_1 | E_2 , B/N, p_2/N) \quad = \quad B/p_2 \qquad (3\text{-}51)$$

Or, in a more condensed form, using k to denote the information conveyed by the terms $B/N, p_2/N$, this result can also be expressed as:-

$$\mathcal{P}(E_1 | E_2, k) \quad = \quad \frac{\mathcal{P}\,((E_1 | k) \wedge (E_2 | k))}{\mathcal{P}\,(E_2 | k)} \qquad (3\text{-}51a)$$

or, expanding $\mathcal{P}\,(E_1 \wedge E_2)$ by Proposition 3:-

$$\mathcal{P}(E_1 | k, E_2) \quad = \quad \frac{\mathcal{P}(E_1 | k) \times \mathcal{P}(E_2 | k, E_1)}{\mathcal{P}\,(E_2 | k)} \qquad (3\text{-}51b)$$

Further, as we shall see in later chapters, there are numerous applications in which it is helpful to expand the denominator in *(3-51b)* to cover explicitly certain kinds of trial in which there are marked correlations between the happening or non-happening of the events. Thus, by Propositions 1 and 2:-

$$\mathcal{P}(E_2 | k) = \mathcal{P}(E_2 | k, \sim E_1).\mathcal{P}(\sim E_1) + \mathcal{P}(E_2 | k, E_1).\mathcal{P}(E_1) \qquad (3\text{-}51c)$$

or, if E_1 comprises the happening of one of a set of mutually-exclusive 'sub-events' $E_{1a}, E_{1b}, \ldots\ldots E_{1n}$ to each of which there attaches a probability $\mathcal{P}(E_{1a} | k) \ldots\ldots \mathcal{P}(E_{1n} | k)$ then:-

$$\mathcal{P}(E_2 | k) = \mathcal{P}(E_2 | k, E_{1a}).\mathcal{P}(E_{1a} | k) + \mathcal{P}(E_2 | k, E_{1b}).\mathcal{P}(E_{1b} | k) + \ldots.$$

$$\ldots\ldots + \mathcal{P}(E_2 | k, E_{1n}).\mathcal{P}(E_{1n})$$

which we abbreviate as:-

$$\mathcal{P}(E_2 | k) = \sum_i \mathcal{P}(E_2 | k, E_{1i}).\mathcal{P}(E_{1i} | k) \qquad (3\text{-}51d)$$

or, in the case where E_1 is continuously variable:-

$$\mathcal{P}(E_2 | k) = \int \mathcal{P}(E_2 | k, E_1).\mathcal{P}(E_1 | k) dE_1 \qquad (3\text{-}51e)$$

from which we see that the integrand in the denominator of *(3-51e)* is identical to the numerator and therefore, the probability[1] at a postulated value of $E_1 \approx e$ is given by:-

$$\mathcal{P}(E_1 \approx e \mid k, E_2)$$

$$= \frac{\mathcal{P}(E_1 \mid k) \times \mathcal{P}(E_2 \mid k, E_1)}{\mathcal{P}(E_2 \mid k)}$$

$$= \frac{\mathcal{P}(E_1 \mid k) \times \mathcal{P}(E_2 \mid k, E_1)}{\int_{-\infty}^{\infty} \mathcal{P}(E_2 \mid k_0, E_1 = e) \cdot \mathcal{P}(E_1 \approx e \mid k) de} \qquad (3\text{-}51f)$$

Proposition 6

As Bayes moves to Proposition 6, his style changes and the argument based on expectations is left behind. The proposition is:-

The probability that several independent events shall all happen is a ratio compounded of the probabilities of each.

The demonstration takes the following form:-

It is in the nature of independent events that the probability of one such event happening is not altered by the happening or failing of any other such event. But the probability that a second event E_2 will happen, supposing that some prior event E_1 is known to have happened, is the probability that both events will happen, i.e.:-

$$\mathcal{P}(E_1 \wedge E_2) = B/N = \mathcal{P}(E_1) \times P(E_2 \mid E_1)$$

$$= (p_1 / N) \times (r_2 / N) \qquad (3\text{-}52)$$

But because E_1 and E_2 are independent, we know that $r_2 = p_2$ (cf 3-17 above), whence

$$\mathcal{P}(E_1 \wedge E_2) = (p_1/N) \times (p_2/N) \qquad (3\text{-}53)$$

We then consider E_1 and E_2 taken together as a single event, and introduce further independent events $E_3 \ldots E_n$ whence:-

$$\mathcal{P}(E_1 \wedge E_2 \wedge E_3 \wedge \ldots E_n)$$

[1] For an explanation of our approach to the probability at a point on a continuous distribution, see the preliminary section on *Notation*.

$$= \mathcal{P}(E_1) \times \mathcal{P}(E_2) \times P(E_3) \times \ldots \mathcal{P}(E_n) \qquad (3\text{-}54)$$

The first corollary to Proposition 6

If there are several independent events, E_1. E_2 ..... E_n then the probability that E_1 happens and E_2 does not happen, etc., is:-

$$\mathcal{P}(E_1 \wedge \sim E_2 \wedge \ldots\ldots \sim E_n)$$

$$= \mathcal{P}(E_1) \times \mathcal{P}(\sim E_2) \times \mathcal{P}(E_3) \times \ldots \mathcal{P}(\sim E_n) \qquad (3\text{-}55)$$

The second corollary to Proposition 6

This corollary is then merely a re-expression of the first in slightly different notation, where $\mathcal{P}(E_1) = a$, $\mathcal{P}(E_2) = b$, *etc.,* such that

$$\mathcal{P}(E_1 \wedge \sim E_2 \wedge E_3 \wedge \ldots\ldots \sim E_n) = a \times (1\text{-}b) \times c\ldots\ldots \qquad (3\text{-}56)$$

A further definition

On page 383 of the original text there is a further definition, which is far from easy to grasp on a first encounter:-

If in consequence of certain data there arises a probability that a certain event should happen, its happening or failing, in consequence of these data, I call its happening or failing in the first trial. And if the same data be again repeated, the happening or failing of the event in consequence of them I call its happening or failing in the second trial; and so on as often as the same data are repeated. And hence it is manifest that the happening or failing of the same event in so many different trials, is in reality the happening or failing of so many distinct independent events exactly similar to each other.

Having studied this definition at some length, we conclude that although it appears in the original text to be typographically part of Proposition 6, it is logically a preamble to Proposition 7. Another problem is to determine just what is being defined, and since Bayes does not make this explicit, we conclude from reflection upon the final sentence:- *'the happening or failing of the same event in .... different trials, is in reality the happening or failing of .... independent events'*, that the purpose of the definition is to make clear what is meant when we talk about multiple but independent happenings of the 'same event'; for in common speech, the events which constitute a multiplicity of happenings are, in principle, individually enumerable and are distinguished by an attribute such as a unique ordinal number. Hence, in this context, the term *'same event'* must mean something special, and with this in

Version zh25a.doc at 18:36 on 13 July, 2001

mind, we can infer that Bayes' purpose was to achieve a definition on the following lines:-

When we talk about multiple happenings of the same event, we mean multiple occasions on which we get the same outcome from a given type of trial, where each trial is of an identical probablistic nature, being governed in every relevant aspect by identical specifications and conditions, so that each such trial is independent of, whilst being externally similar to, each other such trial.

Proposition 7

If the probability of the happening of an event 'E' in a single trial is 'p' and that of its failing is 'q' , then the probability of its happening 'm' times in 'n' trials is

$$\mathcal{P}(m : n) \quad = \quad C\, p^m q^{n-m} \qquad (3\text{-}57)$$

where C is the coefficient of $p^m.q^{n-m}$ in the binomial expansion of $(p + q)^n$.

In the demonstration of this proposition, Bayes begins by implicitly invoking the 'further definition' discussed a few lines earlier:

For the happenings or failings of an event in different trials are independent events. Hence, by the second corollary to Proposition 6, the probability of a set of 'n' outcomes forming a series such as 'E ∧~E ∧~E ∧E ∧~E ....'. is given by the product of the appropriate 'n' components of a series such as $p \times q \times q \times p \times q$ ..... $= p^m \times q^{n-m}$, in which the product has the same value for every arrangement having the same number of p's and q's.

And since the number of different arrangements of these p's and q's is the number of different series that can be formed by the outcomes of 'n' independent trials in which the event happens 'm' times and fails on 'n-m' occasions, and because the occurrence of any one such series on any given set of 'n' trials is incompatible with the occurrence of any other series on that same set of trials, the probability that, in one way or another, there will be 'm' happenings and 'n-m' failures is $Cp^m q^{n-m}$, where C is the number of different possible arrangements of the given number of p's and q's.

Here, we reach the end of the first section of the essay. Although the second section is written in a noticeably different style, it makes significant use of the results derived above and it is only by getting thoroughly to grips with the first Section that we can face honestly the questions which are posed in Section 2. Hence we cannot subscribe to the view that Section 1 is merely a recital of well-known formulae, and we shall see, as we examine Section 2 in detail, that we require all the definitions, propositions and corollaries of Section 1.

Chapter 4

The Experiment

It is in Section 2 of the essay, beginning on page 385 of the original text that we encounter the famous experiment which takes place, conceptually, upon a 'square table or plane'. In simple terms, the experiment takes the form of throwing a 'first ball' at random onto the table and then, it having come to rest on the table, we are required to estimate the probability that its distance from a defined side is between any two given values by randomly throwing a number of similar balls onto the table and being told how many of these come to rest between the first ball and the designated side of the table.

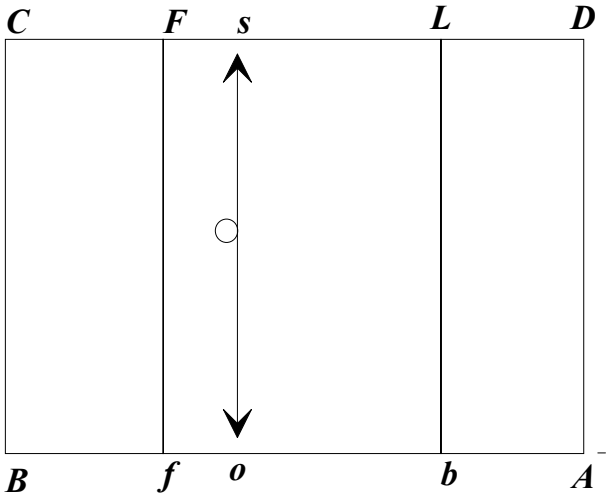


Fig 4.1

Unfortunately, in Bayes' description, there are some difficulties with the notation: Bayes writes of the experiment in terms of balls designated 'o' and 'W' in the text, or 'O' and 'W' in the printed diagram, but identical or very similar symbols are also used with quite different meanings in the discussion; *e.g.* the line through the point at which the ball W comes to rest is designated *os* and there is frequent use of the symbol *'o'* to denote what we would regard as the *x* co-ordinate of the position of the first ball. We have therefore changed the designations to 'first ball' and 'second ball', trusting that this change, and other minor changes to the notation, are of some help in

making clear the finer details of the argument. We must however mention here, as is emphasised later, that the order in which the balls are thrown is not vital to Bayes' argument. At this stage, however, it is easier to think of the experiment in such terms.

The description of the experiment begins with **Postulate 1,** this being an assertion that when a ball is thrown onto the table, the probability that it will come to rest on any one part of the table is equal to the probability that it will come to rest on any other equal part. We assume, although Bayes does not say so explicitly, that the equality is in the areas of the parts, and it later becomes important to assume also that these parts do not overlap. With **Postulate 2**, we suppose that we throw the first ball onto the table and draw a line *os* parallel to the side AD, through the point at which the ball comes to rest. Logically then, and putting aside for the moment Postulate 2, this brings us to:-

Lemma 1
The probability that the point o will fall between any two points on the line AB is equal to the ratio of the distance between the points to the length of the side AB.

Although this may seem evident to ourselves, as a direct consequence of the assertion of equal probabilities over equal areas as stated above, Bayes devotes some two pages to proof of this Lemma. The proof starts by marking points *f* and *b* on the side *AB* and through these points drawing lines *fF* and *bL* parallel to *AD* to intersect side *CD* at *F* and *L* respectively.

Bayes then argues:-

If the rectangles Cf, Fb, LA are commensurable to each other, they may each be divided into the same equal parts, and, the first ball being then thrown, the probability that it will come to rest somewhere upon a given subset of these parts will, because resting upon any given part is inconsistent with resting upon any other such part, be equal to the sum of the probabilities of its coming to rest upon each such part.

An immediate difficulty with this argument stems from the way in which the word *commensurable* is used, since it does not readily accord with its common usage in our own time. Fisher[1] states however, that the sense with which Bayes uses *commensurable* is to be found in Euclid Book V, and if we

[1] Fisher (1956, p 13). Also, *The Penguin Dictionary of Mathematics* [Daintith (1989)], defines *commensurable* uniquely as *'Describing two quantities which are integral multiples of a common unit'.*

look ahead in the essay to p387 of the original we find the words:- *Again; if the rectangles Cf, Fb, LA are not commensurable.* Having studied the arguments which follow, we conclude that, in this context, two rectangles are *commensurable*, in the sense used by Bayes, if each can be constructed, exactly and without remainder, by the non-overlapping conjunction of identical 'tiles', *i.e.* the tiles are *'the same equal parts'* but the number of tiles in any given rectangle is, proportional to the area of that rectangle. There is then a further implication that the probability of the ball coming to rest on any given tile is equal to the probability of it coming to rest on any other tile. We can therefore paraphrase the argument as follows:-
If the rectangles are commensurable, then each may be divided into an integral number of identical tiles, and if we select from these a specific subset, then, because the resting of the ball on any given tile in the subset is inconsistent with its resting on any other such tile, the probability that the ball will come to rest on a tile which belongs to the subset, is the sum of the probabilities of its coming to rest upon each such tile. Hence, because the probability of its coming to rest on each such tile is equal to the probability of its coming to rest on any other such tile, the sum of the probabilities is equal to the probability of its resting on one such tile, multiplied by the number of tiles in the subset.

There then follows a string of arguments which may be confusing on a first reading, although if they are taken slowly, they are found to be very simple, particularly if we denote the three rectangles *Cf, Fb, LA* by the symbols R_1, R_2 , R_3 respectively, and take it for granted that the number of 'equal parts' within a rectangle is the area of the rectangle:-

Hence the probability that the ball will rest on R_1 is proportional to the area of R_1, and because it must come to rest somewhere on the plane, the probability that it will not rest upon R_1 is proportional to the remaining area R_2+R_3 . Hence the ratio of the probability that it rests on R_1 to the probability that it does not rest on R_1 is equal to the ratio of the area of R_1 to the sum of the areas R_2+R_3 : and as these areas are proportional to the lengths of the sides fb, Bf and bA respectively, the ratio of the probability that the ball will rest on R_1 to the probability that it will not do so, is equal to the ratio of the length fb to the sum of the lengths Bf + bA.

The demonstration of the Lemma then continues in a vein which may seem very laboured to ourselves, but may have been necessary in Bayes' day when serious doubts were still current as to the validity of various algebraic techniques which we take for granted. The continuation is however based upon the simple fact, demonstrated in the Corollary to Proposition 1, that the

probability of an event added to the probability of its contrary is the ratio of equality. Unfortunately, the argument in the original text is expressed as algebraic operations performed on English phrases, rather than on the more compact symbolic forms which are commonly used in our own times. To overcome this problem, we therefore use $\mathcal{P}(R_1)$ to symbolise the probability that the ball will rest on R_1 and $\mathcal{P}(\sim R_1)$ to symbolise its contrary. Bayes' argument can then be expressed as:-

$$\frac{\mathcal{P}(R_1)}{\mathcal{P}(\sim R_1)} = \frac{fb}{(Bf + Ab)} \qquad (4\text{-}1)$$

hence

$$\frac{\mathcal{P}(R_1)}{(\mathcal{P}(R_1)+\mathcal{P}(\sim R_1))} = \frac{fb}{(fb + Bf + Ab)} = \frac{fb}{AB} \qquad (4\text{-}2)$$

and therefore the probability that the point o will lie between f and b is the ratio of the length of fb to the side AB.

In terms of the original text, we are now on page 387 and, perhaps to the reader's concern, Bayes continues:-

Again; if the rectangles ..... are not commensurable ..... the last mentioned probability can be neither greater nor less than the ratio fb/AB,

i.e. the probability must therefore be equal to *fb/AB*, which Bayes proceeds to prove by examining the consequences if it were not so:-

We first consider the implications if the probability that the point o will lie between f and b were less than the ratio fb/AB, and were equal to some other ratio, say fc/AB.

To examine the implications we take the segment of line *fb* and on it we select points *p* and *t* such that the length *pt* is greater than the length *fc.* We then construct the three commensurable segments *Bp, pt* and *tA,* which, Bayes asserts:- *can always be done by dividing AB into a number of equal parts, each being less than cb/2 in length, and then taking p and t as the division-points nearest to f and c respectively.*

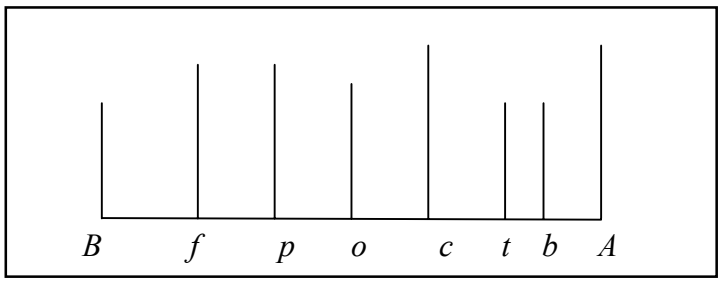


Figure 4.2

The commensurable segments are, therefore, such that each can be constructed exactly by conjunction of an integral number of units of the same finite length, and Bayes' argument is that such lines can always be found by choosing a unit which is less than *cb/2*, and such that the line *AB* is exactly divisible into an integral number of such segments. We are told that the division (*i.e.* end-of-segment) which is nearest to *f* (and is within the segment *fb*), is then chosen as the point *p*, and the division which is nearest to the point *c*, (and is again within the segment *fb*), is chosen as the point *t*. We note that the points *p* and *t* are defined as being upon the line *fb*; hence the first division that occurs to the right of *f* becomes the point *p*, and the first division that occurs to the left of *b* becomes the point *t.* Otherwise the division that is nearest to *f* may be to the left of *f,* and therefore outside the line *fb*. Also, if the segment length is λ, the point *p* could be this distance to the right of *f ;* but if we were then to make the point *t* the nearest division to *c*, this could be to the left of *c*, whence the length of the line *pt* would be less than the length *fc*, which contradicts its definition. The point *t* must therefore be to the right of *c* ; but, for us to be sure that *pt* is greater than *fc*, we must be sure that *t* is further to the right of *c* than *p* is to the right of *f*. We must therefore be sure that the length *ct* is greater than λ, which is not achieved by selecting the division next to the right of *c*, but is achieved, (because $\lambda < cb/2$), if the division next to the left of *b* is selected as the point *t*. This minor correction aside, Bayes' argument then continues:-

Because the line-segments Bp, pt and tA are commensurable, so are the rectangles based upon these segments. Hence from the result shown above, the probability that the point o will fall between p and t is equal to the ratio pt/AB. But if the point o lies between p and t, it must lie between f and b. Hence the probability that it will be between f and b cannot be less than pt/AB, and, since pt is greater than fc, this probability must also be greater

than fc/AB. Hence, the probability must be greater than any ratio which is less than fb/AB. And, since one can prove in the same way that that this same probability must be less than any ratio which is greater than fb/AB, it follows that it must be equal to fb/AB.

Lemma 2 to Postulate 2 begins at the foot of page 387 in the original text, but we need briefly to revisit Postulate 2 for definition of 'the event *M*':-

the event that the second ball comes to rest in the rectangle ADso, between the line os and side AD in a single trial is called the happening of the event M.

Lemma 2 then states that:-

the first ball having been thrown, the probability that the event M will occur in a single trial is equal to the ratio of the length Ao to the length of the whole side AB,

and this is demonstrated by the argument that:-

in the same manner as in the previous Lemma, the probability that the second ball will come to rest upon the rectangle ADso is equal to the ratio of Ao to AB, and, since this is defined as the happening of the event M in a single trial, the lemma is manifestly correct.

With ***Proposition 8***, the style of layout in the original text reverts to that of Section 1, the proposition beginning with a preamble describing how we construct the curve *BghikmA*, (see Fig 4.3), below the base of the plane viz:-

First we mark on the base AB a number of points such as b, c, d, e, f, from each of which we draw perpendiculars downwards to points m, k, i, h, g, respectively. We then define variables y_b, p_b, q_b, etc., with respect to each point we have marked on AB such that:-

$$y_b = \frac{bm}{AB} \qquad y_c = \frac{ck}{AB} \qquad \textit{etc.}$$

$$p_b = \frac{Ab}{AB} \qquad p_c = \frac{Ac}{AB} \qquad \textit{etc.}$$

$$q_b = \frac{Bb}{AB} \qquad q_c = \frac{Bc}{AB} \qquad \textit{etc.} \qquad (4\text{-}3)$$

and also such that each member of each (y,p,q) triplet obeys the relationship:-

$$y_b = {}^nC_m \, p_b^{\,m} \, q_b^{(n-m)} \qquad (4\text{-}4)$$

and

$$y_c = {}^nC_m \, p_c^{\,m} \, q_c^{(n-m)} \quad \ldots\ldots\ldots\ldots \textit{etc.} \qquad (4\text{-}5)$$

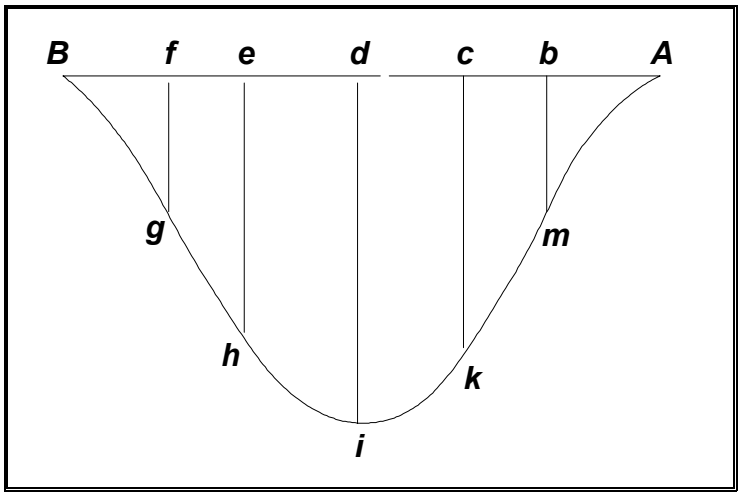


Figure 4.3

The body of the proposition then continues effectively as follows:-

Before the first ball is thrown, the probability that:-
(a) it will come to rest with the point 'o' between points 'f' and b,
and, it having done so, the probability that
(b) the event M will then happen 'm' times in 'n' throws of the second ball,
is equal to the ratio of the area fghikmb to the area of the square ABCD[1].

Bayes' demonstration of the proposition takes a form he seems to favour, *i.e.* a geometrical demonstration that, if the proposition is not true, then a contradiction is entailed. The argument is again somewhat lengthy, and in the following exposition, we try to strike a reasonable balance between fidelity to Bayes' original form and a form which may be easier for our own contemporaries to grasp:-

*(1) Let $\mathcal{P}(m:n)$ denote the probability that the event M will happen on m occasions out of n throws of the second ball, and let us suppose it is equal to a ratio **G**/ABCD, where **G** is some area greater than fghikmb, so that $\mathcal{P}(m:n)$ is greater than the ratio fghikmb/ABCD.* (The bounding outer curve, *B i A* in Fig 4.4 may be envisaged as an example of an area having the property defined for ***G***)**.**

(2) At the points such as e,d,c, we drop perpendiculars from fb to meet the curve AmigB at h,i,k, respectively.

[1] For a diagram showing the full square *ABCD*, see *Ch 2* above, (*p387* in the original text).

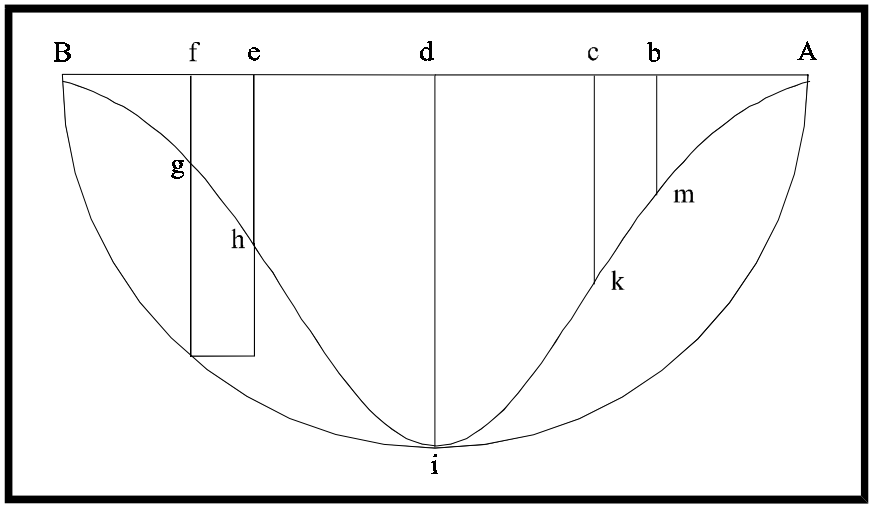


Fig 4.4

(3) Let the point d be selected so that the perpendicular di is the longest perpendicular between the line fb and the curve AmigB.

(4) We then select a number of points such as e and c, either side of d such that the sum of the areas of the rectangles such as bck, cdi, edi, feh differs less from the area of fghikmb than does the area ***G****.*

At this point, Bayes inserts the comment:-

all which may be easily done by the help of the equation of the curve, and the difference between ***G*** *and fghikmb being given,*
i.e., he is asserting that however small may be the difference between the hypothetical area ***G*** and the area of *fghikmb,* we can, by making the number of points on *fb* and the corresponding rectangles sufficiently numerous, cause them to approximate ever more closely to *fghikmb,* and hence for the sum of their areas to differ from the area of *fghikmb* by less than does the area of ***G***. Bayes also notes that, since the line *di* is, by definition, the longest of the perpendicular ordinates from the line *fb* to the curve, the other vertical lines will decrease in length as we move away from *di.* Hence if we consider vertical lines such as *eh,* descending from the base segment AB, we see that the ordinate *eh* is longer than *gf,* and is longer than any other ordinate that can be drawn from the base segment *ef.* The argument then continues:-

Now suppose that the first ball comes to rest so that the point o coincides with the point e; then, by Lemma 2, the probability that, in a single trial, the second ball will come to rest to the right of the perpendicular through e is

equal to the ratio Ae/AB; and, as this is the probability that the event M will happen in a single trial, we may write

$$\mathcal{P}(M) = Ae/AB = p \qquad (4\text{-}6)$$

and, by the corollary to Proposition 1, the probability that, in a single trial, the second ball will come to rest to the left of the perpendicular through e, is equal to the ratio Be/AB; and, as this is the probability that the event M will not happen in a single trial, we may write

$$\mathcal{P}(\sim M) = Be/AB = q \qquad (4\text{-}7)$$

Hence, by Proposition 7, the probability that, in 'n' trials, the event M will happen 'm' times will be:-

$$^{n}C_{m}\, p^{m} q^{(n-m)}$$

It follows that because the curve AiB has been drawn in such a way that the ratio eh/AB = y = $^{n}C_{m}\, p^{m}\, q^{(n-m)}$, the ratio eh/AB is equal to the probability that M will happen on m occasions in n trials.

Therefore, when the point o coincides with the point f, $\mathcal{P}$(m:n) equals the ratio fg/AB, and so on for any other ordinate.

Hence, as eh is the longest ordinate that can be drawn from the line-segment ef to the curve AiB, $\mathcal{P}$(m:n) cannot be greater than eh/AB if the point o is to lie somewhere on or between the points f and e.

Using the above result, together with the formula for the joint probability of consecutive events, as proved under Proposition 3, Bayes now shows that the probabilities of various joint events correspond to ratios which cannot be greater than those of the areas of various rectangles inscribed upon the line *f b* to the area of the square *ABCD*. He does this by considering consecutive events, E_1 and E_2 where:-

E_1 is the event that the first ball comes to rest such that the point 'o' is on or between points 'e' and 'f', the probability of which was shown by Lemma 1 above, to equal ef/AB.

E_2 is the compound event that, following E_1, the event M happens on m occasions out of n trials, the probability of which we have just proved cannot be greater than eh/AB.

It then follows from Proposition 3, that the probability that both E_1 and E_2 will happen cannot be greater than the product (ef/AB)x(eh/AB), which is the ratio of the area of the rectangle feh to the area of the square ABCD.

Likewise:

The probability that the point o will lie on or between points e and d, and that the event M will happen as before on m occasions out of n trials, cannot be greater than the ratio of the area of rectangle edi to the area of the square ABCD and

The probability that the point o will lie on or between points d and c, and that the event M will happen as before on m occasions out of n trials, cannot be greater than the ratio of the area of rectangle cdi to the area of the square ABCD.

The probability that the point o will lie on or between the points c and b, and that the event M will happen and fail as before, cannot be greater than the ratio of rectangle bck to ABCD.

Thus, if we now add together all such probabilities, then Proposition 1 tells us that their sum will be the joint probability that the point o will lie somewhere on or between point f and b, and that the event M will happen on m occasions out of n trials.

Hence, adding together the corresponding ratios identified above, their sum is the ratio of the sum of the areas of the rectangles (feh+edi+cdi+bck) to the area of the square ABCD.

But, because by definition, the area of the unknown ***G*** *is greater than (feh+edi+cdi+bck), the ratio*

$$(feh+edi+cdi+bck) / ABCD$$

must be less than the ratio $\boldsymbol{G}/ABCD$.

Therefore the probability that both E_1 *and* E_2 *will occur must be less than the ratio* $\boldsymbol{G}/ABCD$, *which contradicts the initial assumption that it was equal to* $\boldsymbol{G}/ABCD$ *: hence the assumption cannot have been correct.*

Conversely, by denoting an area less than fghikmb as, say ***L****, and assuming that the probability of* E_1 *and* E_2 *is equal to the ratio* $\boldsymbol{L}/ABCD$, *we can show, by considering rectangles inscribed within the curve, that the probability of* E_1 *and* E_2 *is greater than* $\boldsymbol{L}/ABCD$, *which is again a conclusion contradicting the assumption from which it was derived.*

Therefore, as the required probability can be neither greater than $fghikmb/ABCD$ *nor less than* $fghikmb/ABCD$, *it must be equal to the ratio of the area of* $fghikmb$ *to the area of* $ABCD$.

The corollary to Proposition 8

If, therefore, we make the point f coincident with the point B, and the point b coincident with the point A, so that the segment of interest comprises the whole line AB, then the probability that the point o will lie on the chosen segment and that the event M will subsequently happen on m occasions out of n trials is the ratio of the area of the whole curved area AiB to the area of the square ABCD. Hence, before the first ball is thrown, (and because it is certain that the point o will lie on the line AB), the probability that M will occur on m occasions out of n trials, is given by the ratio of the curved area AiB to the area of the square ABCD.

Proposition 9

If, before anything is known about the position of the point o, we learn that the event M has happened on m occasions out of n trials, and we then guess that the point o lies between any two points, such as f and b, on the line AB, this corresponds to a guess that the probability of M in a single trial is somewhere in magnitude between the ratio Ab/AB and Af/AB, and the probability that we are right in this guess is the ratio of the area fghikmb to that of the whole curved area AiB.

The proof of this, drawing upon Proposition 8 and its corollary, and drawing also upon Proposition 5, is as follows:-

Considering again the consecutive events E_1 and E_2 as above, and that:-
by the corollary to Proposition 8, the prior probability of E_2 is the area-ratio AiB/ABCD, and that
by Proposition 8 itself the probability of both events is the area-ratio fghikmb/ABCD, then, in the terms used in discussion of Propositions 4 and 5,

$$\mathcal{P}(E_1 \wedge E_2) = B/N = fghikmb/ABCD \qquad (4\text{-}8)$$

and

$$\mathcal{P}(E_2) = p_2/N = AiB/ABCD \qquad (4\text{-}9)$$

whence, if, knowing that E_2 has happened, we now guess that E_1 has also happened, then the probability that we are right is:-

$$B/p_2 = fghikmb/AiB \qquad (4\text{-}10)$$

The corollary to Proposition 9

This corollary presents some minor problems of terminology in the original text, the symbol *o* suddenly changing its significance from that of a

Version zh25a.doc at 18:36 on 13 July, 2001

point on the line *AB* to that of a zero numeric value. Hence we present this corollary as:-

If on the same suppositions we guess that the probability of M lies somewhere in magnitude between zero and the ratio of the lengths Ab/AB, then the probability that we are right is equal to the area-ratio Abm/AiB.

This concludes our exposition of Bayes' argument to the end of Proposition 9. In the next chapter, we examine the *Scholium*.

Chapter 5

The Scholium

In Bayes' essay, *The Scholium* follows Proposition 9. Its focus is on the event E_1, which is defined as occurring when P_m has a value in an interval bounded by two arbitrary values which we denote as x_1 and x_2. The value of P_m is the probability that the event M, as defined in the previous chapter, will occur when the second ball is thrown. The purpose of the *Scholium* is to discuss the difficult problem of assigning, or assuming a pre-trial, or *'a priori'* probability for the event E_1, when we have no information whatsoever, prior to the trials, about the probabilities attaching to the many different, but possible, values it could have. The difficulty is acute because, according to *Proposition 5*, even after a set of n trials, the probability that P_m is in the defined range can be determined only by knowing or assuming the pre-trial, *i.e.* the 'prior', probability of its being in that range.

To place this problem in context, we recall the crux of the essay:-

The probability that an event has happened is the same as the probability that I am right if I guess that it has happened. Wherefore:-

Proposition 5

If there are two consecutive events, the probability of the second event being p_2/N, and the probability that both events will occur together being B/N, if it is discovered that the second event has happened, and I then guess that the first event has also happened, the probability that I am right is B/p_2.

That is, we are concerned to evaluate the probability that event E_1 has happened, when it is known that event E_2 has happened, the answer being given in terms similar to *(3-51)* by:-

$$\mathcal{P}(E_1 \mid E_2) \quad = \quad B/p_2 \qquad (5\text{-}1)$$

Version zh25a.doc at 18:36 on 13 July, 2001

where

***'B'** is* *The prior probability that E_1 and E_2 should both occur in a joint trial*

and

***'p_2'** is* *The prior probability that E_2 should occur in an identical joint trial, but without regard to the occurrence of E_1 in that same trial.*

However, in the assumptions leading to the derivation of *(3-51)*, under Prop.4 and its corollary, Bayes does not assume that E_1 and E_2 are time-ordered; he merely assumes, to use modern terminology, that they may be correlated[1]. To give clearer correspondence with the experiment carried out on the table, we therefore define a two-part trial in which *part-A* comprises a single throw to determine the value P_m and *part-B* comprises a set of n individual, probablistic throws, each governed by the value of P_m obtained in *part-A* and in each of which the event M either happens or does not happen.

We now substitute into *(5-1)* propositional values as follows:-

E_1 *Is an event such that the value of P_m obtained in part-A is greater than x_1 and less than x_2 i.e. symbolically $(x_1 < P_m < x_2)$*

E_2 *Is an event such that in a set of n trials, the event M occurs in m of those trials.*

B *Is the probability prior to the test that:-*
(1) The value of P_m determined by part-A will be greater than x_1 and less than x_2
and, given (1), that
(2) The event M will happen m times and fail n-m times in part-B

$\mathcal{P}(m|n)$ *Is the probability that the event M will happen m times, in a test comprising n trials.*

Hence, the problem is to evaluate:-

$$\mathcal{P}\{(x_1 < P_m < x_2) \mid (m,n)\} = B/\mathcal{P}(m:n) \qquad (5\text{-}2)$$

which requires us to provide numerical values for B and $\mathcal{P}(m|n)$, given that, by the definition of the problem, we have no prior knowledge of any of the probabilities involved. We know only the general nature of the test set-up and the values of x_1, x_2 and n which we have arbitrarily selected.

[1] A simple analysis of the general case of correlation is given by Keynes (1921) Ch XXXI.

In considering how the above problem is handled by Bayes, we learn, from the covering letter which Richard Price wrote when he sent Bayes' essay to John Canton, that Bayes deliberately kept the *Scholium* aside from the main body of his mathematical reasoning. Price also tells us that Bayes had quickly perceived that he could solve the general problem addressed by the essay:- *provided some rule could be found according to which we ought to estimate the chance that the probability for the happening of an event perfectly unknown, should lie between any two named degrees of probability, antecedently to any experiments made about it; and that it appeared to him that the rule must be to suppose the chance the same that it should lie between any two equidifferent degrees* [1]. The reader is thus conditioned to expect an encounter with the notorious 'postulate of the uniform prior distribution': but this is a point on which we must ask for caution, particularly when considering the role of the uniformly level table; for the table is, as we show later, distinct from the general issue and serves essentially to demonstrate the application of Bayes' approach in a particular case. Thus, in separating the *Scholium* from the main part of the essay, Bayes seems to have been concerned to separate metaphysical disputes from the mathematics. It is also worth noting that in the definition of *probability*, he adopted a purely pragmatic approach in which the meaning of the term was equated to a rule governing its use: an approach which he appears to have justified in the missing Introduction. Hence it may be significant that we again find the words *'the ... rule ...to be used'* in the *Scholium,* which we now quote, in modern print, paraphrasing slightly, and with some small changes also to the layout[2]:-

Scholium

(1) From the preceding proposition it is plain that in the case of such an event as I there call M, from the number of times it happens and fails in a certain number of trials, without knowing any thing more concerning it, one may give a guess whereabouts its[3] probability is, and, by the usual methods of computing the magnitudes of the areas there mentioned, see the chance that the guess is right. And that the same rule is the proper one to be used in the case of an event concerning the probability of which we know absolutely nothing antecedently to any trials made concerning it, appears from the consideration that, concerning such an event, I have no reason to think

---

[1] p 371 in the original text.

[2] The Scholium is on pp392-394 of the original text. The rendering here paraphrases certain phrases and introduces paragraph breaks and numbers which are not in the original. The paraphrases are marked by footnotes.

[3] *it's* in the original text

that, in a certain number of trials, it should happen any one possible number of times than another. For on this account, I may justly reason concerning it as if its probability had been at first unfixt, and then determined in such a manner as to give me no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another.

(2) But this is exactly the case of the event M. For before the first ball is thrown[1]*, which determines the probability of the event M in a single trial*[2]*, (by cor. prop.8) the probability it has to happen m times and fail n-m times in n trials is the ratio of AiB to CA, which ratio is the same when n is given, whatever number m is; as will appear by computing the area of AiB by integration*[3]*.*

(3)[4] *.... (which shows) ..... that AiB contracted in the ratio of* ${}^{n}C_{m}$ *to 1 is to CA as 1 to (n + 1); from whence it plainly follows that, antecedently to this contraction, AiB must be to CA in the ratio of 1 to n+1, which is a constant ratio whenever n is given, whatever m is.*

(4) And consequently before the place of the point o is discovered or the number of times the event M has happened in n trials, I can have no reason to think it should rather happen one possible number of times than another.

(5) Therefore I shall take for granted that the rule given concerning the event M in Prop.9 is also the rule to be used in relation to any event concerning the probability of which nothing at all is known antecedently to any trials made or observed concerning it. And such an event I shall call an unknown event.

Although the full argument which is put forward by the *Scholium* may not be easy to grasp at first sight, particularly if one has been unwittingly prejudiced by hearsay, it seems clear enough, if one reads carefully what it says, that its basis is the plain fact, stated in *(1)* , that *'in the case of an event concerning the probability of which we absolutely know nothing antecedently ..... I have no reason to think that, in a certain number of trials, it should happen any one possible number of times than another'.* Bayes does not however proceed directly from *(1)* to *(5)*, but in *(2)* he asserts that:- *this is exactly the case of the event M .... before the first ball is thrown ..... the probability it has to happen m times and fail n-m times in n trials is ..... the same when n is given, whatever number m is*. In *(3)* he asserts that this can

---

[1] *i.e.* the ball W in terms of Ch.4.

[2] the original text says simply *it's probability*

[3] In the original text *the method of fluxions*

[4] This section is printed as a footnote to the original text

be proved algebraically. Whence, in *(4)* he reiterates that:- *'consequently ..... I can have no reason to think it should rather happen one possible number of times than another',* and thence to *(5) ..... therefore I shall take for granted that the rule given concerning the event M in Prop.9 is also the rule to be used in relation to any event concerning the probability of which nothing at all is known antecedently*

However, if we take the description of the test up to this point, literally as it stands, it involves a slight contradiction between the definition of an *unknown event* and the fact that it is only on completion of the test that we know the value of m - knowledge which is highly relevant to the estimation of the probability in question. Our solution to this difficulty is therefore to re-formulate the problem so that we are required to find, prior to conducting the test, a rational basis on which to estimate, for each possible value of m in the range 0 - n inclusive, the probabilities:-

(1) That E_1 will happen in *part-A*,
(2) Given E_1, that the event M will happen m times in *part-B*.
(3) That, regardless of E_1, the event M will happen m times in any trial-set of size n.

To (1), the answer of the *Scholium* is that:- *'the same rule'*, (*i.e.* the rule given under Prop.9), *'is the proper one to be used'*. The problem posed by (2) is that addressed and solved in Prop.8, albeit the technical problem of algebraic determination of the area *fghikmb* was one of some difficulty, with which Bayes struggled in the final part of the essay, and which is outside the scope of our enquiry[1]. The problem with (3) is then to find a rational basis on which to evaluate $\mathcal{P}(m \mid n)$ for each value of m in the range 0 - n inclusive, *i.e.* the prior probability that, in *part-B* of the test, and knowing nothing whatsoever concerning the actual or probable outcome of *part-A*, the event M will happen m times. Let us suppose, for example, we are told only that the set will comprise *100* trials and we are asked to assess the probability that the observed value of m will be *3*: can there be any rational basis on which to provide the requested estimate? Superficially, the answer appears to be that, unless we make further assumptions or are given further information, there can be no rational basis on which to give any estimate of the requested probability.

This does not, however dispose totally of our ability to grapple with the problem; for an alternative response, as Bayes perceived, is that we can also express the situation in terms of having no reason to rate any given

[1] See Stigler (1986) p 130

Version zh25a.doc at 18:36 on 13 July, 2001

value of the probability more highly than any other possible value. Bayes also perceives that we know that P_m is constrained to fall in the range *0 - 1* and that this is logically equivalent to the situation addressed in the Corollary to Prop.8. There, Bayes argues, the value of $\mathcal{P}(m \mid n)$ is given by the ratio of the curved area *AiB* to the area of the square *ABCD* . Referring back to Prop.8, we find that the curve *AiB* was constructed by means of the formula *(4-4)*:-

$$y_b \quad = \quad {}^{n}C_{m}\ p_b^{m}\ q_b^{(n-m)}$$

which, when integrated with respect to *p*, appears to show, as Bayes then points out, that the area *AiB* is a function only of *n*, and is independent of *m*[1]. Hence, he argues, the value of $\mathcal{P}(m \mid n)$ is itself a function only of *n* and is identical for all values of *m* in a trial-set of a given size. Bayes suggests that this corresponds with a situation in which the value of $\mathcal{P}(m \mid n)$ is fixed by a prior process in which all values of $\mathcal{P}(m \mid n)$ are equally probable, and as a result of which we have no reason to rate any one value more highly than any other value. He suggests therefore that, the rule developed under Prop.9, is the rule we should use to find the probability that P_m lies between the given values when we are faced by an 'unknown event'.

Having expounded the argument, we now seek to examine, for ourselves, its validity, particularly with regard to the assumption of the uniform prior distribution. It is not however vital to the logic of the experiment that different balls shall be used in the two parts of the experiment, nor indeed is the order in which the balls are thrown. For we could construct an experiment which is logically equivalent to that performed by Bayes, in which a single ball is thrown *n+1* times, and, after each throw, we mark the position in which the ball comes to rest. We then select one of these positions as the reference position to establish the point *o*, as if it had been established by a throw of the ball in *part-A*. It is also worth noting, at this point, because of the important part played by the uniform prior distribution of P_m in Bayes' argument, that there is a temptation to assume that the uniformly level table is correspondingly important. However, a uniform prior for P_m does not require a level and unbiased table but can be achieved by any cumulative probability function which increases across the table and which equally affects the ball in both *part-A* and *part-B*, such that a one-to-one mapping can be achieved between the uniform continuum defined by Bayes and the space in which any given experiment takes place.[2] The situation is however

---

[1] The assumptions hidden in this assertion are discussed later in this chapter. See also refs. to Molina in later chapters.

[2] See also Edwards (1974, 1978).

quite different if, say, an iron ball is used and a magnet is held under the table to produce a bias in *part-A* but is removed in *part-B* in order to restore a uniform distribution of the point at which the ball comes to rest.

Returning to Bayes' argument, a point of great importance in the context of its time, is simply the emphasis which he places on the rôle of the prior distribution[1]. However, as Stigler pointed out[2], and as we have seen above, a careful study of the Scholium clearly shows that the thrust of Bayes' argument is, repeatedly, from having *'no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another'*, to the conclusion in *(5) 'that the rule given ..... in Prop.9 is also the rule to be used in relation to any event concerning the probability of which nothing at all is known antecedently to any trials ....'..* Unfortunately, the words of the Scholium on this matter were heavily pre-empted in the covering letter which Richard Price wrote when he sent Bayes' essay to John Canton. That letter was printed as an introduction to the essay and seems, alas, to have set alight a smoke-screen of prejudice which has only been challenged in recent years[3]. Yet, if Richard Price is to be believed, the obstacle originated with some remarks made by Thomas Bayes himself, in an Introduction to the essay which he had written at some point, but which was not published with the essay, and which cannot now be traced[4]. Price, however, tells us in his covering letter, that Bayes, perceived that he could solve the problem he had posed:- *by supposing the chance the same that the probability P_m should lie between any two equidifferent degrees*[5], and against which the critics have unceasingly railed. This is, however, not perhaps surprising, for Price continues:- *'(but Bayes) afterwards considered that the postulate on which he had argued might not be looked upon by all as reasonable; and therefore he chose to lay down in another form the proposition in which he thought the solution of the problem is contained, and in a Scholium to subjoin the reasons why*

---

[1] A point which Bayes had emphasised also in a letter to John Canton concerning observational errors, *'he supposes that the chances for the same error in excess or defect are exactly the same, and upon this hypothesis only has he shown the incredible advantage'*. Canton papers, vol. 2, p32. Quoted in Stigler (1986) p94.

[2] Stigler (1986)

[3] Notably by Edwards, (1974,1978) and Stigler(1982, 1986), to whom thanks are due for tackling an obstacle which had long impeded access to Bayes' argument. But see also Dale (1991) p39, section 3.5.

[4] See Barnard (1958)

[5] This and the following quotation are slight paraphrasings. The original version is given in Ch. 2 above.

Version zh25a.doc at 18:36 on 13 July, 2001

he thought so, rather than to take into his mathematical reasoning any thing that might admit dispute'.

The first point to note in considering this quotation is that we have here a statement attributed to Bayes by Richard Price, and although we have no reason to doubt the honesty or accuracy of that attribution, we have no means of knowing when, in relation to the other parts of the essay, those views were expressed by Bayes. As we have noted previously, there are marked stylistic differences and these may well indicate that various parts of the essay were written at different times, and possibly in quite different order from that in which they are now printed. Hence the reservations to which Price refers may well have stemmed from Bayes' views when he wrote the *Scholium*, his death having prevented any later revision. There is, however, no firm evidence that Bayes himself entertained any doubts upon the point in question: the only doubts of which we have any evidence are Bayes' doubts concerning the opinions of others. A corollary to this is, however, that at the time Bayes wrote the Introduction to which Price refers, Bayes had not seen how to show beyond doubt that the assumption of the uniform prior distribution is the rule to use in the general case. This is perhaps a point of some significance for our understanding of the human, chronological and philosophical aspects of the matter. Hence, putting aside the possibility that Bayes' use of the phrase *'I shall take for granted'* may, (or equally may not), indicate some residual doubt in Bayes' own mind, the question we have to address is whether we can agree that *'the rule given concerning the event M in Prop. 9 is also the rule to be used in relation to any event concerning the probability of which nothing at all is known antecedently to any trials'* ?

A particular difficulty, however, is the presence of the argument that since we have no reason to suppose any particular value of m to be more probable than any other value, our situation is logically equivalent to that which would follow from the value of P_m being determined by a prior process in which all values of m are equally probable. Certainly if we take also into consideration the physics of an actual experiment, we see that Bayes has pointed out an additional element which, as a matter of physical logic, has an essential role in any such experiment, *i.e.* the necessary existence of a prior process by which the value of P_m is determined. But the reasoning is fallacious if we equate a state of ignorance concerning the prior probabilities of different hypotheses, with the state of our knowledge if we know that the probabilities have been made equal by the operation of a prior process. Philosophically, and indeed in common life, we cannot, and we do not, argue that because a state of ignorance on a particular matter has a consequence in common with a state of factual knowledge, that the former state can be

equated with the latter ! We might not know an Emu if we saw one, and we might not know the name of the bird on the lawn, but we do not thereby infer that it is an Emu.

But Bayes, in fact makes no such inference, and it is inconceivable that he would have committed such a patent fallacy. All he says is that, from a pragmatic point of view, in performing the calculations, we use the same rule in both cases. (If we have heard that Emus are dangerous, we might be well advised to give the unknown bird a wide berth, just in case it should indeed be an Emu.) Nevertheless, it seems that all who read the *Scholium* feel uneasy at the lack of any clear, formal justification, to the standard found elsewhere in the essay, for the use of the rule, and for many years it has been commonplace to believe that Bayes' conclusion was at best of very limited application, and, at the worst, fallacious[1]. Also, as Murray and Molina later showed,[2] a uniform prior is necessary if the value of $\mathcal{P}(m \mid n)$ is to be uniform over all values of m. Although the tone of their papers is sympathetic to Bayes, it seemed thereafter clear that any reliance by Bayes upon the prior uniformity over the possible values of m, must weaken his case. Although the proofs used by Murray and Molina involve some fairly difficult mathematics, it is easy to see that, if one were able to bias arbitrarily the resting place of the ball in *part-A*, but not in *part-B*, then it would destroy the topological equivalence to uniformity over both parts of the trial and thereby destroy the uniformity of probability over the possible values of m. However, although there are, in the *Scholium*, strong hints in the direction of positively equating the unknown probabilities with equal probabilities, Bayes nowhere makes that equation explicit. Which is why, perhaps, some sixty years after Bayes' death, Laplace was moved to write:- *'Bayes produced, albeit with some embarassment, a solution which was both elegant and ingenious '*.[3]

The difficulty is that we are, by the very definition of the problem, in a state of absolute ignorance concerning P_m, and we have so far seen no reasons for believing, nor even for assuming as a pragmatic rule, that all the possible values of P_m in the range *0 - 1* are equally probable. Therefore, we really have to consider the matter in some depth. So, let us define the prior probabilities over the values of P_m by an increasing cumulative distribution function, *F(x)* such that:-

---

[1] *cf* Hacking (1965) pp 200-201.

[2] Murray F.H. (1930) and Molina E.C. (1931). See also the discussion in Ch 8.

[3] *'il y a parvenu d'une manière fine et très ingénieuse, quoiqu'un peu embarassée'* Laplace (1820) p iv (Preface)

Version zh25a.doc at 18:36 on 13 July, 2001

$$\mathcal{P}(E_1) \quad = \quad \int_{x_1}^{x_2} dF(x) \quad = \quad \left[F(x) \right]_{x_1}^{x_2} \qquad (5\text{-}7)$$

Then, in general, we know, or assume, by Bernoulli's theorem, that when we are given P_m, the probability of m events in n trials is[1]:-

$$^nC_m \, P_m{}^m \, (1 - P_m)^{(n - m)} \qquad (5\text{-}8)$$

so that, explicitly introducing $F(x)$ as a prior probability weighting function over the values of x, the numerator in *(5-1)* becomes:-

$$\mathcal{P}((E_1 \wedge E_2) \mid F(x)) \quad = \quad {}^nC_m \int_{x_1}^{x_2} x^m \, (1 - x)^{(n - m)} \, dF(x) \qquad (5\text{-}9)$$

and the denominator in *(5-1)* becomes the value given by integrating *(5-9)* over all the possible, and duly weighted values of P_m such that:-

$$\mathcal{P}(m \mid (n, F(x))) \quad = \quad {}^nC_m \int_{0}^{1} x^m \, (1 - x)^{(n - m)} \, dF(x) \qquad (5\text{-}10)$$

whence

$$\mathcal{P} \{ (x_1 < P_m < x_2) \mid (m,n, F(x)) \}$$

$$= \frac{{}^nC_m \int_{x_1}^{x_2} x^m \, (1 - x)^{(n - m)} \, dF(x)}{{}^nC_m \int_{0}^{1} x^m \, (1 - x)^{(n - m)} \, dF(x)} \qquad (5\text{-}11)$$

At this point, as the grounds for assuming a uniform prior distribution look remarkably bleak, it may be appropriate to ask whether our quest is reasonable: do we truly have any grounds for believing that an answer is possible ? Is it indeed reasonable that, when we have conducted n trials and obtained m occurrences of the previously unknown event M, that we should feel justified in having some idea of the magnitude of P_m over this set of trials? And do we have any justification for believing that we ought to be able also to quantify the uncertainty concerning the true whereabouts of P_m on the basis of the evidence provided by the trials? From this sceptical point of view, therefore, we can see no grounds on which to assess our uncertainty, and it would seem reasonable indeed to doubt that grounds can be found.

[1] See Keynes(1921) Ch XXIX (p337 *et seq.*), for an exposition of the limitations which apply to the validity of Bernoulli's theorem .

Yet, from another point of view, if we accept Bernoulli's theorem and believe that, as n increases, so $m/n \rightarrow P_m$, it seems an insult to our reason to suggest that we can conduct a lengthy test of the form described, obtain the results, and yet be no better informed on the value of P_m than before we undertook the trials. This leads us, therefore, to look further into the implications of asserting that, on the evidence available, it is more probable that P_m is equal to the ratio m/n than that to any other value, for there is indeed no evidence to support any other value, and the only evidence which we have, is that which supports the value m/n.

To examine this point in more detail[1], we consider the cases when $F(x)$ is at least twice differentiable so that, in terms of probability density, the post-trial probability that P_m is within an infinitesimal element[2] dx at any point in x is given by:-

$$\mathcal{P}(P_m \approx x) = \frac{x^m (1 - x)^{(n - m)} dF(x)}{\int_0^1 x^m (1 - x)^{(n - m)} dF(x)} \qquad (5\text{-}12)$$

Hence, to accord with a belief that, relative to our knowledge and reasonable assumptions, (for there can be no other rational basis for any belief on such matters), m/n is the most probable value of P_m, we require that *(5-12)* shall have a unique maximum when $x = m/n$. However, because the denominator is integrated over all possible values of x, its value will not vary with different values of x in the numerator. Hence the result which we require is that which maximises the numerator with respect to x. Further, since the value of the term $x^m(1 - x)^{(n - m)}$ is determined, for each value of x, by the result of the test, any additional constraints must operate on $dF(x)$ to ensure that the maximum occurs at the required point. To resolve this, in cases where $F(x)$ is such that its second differential with respect to x exists, and to keep the notation tractable, we define a function $g(.)$ such that:-

$$g(x,m,n) = x^m (1 - x)^{(n - m)} \qquad (5\text{-}13)$$

and, hence we can re-express *(5-12)* as:-

---

[1] The next few paragraphs contain some rather arid mathematical pedantics, the essence of which is to show that an assumption of a non-uniform prior will, in general, produce an indefensible result with an apparently most-probable value other than at m/n. The general reader may proceed without loss to the paragraph marked with an asterisk.

[2] In this case, $\int (dF(x)/dx)dx = \int dF(x)$: See e.g. Rao (1973)

$$\frac{g(x,m,n)\,dF(x)}{\int_0^1 g(x,m,n)\,dF(x)} = \frac{x^m\,(1-x)^{(n-m)}\,dF(x)}{\int_0^1 x^m\,(1-x)^{(n-m)}\,dF(x)} \qquad (5\text{-}14)$$

Then, denoting the differentials of $dF(.)$ and $g(.)$ with respect to x by $F''(.)$ and $g'(.)$ respectively, we require, at the point where *(5-14)* has its maximum, that:-

$$g(x,m,n)F''(x) + g'(x,m,n)dF(x) = 0 \qquad (5\text{-}15)$$

However, $g(x,m,n)$ already has a unique, positive, non-zero maximum when $x = m/n$. We therefore let $\xi = m/n$, whence

$$g'(\xi, m, n)\,dF(\xi) = 0 \qquad (5\text{-}16)$$

and, to satisfy *(5-15)*, we require also that

$$g(\xi,m,n)\,F''(\xi) = 0 \qquad (5\text{-}17)$$

which implies that $F''(\xi) = 0$ whatever the value of ξ, *i.e.* we require that $dF(x)$ shall have the same value for every value of x over the open set $(0 - 1)$.

For completeness, however, we must also consider the possibility that $F(x)$ might not be twice differentiable, as assumed above, but might contain 'steps'. Physically, such a situation could be created by placing an iron bar under the table and placing a magnet inside the ball used in *part-A*. We would then have a concentration of probability along the line of the bar, producing a sharp step in the cumulative probability function at the position of the iron bar as illustrated in figure 5.1:-

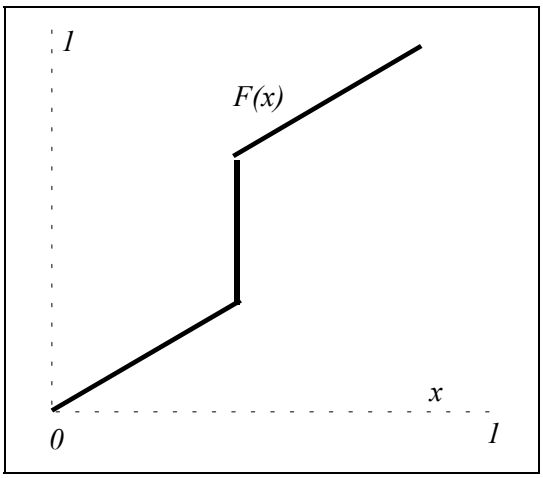


Fig 5.1

If we then consider the impact of applying such a function as a multiplier over *g(x,m,n)*, it is immediately clear that any non-uniform prior, would, in general, seriously disturb the natural maximum of *g(x,m,n)* at $x = m/n$, illustrated in figure 5.2[1].

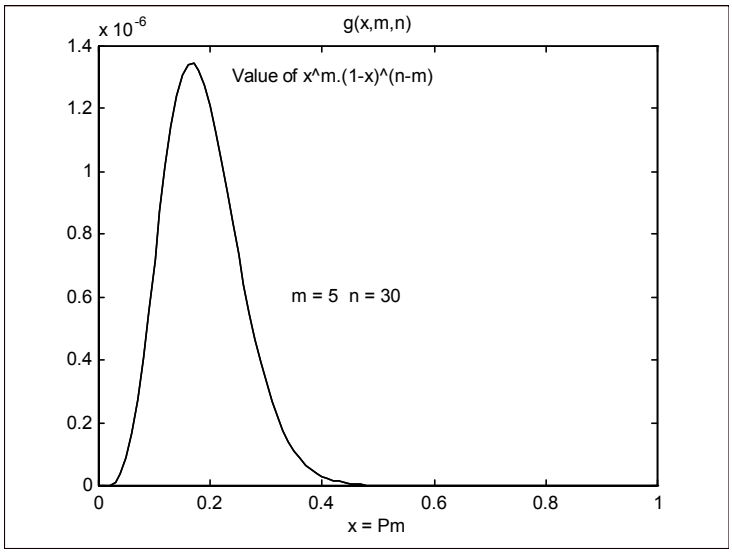


Figure 5.2

This can be shown more formally[2] by using the concept of alignment between elements in a given space and in its dual, *i.e.* between *g(x)* and *dF(x)*, and hence showing that because *g(x)* has a unique maximum when $x = m/n$, it is necessary that *dF(x)* shall be constant for all values of x in the open interval $(0,1)$ if that maximum is to be preserved. We then have two special cases to consider where spikes could be present in *dF(x)* without affecting the maximum of *g(x)*. The first such case occurs if *dF(x)* has a finite[3] spike at $x = 0$. This would have no effect upon the maximum of *g(x)* for values of $m/n > 0$, because in all such cases the value of *g(x)* is itself zero. In the particular case where $m = 0$, and *g(x)* has its maximum at $x = 0$, the effect of the spike at $x = 0$, is simply to reinforce the maximum of *g(x)*. Conversely, the only effect of a spike at $x = 1$ would be to reinforce a maximum of *g(x)* at that point due to a test returning a value of m equal to that of n. However, while such spikes at the extremes would not contradict a requirement to infer

[1] See also Ch 12, Figs 12.2 and 12.3

[2] Luenberger (1969) Ch 5.

[3] *i.e.* the total probability within the spike must be less than *1*. If it is equal to *1*, it implies no possibility of $m > 0$.

the value of *m/n* as the most probable value of P_m, an assumption of their presence would be totally at odds with the assumption of prior ignorance.

* We conclude therefore that, when we interpret the results of Bayes' experiment, the assumption of a uniform prior distribution of probability for P_m, over the closed unit interval, is a necessary condition for any rational inference regarding the probable value of P_m. If we refuse to make that assumption, then we are making an implicit assertion that observation can tell us nothing about the probable value of such a parameter. A case indeed of 'Hobson's Choice[1]'. Which, in itself, proves nothing. But Nature, in the form of experiments, rarely - arguably never - proves anything. Nature provides evidence - which we are beholden, rationally, to interpret.

However, if we now go back to the *Scholium*, and read it very carefully, we find that it can be understood in several different ways. Stigler[2] suggests an interpretation in terms of a uniform distribution over the possible outcomes of a Bayes trial, *i.e.* in terms of the observables, rather than in terms of the unobservable *a priori* probabilities. But Stigler's argument is not, we find, compelling. Indeed, another interpretation is that Bayes is actually arguing by analogy between the probabilities of outcomes in the experiment and the probabilities of outcomes in the case of the unknown event. On this interpretation, his argument by analogy is, precisely as Murray and Molina showed by analysis, that a uniform prior over the possible values of P_m is a necessary condition for uniformity of probability over the outcomes. While it is arguable, as so many have indeed argued, that the assumption of a uniform prior for P_m is a contradiction of our axiomatic ignorance on this very point, it is also arguable that Bayes may have actually been asserting that it is precisely the uniformity which represents our ignorance - *i.e.* the prior uniformity necessarily and uniquely entails that *'before the place of the point o is discovered or ....(before).... the number of times event M has happened in n trials .... (is discovered) ..... , I can have no reason to think it should rather happen one possible number of times than another'*.

Before we leave this discussion of the *Scholium*, it is as well that we consider a further consequence of assuming a uniform prior, for we can ask, and perhaps we ought to ask, how good is the estimate which it yields? Clearly, in cases of total ignorance concerning the prior itself, we shall be unable to answer this question, taken simply as it stands. That inability

---

[1] In the days before cars and trains, Mr Hobson hired horses to travellers. The choice of horse offered to his customers was "Take it or leave it."

[2] Stigler (1982)

might seem to indicate a fairly serious weakness in the argument. Yet, if we consider a case where we are given prior information about the value of P_m but we ask to be told also the probability that the information is in some way defective, we are likely to receive, at best, a vague answer. For, in general, by asking to be told the recursive values of probabilities, *i.e. 'What is the probability that the probability that the probability ..... ',* one will very quickly push any respondent into an admission of ignorance; and this, it seems, is likely to apply to even the best constructed and authenticated experiment. Hence, we are led to suggest that, in the case of Bayes' experiment an acceptable evaluation of our estimate of $\mathcal{P}(x_1 < P_m < x_2)$ is the simple statement that it is the 'best possible'. This is, we would suggest, a completely objective statement of a kind which is widely accepted in a great range of human affairs. Indeed, in the case of Bayes' experiment, the strength of the statement is remarkable, for not only is it the best possible estimate, but, denoting our assumption of the uniform prior by $P_u(x)$, the estimate of probability given by Bayes' rule, *i.e.*

$$\mathcal{P}\{ (x_1 < P_m < x_2) \mid (m, n, P_u(x)) \}$$

$$= \frac{\int_{x_1}^{x_2} x^m (1 - x)^{(n - m)} \, dx}{\int_{0}^{1} x^m (1 - x)^{(n - m)} \, dx} \qquad (5\text{-}18)$$

is the only rational estimate. This does not mean that we have here a good estimate, for as Ehrenberg[1] asks in a not wholly dissimilar context, we also have to ask *'How good is best ?'*. While the best can be indeed pretty awful, the alternative, here, is silence.

[1] Ehrenberg (1982)

Chapter 6

Probability and Expectation

The purpose of this chapter is to discuss some historical and logical points concerning the formal 'Problem in the Doctrine of Chances' addressed by Bayes, and, in particular, the concepts of Probability and Expectation which he employs in its solution. The problem is also of historical interest and seems to have been posed first by James Bernoulli. Although this is not mentioned in the main body of Bayes' essay, it is noted quite prominently by de Moivre, who commends Bernoulli thus:- *'I .... shall conclude this remark with a passage from the 'Ars Conjectandi' of Mr James Bernoulli, Part IV Cap 4, where that acute and judicious writer thus introduceth his solution of the problem for:- Assigning the limits within which, by the repetition of Experiments, the probability of an event may approach indefinitely to a probability given, 'Hoc igitur est illud Problema &c'. This says he is the problem which I am now to impart to the Public after having kept it by me for twenty years: new it is, and difficult, but of such excellent use that it gives a high value and dignity to every other branch of this doctrine'.* [1]

Richard Price, however, in his introduction to Bayes' essay, is careful to point out the limitations in the problem addressed by Bernoulli and de Moivre, and goes on to suggest that Bayes' approach is superior, in that it is *'more directly applicable..... (to) .....confirm the argument taken from final causes for the existence of the Deity'*: a justification which could easily strike one, today, as hollow if not indeed thoroughly disingenuous. The sincerity and ability of Richard Price are not, however to be doubted, for it is well-established that his arguments on such matters were treated seriously by many people, including David Hume[2]. A related point is the concern of Price, in his covering letter, to assert the serious nature of the study of probability. De Moivre also took pains in the Dedication of his work to contradict *many who are possessed with an opinion that the doctrine of chances has a tendency to promote Play*[3]. However, nowhere in the essay does Bayes

---

[1] de Moivre (1756, p254).

[2] See Gillies (1987), Thomas, D.O. (1977), and Mossner (1954). Bernstein (1996) gives an admirably succinct picture of the respect accorded to Price. See also Broad, C.D. (1918).

[3] *i.e. gaming or gambling for money*

Version zh25a.doc at 18:36 on 13 July, 2001

even hint at the relevance of his subject to gaming, nor indeed does Price in the covering letter. The omission from the essay could be explained on the grounds that its concerns were purely with the logic and mathematics of the problem, and required no external justification. Indeed, inclusion of such reasoning within the main body of the paper could have been deemed improper, (*cf* the reason given by Price for the separation of the Scholium), but no such restriction would have applied to Price's covering letter. De Moivre, in contrast, has no hesitation in acknowledging the social problem presented by the gaming fever, which, in England, at that time, had reached, together with gross consumption of gin and port wine, epidemic proportions[1].

With regard to the technical content of the problem, Bayes' formulation advances on that of Bernoulli in several respects, and a full discussion of the differences is given by Stigler[2]. In brief, Bayes' problem is better defined and is more general than that of Bernoulli. Bayes also is strong in his insistence that the trials relate to an *unknown event*: a point not mentioned by Bernoulli. Although Bayes does not define 'an unknown event' in the Definitions, he does so at the end of the Scholium *viz:- '..... any event concerning the probability of which nothing at all is known antecedently to any trials made or observed concerning it*. However, it is ironic that although Bayes addresses only those problems in which we have absolutely no *a priori* information, we find that nowadays people often use the term 'Bayesian' for problems of precisely the opposite kind, *i.e.* in which *a priori* information is indeed available: or is invented.

Turning now to Bayes' Definitions, the first four, concerning *Inconsistent events, Contrary events, Failure of an event, and a Determined event,* are straightforward. Definition 5 is however surprising and is central to the concern of this chapter:- *The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening.* In Definition 6 Bayes then tells us that when he uses the word *chance*, he means the same as *probability*, thus showing that in his time, as in ours, the tendency to use any one of a number of words to denote this same underlying concept was already quite strong. An interesting contrast however to the meaning which Bayes assigns to *expectation,* and in a context practically identical to that set by Bayes, is provided by Hartley [3]:- *An ingenious*

---

[1] Pearson (1978). See also the reference to Turberville (1926) on p26, above.

[2] (1986, Chaps 2,3) See also Hacking (1975,Ch 17).

[3] Hartley, D. (1749), p339

friend has communicated to me a solution of the inverse problem, in which he has shown what the Expectation is, when an event has happened p times and failed q times, that the original ratio of the causes for the happening or failing of an event should deviate in any given degree from that of p to q[1]. Hence it is clear that, in Bayes time, the word *expectation* had a variety of meanings, but, in Bayes essay, the meaning seems to be consistently that of a monetary value.

In more recent times, there have been, various attempts to assign other, associated, words to particular aspects of the underlying concept. Possibly the most successful of these attempts has been that which would reserve the word *'Likelihood'* to denote a measure of the relative degrees to which various hypotheses are supported by some given set of evidence. In specialist writings, such as those of Fisher and his followers[2], a semantic convention of this kind is acceptable, provided it is clearly defined for the sake of readers not privileged to be of the *cognoscenti*: but this does not alter the fact that, in the general use of the English language, there is no such convention. In that wider context, *probability* and *likelihood* are as close in meaning as are *probability* and *chance,* and every person is entitled to use these words in this way, in normal discourse, without apology. Theoretically, we could argue that scientific discourse would benefit considerably if we could achieve general agreement to reserve these various words for particular uses. Thus, we might reserve use of the word *chance* for situations involving, or analogous to, the throwing of dice. The word *credibility,* we might reserve, as suggested by Bertrand Russell[3] to denote the degree to which it is rational to believe an hypothesis or assertion. For, although it is often doubtful whether we can rationally form any view of the absolute degree to which an hypothesis may be credible, the fact is that, in ordinary speech, people do use the word *credible* in this sense, perhaps indicating thereby a belief in some common body of knowledge, or assumptions, and of proper modes of inferring some measure of credibility which, if not absolute, is at least generally accepted.

Another aspect of Bayes' equating of *chance* with *probability,* concerns the recursive use of these words, as mentioned in Chapter 3, *i.e.* we find it awkward to talk about the *'probability that the probability .....',* but we have no difficulty with a phrase such as *'the chance that the probability'.* However, while this variation may be soothing for our minds, it distracts

---

1 Cited by Stigler (1983)

2 See also the discussion in Chapter 7 below.

3 Russell (1948), p359. See also Hacking (1975), p158.

attention from the remarkable recursive property of the word *probability* and its synonyms. That is, it seems intuitively clear that, if we are concerned to establish the value of some underlying probability, such as of the kind which Bayes is concerned to measure, then it makes sense to talk of 'the probability that the probability has a value in some specific range', This can however obscure the fact that in this single phrase, we can use the word *probability* in two different senses, one sense being 'epistemological', *i.e.* concerned with how certain or uncertain we are on a matter of knowledge, the second being 'aleatory' and concerned with the frequency of an occurrence in a dice-like trial[1]. More obscure, however, are the rules and conditions under which we can validly, and meaningfully, use recursive expressions such as *'the probability that the probability ..... '.*

Bayes' concept of the *'value of an expectation'* has however little direct connection with the modern concept of an *'expected value',* for the meaning which Bayes attributes to the word 'value' is very close to that which we mean when we speak of monetary value and he links it closely to the probability of a singular happening of an event in a trial. In mathematics of our own time, the *expected value* is often taken to mean a summation of parametric magnitudes multiplied by the corresponding probabilities *i.e.*:-

$$\mathcal{E}(x) = \Sigma [\, x_i \times \mathcal{P}(x_i) \,] \;;$$

the units of which are the units of x and are not necessarily similar to a monetary value: a point which Jeffreys misses in his brief discussions of this matter[2]. Historically, however, the matter is not straightforward, for we find that de Moivre, whose writings, we would expect to have been known to Bayes, says forthrightly [3]:- *'The method of Huygens ..... I was absolutely resolved to reject',* drawing our attention to the fact that, in the *'De Ratiociniis in aleae ludo'* of 1657, Huygens takes the notion of an expectation as primitive and from it derives the value of a probability. In Bayes' case, however, it seems clear from his definition of probability that the values of expectations are scalar monetary values, amenable to multiplication, and hence, we may assume, amenable also to addition, subtraction and division: properties which are, as we see later, vital to the cogency of the argument. Hence, by defining this simple relationship between probability and monetary value, Bayes creates an instrument of remarkable power by which we can judge the degree to which propositions concerning mathematical opera-

---

[1] See Hacking (1975), p 149.

[2] Jeffreys, H. (1983), p53. Hacking (1975) devotes a chapter to discussion of expectation (Ch 11).

[3] In the Preface to the 1717 edition of *The Doctrine of Chances*

Version zh25a.doc at 18:36 on 13 July, 2001

tions on probabilities merit our assent. Indeed, it is notoriously the case that, without a tangible yardstick of this kind, some remarkably fallacious propositions concerning probabilities can be made to appear mathematically reasonable, a trap into which many have fallen[1]. Jeffreys, for example, points out that the modern definition of expectation can lead to results which, from the viewpoint of common sense, are sheer nonsense: *e.g.* the mathematically 'expected value' from a trial can be a value which cannot possibly occur in practice. In Bayes' terms, however, no sane person would give any value whatsoever to the expectation of such a result, and Bayes, throughout the demonstrations of the Propositions, repeatedly appeals to an assumed self-evidence of propositions concerning financial prudence in order to seal the argument. From a purist mathematical point of view, this appeal may be questionable: yet who is to say that Bayes' concern is with a matter of 'pure mathematics' ? On the contrary, the whole object of the essay is to discuss what is essentially a matter of natural philosophy concerning inference from experiment[2]. It is interesting therefore to note the contrast with the view of Ramsey[3] that:- *The laws of probability are ... an extension to partial beliefs of formal logic. ... based ... on the idea of mathematical expectation*. Ramsey also rejects money as a measure because of its alleged diminishing marginal utility[4]. He then resorts to a concept of 'values' which can be manipulated in accordance with certain axioms, prior to which he has nominally discarded the assumption that values are additive[5]. We are however extremely doubtful as to the validity and consistency with which he has carried this through, for he also introduces notions of *worlds*, and *values of worlds*, designated α and β, such that α being preferable to β corresponds with the relationship 'the value of α is greater than the value of β'. This implies therefore that we could make some change in the world-state β such as to make it identical with world-state α and it is hard therefore to see why the making of such a change should not be regarded as precisely equivalent to an operation of 'adding value' exactly as with money.

---

[1] cf Keynes (1921) Ch 30 and particularly the statement:-*'we will endeavour to discredit the mathematical charlatanry by which, for a hundred years past, the basis of theoretical statistics has been greatly undermined'.* (p367).

[2] cf Hacking (1975), p93: *'The doctrine of chances is applied mathematics arising from vulgar practice'.*

[3] Ramsey (1931), p183

[4] Many measuring devices suffer diminishing utility outside certain magnitudes but are perfectly useful within their designed limits. A glass thermometer has a zero marginal utility above the temperature at which it explodes. An excellent and concise discussion of utility is given by Savage (1954), pp 91-104

[5] Ramsey (1931), p 176

However, as we remarked earlier, another difficulty with Bayes' Definition 5 is that it bears no obvious or direct relationship to the ways in which we normally use the word 'probability', and Bayes actually treats his definition, not as a definition of 'probability in itself', but as a definition of the manner in which a known value of a probability ought to be used in certain types of situation. Evidence that Bayes took this approach with some deliberation is provided by Richard Price in the covering letter:- *Mr Bayes ..... has also made an apology for the peculiar definition he has given of the word Chance or Probability . His design herein was to cut off all dispute about the meaning of the word, which in common language is used in different senses by persons of different opinions, and according as it is applied to past or future facts. But whatever senses it may have, all (he observes) will allow that an expectation depending on the truth of any past fact, or the happening of any future event, ought to be estimated so much the more valuable as the fact is more likely to be true, or the event more likely to happen. Instead therefore of the proper sense of the word probability, he has given that which all will allow to be its proper measure in every case where the word is used.*[1]

Clearly, therefore, Bayes is only concerned with probabilities which are, at least in principle, numerically comparable. We may infer from other sources that philosophers, their concepts, and *a fortiori* their disputations, were not to his taste[2]. Hence we may guess that the concept of non-comparable probabilities propounded by Keynes[3], would have held little interest for Bayes, and would certainly have had no rôle in the essay. Thus, while we may agree with Keynes that many probabilities may be neither enumerable, nor relatively comparable, there is no place here for a digression into that argument. Bayes, it would seem, decided to treat *probability* as a primitive and essentially numeric concept, requiring no further definition, and then to define a rule of how we ought to use the numeric value of a probability in the type of financial operation on which his argument is based.

Hence, from a philosophical point of view, Bayes' approach is a remarkable and early expression of a pragmatic and positivist attitude: he phlegmatically 'gets on with the job' leaving others to waste their time in disputes which can be never resolved. Although such attitudes may now be common among people who equate the meaning of a term with its operational use, it is very much to be doubted that such attitudes were common in the time of Bayes.

---

[1] p375 in the original manuscript. Reprinted in Ch. 2 above

[2] cf Barnard (1958)

[3] Keynes (1921), Ch 3

Version zh25a.doc at 18:36 on 13 July, 2001

Fisher[1] also comments pointedly on Bayes' understanding of 'probability' seeming convinced, and perhaps excessively concerned to assert that Bayes' view is essentially 'frequentist', being based on the ratio of actual occurrences of an event to the number of trials in which it could possibly have occurred. Bayes however does not make this equation of meaning, and we can find nothing in the essay to suggest that Bayes believed the meaning of probability to be constrained in this way. Clearly, Bayes' view of probability includes those probabilities which can be equated to occurrence-ratios, for it is at the estimation of exactly this type of probability that the essay is directed; but had his understanding of probability been confined in this way, he could easily have said so: yet he did not, and his definition, (or perhaps we could better say 'rule'), has the countervailing advantage that it can be used in unique situations where there is no possibility of any repetition of the trial, yet where we are able to form an estimate of the probability in question and are required to take corresponding monetary action, *e.g.* for investment or insurance purposes, or simply as a gamble.

A much later work which also takes expectation as fundamental, is that of Whittle[2] but it seems to us misleading[3] to equate Whittle's approach with that of Bayes, for Whittle quite explicitly bases his concept of Expectation on a frequentist model. In certain cases, we can apply Bayes' general principle to multiple trials of dice-like events, exactly as does Bayes himself, and in those cases it seems entirely reasonable to expect the observed frequencies of occurrence to be distributed about the true value of the probability. There are however in real life many cases of a quite different type where we are, for example, uncertain about the precise state of a system or about the dynamics of a trial, but our knowledge is sufficient for us rationally to assign different probabilities to different outcomes and therefore to evaluate, in Bayes' sense, the expectation. Yet the trial may be of a kind which is entirely deterministic, and although we may be totally uncertain, prior to the first such trial, as to the outcome, we may equally know that, however many times it may be repeated, the result will be invariably the same: the essence of the problem being in our uncertainty as to the initial state and dynamics of the trial. Real life, presents us with innumerable instances of this kind, and with many others, which are similar in terms of the nature of our uncertainty, but where the trial is unique and there can be no repetition.

---

[1] Fisher (1973), p14

[2] Whittle (1970)

[3] cf Hacking (1975), p 97

The differences between these types of trial are also important for the logic of Bayes' general argument which is critically dependent upon the concept of the same event happening in a probablistic manner in independent trials. As we saw in Chapter 3, the concept of an event being 'the same' in independent trials seems to require an identical specification governing each trial. Hence, if the outcome of the trial is fully determined by the specification, albeit we may know neither that this is the case nor what the outcome will be prior to the first trial, yet we observe after some number of trials that the outcome on each occasion has been the same, then we may entertain strong suspicion that the trial is indeed of the fully deterministic kind, and that $\mathcal{P}(E)$ is unity, or zero. We can, however, never be absolutely certain that this is the case if the sum of our relevant knowledge is confined to knowing the results of trials, and the whole aim of Bayes' endeavour is to enable us to calculate at each point a measure of our uncertainty.

In sum then, Bayes' rule for the use of known numeric probability defines the rational use of such information in uncertain real-life situations. Hence, although there can be no confining of Bayes' view of probability to the 'frequentist' or to the occurrence-ratio view, it is interesting to note that Bayes' view is directly applicable to the situation of a financier who has to take many investment decisions, each unique, and provides an entirely rational basis for such decisions, both individually and collectively. Certainly we could not express disapproval of a financier who behaved in such a way, and indeed, we may well believe that this is the way in which financiers, in making investment and insurance decisions, ought to behave. The 'St Petersburg' problem can then be overcome by adding further rules governing financial prudence, *e.g.* to the effect that one ought not to enter any deal where the effect of a loss would be ruinous[1].

It is remarkable that little attention has been given to Bayes' use of the word 'ought' in this context, with its strong connotations of ethical objectivity, and hence an implication that all rational persons, finding themselves in a given position, should, behave in the same way[2]. It is however notable that even Ramsey[3] is concerned to avoid ways of measuring degrees of belief which are variable from individual to individual[4]. Our own view is that

---

[1] cf Keynes (1921), pp 316 *et seq.*; Ramsey (1931), p172; Jeffreys, H. (1983), p31

[2] Earman (1992) p8 does however comment on this point.

[3] Ramsey (1931), p172

[4] A difficulty arises however if we are dealing with an 'unknown event' whose probability can be estimated, but only in a probablistic manner, by means of repeated trials: *i.e.* in such cases, how do we determine which is the value of probability that we ought to use, when there is a whole range of values from which we can choose ?

there is considerable danger in the use of the term 'subjective probability' and that it is not acceptable to apply this term to Bayes' approach. It may be true enough that many a punter at the races will be influenced in the placing of bets by an entirely personal and subjective feeling: but one has no right to expect that feeling to be shared by the next person. In Bayes' essay, the use of the word 'ought' clearly indicates that the estimation of probability ought to be objective, relative to the evidence and to rationally demonstrable assumptions. Thus everything in Bayes' argument is in line with the view of W.E.Johnson, Keynes, Jeffreys, and others, that the probability relationship is logical, objective and that strictly, we ought always to speak of a probability relative to specific evidence and assumptions[1].

It is also important to note that the assessment of probability in a given situation depends, not upon the truth about that situation, but upon what is believed to be true in the making of the assessment. Although Bayes gives no prominence to this point, we find that, in the *Corollary to Proposition 4,* he writes:- *'suppose that before we know whether the first event has happened, we find that the second event has happened; then purely on the basis of this information, we can infer that the first part of the trial has taken place but we do not know its outcome, and therefore we have no reason to value the expectation either greater or less than it was before*[2]'. It follows that, to support this deduction, we need to assume that the value of the expectation (and therefore of the corresponding probability) depends, not upon the 'true situation', but rather upon the facts which we know, or believe to be true, about the situation.

However, a difficult problem, not addressed by approaches to probability which are based on the concept of expectation, does concern objectivity. Putting aside the philosophical problem of defining exactly what we mean by 'objectivity', there is, in our time, the difficult practical problem of designing automata to 'act rationally' in probablistic situations. In certain types of situation, one may expect automata to accumulate experience based on repetitions of trials concerning an 'unknown event'. In general, however, automata are likely to be faced with series of events, each being in certain respects unique, and in other respects similar to other events which have been recorded. In such situations, it is hard indeed to see how an automaton could resort to a primitive notion of expectation within itself, and certainly with the outlook which is likely to be common to most readers of this book, we would expect automata to be programmed to look for relevant evidence,

---

[1] See also the discussion of this topic in Chapter 11, below

[2] See Chapter 2, above, also Chapter 3.

and, on that basis, to form a view of the spectrum of probabilities relevant to the situation[1]. Reflecting therefore upon the way in which people proceed when required to form probablistic judgements in 'unique' situations, we would suggest that a normal approach is to look for similarities with situations previously encountered and to form judgements of probability based upon the numbers of such similarities, their degrees of similarity, and our view as to the weights they should carry. In such cases, we are, in principle, quite close to the populations from which Whittle[2] derives expectations by means of 'indicator functions' and one can, in principle, envisage the programming of an automatic device to operate on such a basis. This approach also has the attraction that it allows for the fully deterministic trial in which the result will always be the same, and our uncertainty as to the result stems from our imperfect knowledge of the situation and the dynamics of the trial. In a totally unique situation, having no similarities with anything of which we have previous experience, we may be totally unable to form any prior estimate of probabilities and therefore, as we show in later chapters, the only non-irrational path will be the assumption of the uniform prior.

We now turn to the questions which arise concerning the addition of probabilities. In Proposition 1, Bayes deals with the additivity of the probabilities of mutually exclusive events. This is not demonstrated: it is simply asserted that, in the case of three mutually exclusive events, the values of the expectations $\$V_1$, $\$V_2$ *etc.* are additive, and thence, by Definition 5, that the probability of one or other of such events happening is $(p_1 + p_2 + p_3)$, where $p_1 = \$V_1 / \$N$, etc. It is remarkable therefore, in view of the great detail of Bayes justifications of later points, many of which we could today accept as directly obvious, that he makes no attempt to justify the equation which he makes between the sum of the values of the individual expectations, and the value of a joint expectation dependent on mutually exclusive events. Of many possible reasons for this omission, one is that Bayes considered the assertion so obvious as to need no demonstration; another possible reason is that, when he died, he had not fully revised the essay: and indeed, there are other signs that the essay, as we have it, is a preliminary joining together of sections, written at different times, and needing further work to make it complete and fully coherent. Additivity, however, seems to have given trouble to many writers: we see later how Ramsey attempted, (and in our view failed), to avoid an assumption of additivity: Jeffreys[3] makes no at-

[1] cf. Popper (1973) p.43.
[2] Whittle (1970), p28
[3] Jeffreys, H. (1983), p19

tempt to give a logical justification of additivity in the case of probabilities but claims it is merely a convention.

With Bayes' approach, however, of using a linear mapping between probability and financial expectation as a fundamental concept, it is possible to show that the additivity of the probabilities of mutually exclusive events is a consequence of the amenability of expectation to additive union, and conversely to subtractive partitioning. This can be illustrated by the following example:- We take, in the first instance, and as Bayes does in the Corollary to Prop.1, a pair of complementary events, (in the terminology of Bayes' Definition 2, 'contrary' events), these being alternative and exclusive outcomes possible from a single trial, where one person, let us call her Sarah, has an expectation of winning a prize $\$N$ dependent upon the happening of the event, and her expectation is valued at $\$V_1$. Another person, Jack, has a contrary and complementary expectation, valued at $\$V_2$, dependent upon the not-happening of the event. Clearly in such a situation:-

$$\$V_1 + \$V_2 = \$N \qquad (6\text{-}1)$$

We now extend the situation to a second trial and two further persons, Sue and Sam who have expectations from the second trial if, in the first trial, Sarah wins. The expectations of Sue and Sam are however contrary and complementary, *i.e.* if Sarah wins, then Sue has an expectation valued at $\$V_{11}$, and Sam has an expectation valued at $\$V_{12}$. Clearly in this case:-

$$\$V_{11} + \$V_{12} = \$V_1 \qquad (6\text{-}2)$$

and

$$\$V_{11} + \$V_{12} + \$V_2 = \$N \qquad (6\text{-}3)$$

whence following Definition 5, we find the corresponding probability values by dividing each term in (5-3) by $\$N$ giving:-

$$p_{11} + p_{12} + p_2 = 1 \qquad (6\text{-}4)$$

This result is however particular to the type of trial illustrated. To achieve the result which we need for Proposition 1, we have to show good reasons for believing an assertion that the joint value of a set of mutually exclusive expectations is invariant under operations of partition and combination[1]. For it is obvious that this is not true of all value-objects: life provides many examples, a precious vase, for example, where the value of the original union greatly exceeds the joint value of the parts after a partition. And there are converse cases where the joint value of separate parts exceeds that of their combination: as in the case of a feast at which the guests eat the

[1] *'if not to force the assent of others by a strict demonstration, (then) at least to the satisfaction of the Enquirer'* de Moivre, (1756), p254.

courses of the meal in succession, in contrast to a single course in which all is mixed together[1].

Continuing therefore the illustration, we introduce a further person, Sally, whose position entitles her to receive whatever gains are received by Sue and Sam, and whose expectation is therefore equal to that of Sarah: *i.e.* it is realistic and acceptable to assume that, if it is valid to partition an expectation into sub-expectations, each depending upon the happening of one of a set of mutually exclusive events, then it is conversely valid to combine a set of sub-expectations into a greater expectation depending upon the happening of any one of the corresponding events. If this is granted, then it follows by Definition 5 that if p_{X1} , p_{X2} , p_{X3} *etc.* are the probabilities of mutually exclusive events e_{X1} , e_{X2} , e_{X3} *etc.,* then the probability of the event e_X, which we define as an event in which one or other of the subsidiary events occurs, is given by:-

$$p_X = p_{X1} + p_{X2} + p_{X3} + \ldots . \text{ etc.} \qquad (6\text{-}5)$$

and Proposition 1 is demonstrated[2].

We now discuss a minor, but intriguing hitch which arises when we compare Bayes' Problem with the rule which he gives for the computation of a probability. In the problem, he is concerned with the chance that a given hypothesis is correct. In the latter case he gives us a rule by which we should determine the probability of an event. It would therefore be stretching the meaning of words too far to argue that these two concepts can be equated by some expedient such as defining a case in which an hypothesis is true as being an event. It seems therefore, that we have here a point on which Bayes requires some posthumous assistance, *i.e.* we have to find a way of converting the merely probablistic, and therefore only arbitrarily decidable, hypothesis which is the subject of the problem, into a form which is objectively decidable by inspection of agreed evidence as to whether a defined event has happened, or has not happened. That is, merely probable results are not acceptable, nor, we feel, are degrees of belief as suggested by Earman[3].

To deal with this matter, we first re-formulate the problem, slightly, as follows:- *Given that an unknown event has happened m times and failed n-m*

---

[1] cf Jeffreys, H. (1983), p32

[2] It is salutary to compare the logic which governs a joint expectation dependent upon the happening of any one of a number of *mutually exclusive* events, with that which governs a joint expectation in the case of *independent* events, where the value of the joint expectation is also equal to the sum of the parts.

[3] Earman (1992) pp8 et al.

times in n trials : Required the chance that the probability, p, of its happening in a single trial lies somewhere between any two values, α and β, that can be named. We then translate the problem into pragmatic and determinate terms as follows:- *Given the number of times m that an unknown event has happened in a set of n independent, identical trials, what is the probability that the event will happen not less than u times and not more than v times in a further set of s identical, independent trials?* However, if we read probability in this formulation in its normal sense, it will seem that this formulation fails to meet the stated objective. To achieve that objective it is necessary to substitute for *probability* Bayes rule for the computation of probability *viz:- Given the number of times m that an unknown event has happened in a set of n independent, identical trials, what is the ratio at which the value of an expectation dependent upon an event of the same kind happening not less than u times and not more than v times in a further set of s identical, independent trials ought to be computed in relation to the value of the thing expected ?* This is not identical to Bayes' original problem but it does have the advantage that, given the results *m* and *n* from the first set of trials, followed by specification of logically compatible values of *u,v,s*[1], we can compute the required ratio, precisely. We can then carry out the further set of *s* trials in which we observe the number of times *t* on which the event happens. From this we can decide by inspection, whether the assertion *'u < t < v'* is true or false. Hence, although it is always possible that a true value of *p* that is less than *u/s* or is greater than *v/s* will be capable of giving a result in the specified range, we can make the probability of this as small as we please by making *s* as large as we wish[2]. Conversely, it is always possible that the true value of *p* satisfies the assertion *'u/s < p < v/s'* but that, in the second set of trials, *t* turns out to be either less than *u* or greater than *v*. Again, however, we can make the probability of this being the case as small as we please by making *s* as large as we wish.

Although this approach is, we believe, logically sound, it does not lead directly to a simple wager. If the reward for predicting the outcome of the second set correctly in terms of *u,v* is fixed, then one could be always certain of success by selecting *u = 0* and *v = s* and the bookmaker would then rightly insist upon a stake being paid which was equal to the total payable on conclusion of the trial. If however the client selects less-certain values for *u* and *v*, then a more conventional kind of wager becomes possible. Thus, despite the fact that our modified specification of Bayes' Problem leads to a

[1] Values which entail contradictions are not acceptable.

[2] The quantification of this probability is discussed further in Ch 11 below.

somewhat unusual form of wager, it is nonetheless viable. Hence we conclude that, from a pragmatic point of view, Bayes Problem and his rule for the computation of a probability are valid and acceptable, if not, perhaps, to the most stringent formal analyst, then at least to those who have to form conclusions in practical situations, and act accordingly.

A final matter which we have to address in this chapter is the need to deal with 'probability' as a 'degree of rational belief'; a view which receives no explicit mention in Bayes' essay. It is our view, however, while one can neither prove the correspondence analytically, nor test it empirically, the semantic correspondence is close to being tautological, or at least axiomatic, and the numerical correspondence is exact. That is, given evidence and assumptions k, in relation to which the degree to which it is rational to believe an assertion A to be true, is some function $\mathcal{F}(A \mid k)$, then it seems self-evidently true, referring back to equation *(3-1)* and Bayes' argument in support of the corollary to *Proposition 4*, that to value an expectation dependent upon the truth of the assumption as anything other than

$$\$V = \$N \times \mathcal{F}(A \mid k) \qquad (6\text{-}6)$$

would be self-contradictory. That is, the degree to which we rationally believe the assertion to be true is given by:-

$$\mathcal{F}(A \mid k) = \$V / \$N \qquad (6\text{-}7)$$

which is numerically identical with Bayes' measure of the probability that the assertion is true.

Implicit therefore in Bayes' reasoning are the assumptions that probability is objective, relative to the evidence, and represents the degree to which it is rational to believe the assertion. The probability is objective in the sense that every rational agent, given the same facts, assumptions and rules of inference will form the same estimated value for the probable truth of the assertion. A change in the probability can only be brought about by a change in the information or assumptions which are believed to be true. The degree of the agent's rational belief is measured by an expectation ratio which is numerically identical, and is, arguably, conceptually identical, with the probability as defined by Bayes.

Chapter 7

Critics and Defenders

In this chapter, we discuss some important opinions about Bayes' essay which have been expressed by various scholars, and which we do not address elsewhere. However, in attempting to assess the opinions which have been expressed over the past 250 years, two notable difficulties are encountered. First, many criticisms have been levelled against Bayes, relating to things which he neither says nor implies, but which do concern valid and important further questions which stem naturally from the essay. Other criticisms are framed in terms that we have found deeply obscure, defying rational analysis and, in some cases, verging upon the mystical. To deal with all the published opinions would be a large task and would take us down many paths. We therefore concentrate, in this chapter, on comments which directly concern Bayes' essay, and on matters which are particularly relevant to our own exploration.

Concerning therefore the content of the Essay, study of the critics, favourable and unfavourable, must begin with Richard Price, whose reporting of Bayes' views on 'The Postulate' we have noted in Chapter 5. But also, Price went on, in his covering letter, to claim that:- *'[T]he problem now mentioned is by no means a curious speculation .... , but necessary to be solved in order to assure foundation for all our reasonings concerning past facts, and what is likely to be hereafter. Common sense is indeed sufficient to shew us that, from the observation of what has in former instances been the consequence of a certain cause or action, one may make a judgement what is likely to be the consequence of it another time, and that the larger number of experiments we have to support a conclusion, so much the more reason we have to take it for granted. But it is certain that we cannot determine, at least not to any nicety, in what degree repeated experiments confirm a conclusion, without the particular discussion of the beforementioned problem; which, therefore, is necessary to be considered by any one who would give a clear account of the strength of analogical or inductive reasoning; concerning which, at present, we seem to know little more than that it does sometimes in fact convince us, and at other times not; and that, as it is the means of acquainting us with many truths of which otherwise we must*

Version zh25a.doc at 18:36 on 13 July, 2001

have been ignorant; so it is, in all probability, the source of many errors, which perhaps might in some measure be avoided, if the force that this sort of reasoning ought to have with us were more distinctly and clearly understood. After a brief digression to comment on certain aspects of de Moivre's work, Price continues:- *The purpose, is to shew what reason we have for believing that there are in the constitution of things, fixed laws according to which events happen, and that, therefore, the frame of the world must be the effect of the wisdom and power of an intelligent cause; and thus to confirm the argument taken from final causes for the existence of the Deity..... (and) ..... it will be easy to see that the problem solved in this essay is more directly applicable to this purpose; for it shews us, with distinctness and precision, in every case of any particular order or recurrency of events, what reason there is to think that such recurrency or order is derived from stable causes or regulations in nature, and not from any of the irregularities of chance.*[1]

Price has therefore here presented claims for what is achieved in the Essay, which go far beyond anything claimed by Bayes himself, and far beyond the terms of reference established by Bayes in the clear and precisely-bounded definition of the problem with which he was concerned. There are therefore solid grounds on which one could take exception to the claims adduced by Price and whose sincerity or motives one might have thought were rendered questionable by the fact that he could put forward an argument so riddled with seeming *'non sequiturs'* at a time when Hume had already shown the extreme difficulty, if not impossibility, of justifying inductive generalisations by deductive logic[2]. It is however clear from other sources that Price was a man of ability and integrity[3]. He was known personally to Hume, and their positions on matters of probability and induction were remarkably close, while Hume appears to have had no ready answers to certain points made in support of the religious orthodoxy of the time by Price[4]. We have also to bear in mind that the above quotation is taken from a private letter written by Price to John Canton, albeit the letter was later used as an Introduction to Bayes' essay. Price is therefore well within his rights to express such views in such a communication, where he may have thought it prudent to suggest a politico-religious agenda which was not even mentioned by Bayes. Nevertheless, it does appear that, in doing so, Price

---

[1] The original text is reprinted in Ch. 2, above. The version given here is a mild paraphrase with correction of occasional peculiarities in the spelling and punctuation of the original.

[2] Hume (1748) Sect IV, Part II.

[3] Thomas, D.O. (1977), Gillies (1987) , Bernstein (1996)

[4] Mossner (1954)

Version zh25a.doc at 18:36 on 13 July, 2001

has put aside Bayes' clear, albeit un-stated assumption, that throughout the series of the observations, there is a fixed and stable probability. Nowhere does Bayes even hint at the possibility of proving such stability by inference from observation. For, if the probability of success in any given trial were to vary in any way, Bayes' argument would collapse. In his discussion of the Experiment, Bayes takes for granted the stable probability which Price would seem to claim could be inferred from the results. But we should not under-estimate the significance of the point being made by Price. For, by means of Bayes' result, it became possible for the first time to address empirical questions concerning the statistical stability of probablistic phenomena. More serious perhaps were the adverse effects of the connection which Price implied between Bayes' result and the massive questions, philosophical and empirical, which concern probability and causation. This was possibly the first step towards a path of reckless mathematical fantasy in which connections with reality became remote and weak[1].

Serious analytic criticism of Bayes' approach seems however not to have appeared until some hundred years after the publication of the Essay. One of the earlier of these Victorian critics was George Boole[2], in Chapter 20 of *The Laws of Thought*[3], although he, remarkably, makes no direct mention of Bayes. This omission may have occurred because Boole may not have had direct access to the relevant volume of the *Philosophical Transactions*, and may have been working therefore upon hearsay and second-hand evidence. This surmise is perhaps supported by the fact that, in the discussion of Michell's paper[4], the name is mis-spelt as *Mitchell* and the problem discussed has only a loose correspondence with Michell's published paper. Chapter 20 of Boole's work is however entitled *'Problems Relating to the Connexion of Causes and Effects'* and the greater part of that chapter is devoted to the calculation of *a priori* probabilities when the values of all necessary parameters are given in the formulation of the problems. Some of these problems are decidedly complex and Boole concludes the discussion of them thus[5]:- *It is remarkable that the solutions of the previous problems are void of any arbitrary element. We should scarcely, from the appearance of the data, have anticipated such a result. It is however to be observed, that in*

---

[1] *cf. 'The mathematical charlatanry by which, for a hundred years past, the basis of theoretical statistics has been greatly undermined'.* Keynes (1921) p367.

[2] George Boole is a salutary example of what could be achieved by self-education in the nineteenth century.

[3] Boole G. (1854) .

[4] Michell (1767).

[5] Boole (1854) pp 379 *et seq.*, here paraphrased slightly.

Version zh25a.doc at 18:36 on 13 July, 2001

all those problems, the probabilities of the causes involved are supposed to be known 'a priori'. In the absence of this knowledge, it seems that arbitrary constants would necessarily appear in the final solution, and some confirmation of this is provided by a class of problems to which considerable attention has been directed, and which are now briefly considered.

At this point, Boole turns to the type of problem addressed by Michell, Laplace[1], de Morgan, and others who were concerned to know if it were possible to determine the chances that phenomena such the clustering of stars in the constellation of the Pleiades, the co-alignments of the planes of circulation of the planets around the sun, or the co-alignments of the axes of polarisation in rock crystal *etc.* were attributable to random dispositions, as opposed to particular causes. Our prime concern at this present point is more, however, with Boole's attack on attempts to assign probabilities to hypotheses in the absence of prior information. In the course of this, he lays down the lines of what were to become a standard reason for rejecting Bayes' solution[2]. The attack opens with a broadside: *'The general problem, in whatsoever form it may be presented, admits only of an indefinite solution'.* He continues:- *'Let 'x' represent the proposed hypothesis, 'y' a phenomenon which might occur as one of its possible consequences, and whose calculated probability, on the assumption of the truth of the hypothesis, is 'p', and let it be required to determine the probability that if the phenomenon 'y' is observed, the hypothesis 'x' is true. The very data of this problem cannot be expressed without the introduction of an arbitrary element. We can only write*

$$Prob.\ x = a$$

$$Prob.\ xy = ap$$

'a' being perfectly arbitrary, except that it must fall within the limits *0 - 1* inclusive. If then *'P'* represents the conditional probability sought, we have

$$P = \frac{Prob.\ xy}{Prob.\ y} = \frac{ap}{Prob.\ y} \qquad (7\text{-}1)$$

Boole's wording is not, however, as clear as we need in order to effect a comparison with Bayes' analysis. For a phenomenon will not, in general, be observed as a consequence of an hypothesis but as a consequence of an event. The hypothesis will generally take a form such as an assertion that, on a given occasion, a defined event was the cause of an observed phenome-

[1] Laplace (1820) p 276
[2] Boole (1854) p 381

non. We therefore re-phrase the problem posed by Boole, using the notation and definitions as in our previous discussion of Bayes' analysis:-

h : an hypothesis that an event E_1 occurred on the occasion of a particular trial.
E_1 : an event , the occurrence of which creates a probability r_2 that a further event E_2 will occur[1].
p_1 : the probability that the event E_1 will occur in any given trial of the relevant kind.
r_2 : the conditional probability that the event E_2 will occur in a particular trial, if E_1 has already occurred in that trial.
p_2 : the unconditional probability, i.e. without regard to E_1, that in any trial of the relevant kind, the event E_2 will occur.
$\mathcal{P}(h \mid E_2)$: the probability that the hypothesis h was true of a particular trial in which the event E_2 occurred.
This therefore gives us, by substitution into (*7-1*):-

$$\mathcal{P} \quad = \quad \frac{\mathcal{P}(E_1 \wedge E_2)}{\mathcal{P}(E_2)} \quad = \quad \frac{p_1 r_2}{p_2} \qquad (7\text{-}2)$$

Boole then analyses this equation by means of his calculus of logic and concludes that:-

$$\mathcal{P}(h \mid E_2) \quad = \quad \frac{p_1 r_2}{p_1 r_2 + c\,(1 - p_1)} \qquad (7\text{-}3)$$

where *c* is the probability that the event E_2 will occur in a trial in which the event E_1 has failed to occur. Boole also asserts that p_1 and *c* are arbitrary constants, but he does not analyse these terms in relation to Bayes' problem, where:

E_1 is the event that $x_1 < P_m < x_2$
p_1 is the prior probability that the assertion '$x_1 < P_m < x_2$' will be true of any given trial.
E_2 is the outcome of a trial-set of size n in which the event M occurs on m occasions and fails to occur on n-m occasions.
c a parameter in (7-3), represents the probability that the event M will occur on m occasions in n trials under the condition that the assertion '$x_1 < P_m < x_2$' is false.

We know, however, from Chapter 5, that the assumption of the uniform prior is a necessary condition for rational quantitative inference concerning the probable value of $\mathcal{P}(m \mid n)$, and we know from the work of Murray and Molina, that, given a uniform prior, the unconditional prob-

[1] The expression *'will occur'* includes, for brevity *'or will be found to have occurred'*.

ability of E_2 is a function only of n and is identical for all values of m; we therefore denote this number as K_n. Further, the unconditional probability of E_2 can be partitioned into the mutually exclusive and exhaustive conditional cases:-

i. when the assertion '$x_1 < P_m < x_2$' is true
ii. when the assertion '$x_1 < P_m < x_2$' is false.

It follows that:-

the probability that '$x_1 < P_m < x_2$' is true,
the probability that '$x_1 < P_m < x_2$' is false
and that E_2 will not occur $= (1 - p_1)c$ *(7-4)*

thus giving the identities

$$\mathcal{P}(E_2) = K_n = p_1 r_2 + (1 - p_1)c \qquad (7\text{-}5)$$

Boole's expression is therefore exactly equivalent to that of Bayes. Hence, if our view of Bayes' result is valid, Boole's rejection must either collapse or be as valid as a refusal to accept that Bayes' Experiment tells us anything about the probable value of P_m.

The views expressed by Murray in 1930, Molina in 1931 are of interest in that, while they indicate reservations similar to those of Boole, they are essentially supportive, rather than destructive in their attitudes towards Bayes' argument. In Murray's case, the analysis shows that a uniform prior distribution of probability over the possible values of P_m is a necessary condition for the prior probability of each possible number of successes, *i.e.* m, from n trials, to have the constant value $1/(n+1)$ for all legitimate values[1] of m and of P_m. Murray interprets this result as showing that the assumption of all values of P_m being equally likely is equivalent to the assumption of any given legitimate number of successes in n trials being just as likely as any other legitimate number[2]. This view is also supported by Molina who writes:- *' if any outcome of throws not yet made is as likely as any other, then any value of x is as likely as any other. This ..... theorem was submitted to Dr F.H.Murray, who obtained an elegant proof[3]'* Molina, however, is cautious. In his summary he states:- *'Bayes' theorem is the answer to a special case of the general problem of causes. The special case postulates that the 'a priori' .... probabilities ..... are equal. ..... To justify this ... Bayes takes the attitude that a state of total ignorance regarding the causes of an observed event is equivalent to the same .... igno-*

[1] The number of successes in n trials can be $0,1,2 \ldots n$, *i.e.* $n+1$.
[2] Murray (1930) p129, footnote.
[3] Molina (1931) pp34-5

rance as to what the result will be if the trial .... has not yet been made. Laplace, Poincaré and Edgeworth have shown that the 'a priori' existence function w(x) , which appears in the Laplacian generalisation of Bayes' theorem, is of negligible importance when the numbers N and T are large. Therefore, when this condition holds, one need not hesitate to use Bayes' restricted formula for the solution of a problem of causes'. Stigler, whose views[1] were particularly instrumental in motivating our own work, takes a much more positive attitude, as we discussed in Chapter 5.

We now come to the criticisms which have been levelled by Shafer at Bayes' treatment of the basic theorems of probability[2]. Shafer argues that:- '*the temporal order of Bayes' 'subsequent events' is crucial to the validity of the argument*[3]*'.* and that '*the ..... argument which leads to Bayes' fifth proposition does not stand up to scrutiny*[4]*'.* Although Shafer directs his open attack against Bayes, he also hints at a deeper and very important implication:- '*Our current ways of thinking about conditional probability seem very deeply entrenched, and the thesis that these ways of thinking need to be or even can be changed will leave many readers incredulous. ..... Our current ideas are not as old and may not be as permanent as their constant repetition in textbook after textbook makes them seem*[5]*'.* Unfortunately, rather than concentrate on this deeper issue, Shafer switches repeatedly into making contentious assertions against Bayes. This makes it extremely difficult to follow either argument or to feel that either is convincing. That is a pity, for Shafer's use of 'rooted trees' seems to be potentially useful in areas of Bayes' essay which are certainly not easy to follow and where there may well be subtleties which could benefit from some clear new light.

The gist of Shafer's argument against Bayes seems to be contained in the following three consecutive paragraphs [6]:-

- *Why does Bayes, in the statement of his third proposition, refer to the two events he is considering as 'subsequent'?*
- *Imagine a situation in which it is not known beforehand which of two events A and B will happen (or fail) first. In such a situation we cannot say beforehand what the probability of B will be immediately after A hap-*

---

[1] Stigler (1982), (1983), (1986).
[2] Shafer (1982).
[3] Shafer (1982) Section 1 para 5
[4] Shafer (1982) p1075 - Synopsis.
[5] Shafer (1982) Section 1 para 7
[6] Shafer (1982) Section 3, paras 2,3,4.

pens; that probability will be one if B has already happened by then, zero if B has already failed by then, and something in between if B has not yet happened or failed. Saying that B is subsequent to A may be an attempt on Bayes's part to resolve this ambiguity.

- *But the attempt is not fully successful. Once we admit that the timing of events may be contingent, we must guard not only against the possibility that B itself may or may not have happened by the time A happens but also against the possibility that other events affecting the probability of B may or may not have happened by then.*

However, although Shafer has possibly identified a point of some interest in terms of the effects of temporal order on probably-causative relationships in the physical world, we can see no justification for imposing an assumption of a probably-causative structure on Bayes' argument, nor for assuming that such a structure is a necessary part of all probablistic trials in which events are ordered in time. For, although the general issue raised by Shafer is much more important than the question as to why Bayes chose to write of *'subsequent events'*, a simple, and possibly correct explanation of Bayes' phraseology, is that he needed this phrase to picture the possibility of *correlation* between the events in question, and to make the argument readable. As we have noted previously, Bayes' essay is written in that noble tradition of philosophy, going back at least to Plato, where the aim is to establish a lively communion with the reader, and decidedly not to numb one's senses with incomprehensible sophistry. It has also to born in mind, when attacking Bayes, that it was Price, not Bayes, who sent the unfinished essay for publication. Heaven knows what improvements Bayes might have made, had he been spared a while longer.

Shafer seems however to have rather seriously mis-read, or misunderstood, what Bayes wrote at a number of points. A notable instance is where Shafer writes [1]:- *Bayes puts the question of when events happen at the very base of his concepts. In Definition 5, for example, he makes it clear that the 'value of a thing' depends on what events have happened at the time the value is to be computed.* The truth is that Definition 5 reads:- *The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening*. We find it hard indeed to interpret such words in any sense other than that *the value of the thing expected* is utterly invariant with respect to the happening of events. In fact, the sequence of events which Bayes invokes in support of propositions 3, 4 and 5 is an

[1] Shafer (1982) Section 2 para 2.

artificial, but entirely rational, construction. The aim is to satisfy a reasonably intelligent and objective reader that the propositions are valid. The spirit of Bayes' demonstration is entirely in line with that of Euclid: the temporal sequences which Bayes uses are as irrelevant to the validity of the argument as are the temporal sequences in a construction used to illustrate a geometrical proof. Bayes' examples are merely typical of the artificial challenges enjoyed by the gambling fraternity. The order in which dice are rolled, or in which cards are drawn from a pack, is utterly irrelevant if they are concealed from us. What may matter, is the order in which they are made known to us, or the order in which we choose to inspect them, and the wagers that we may choose to make at these various points. All that is needed to make Bayes' argument viable on this score, is the reasonable acceptance of his Definitions, and a degree of correlation between the events.

Indeed, as we have shown in Chapter 5, the temporal ordering in the Experiment, is not essential to the argument, which can equally be couched in terms of a single ball which is thrown $n + 1$ times. In real life there are countless cases which are analogous to the generation of a governing parameter by the throwing of the first ball and the series of trials conducted with the second ball. It is however of the essence of the Experiment as defined by Bayes that all other effects are constant in their influence, so that the two parts of the trial form a close-coupled pair into which nothing further can intrude. Although Shafer specifically excludes consideration of Part II of the essay, (*i.e.* the Experiment), from his paper, we find nothing there to raise doubts concerning the validity of the experimental set-up as an example that falls within the scope of Bayes' demonstrated propositions.

Turning next to Keynes[1] and Fine[2], we here encounter critics of outstanding calibre, whose views have much in common and are cogently expressed. Keynes' criticisms are expressed mainly in the superbly constructed *Treatise on Probability*[3], the final part of which provides a detailed philosophical critique of statistical inference[4]. Fortunately, the critique does not depend to any significant degree upon Keynes' own view of probability: a view utterly different from that adopted by Bayes. However, the criticisms in Keynes' treatise occupy several chapters and we therefore have to be highly selective in our reflection of them. Essentially, two aspects seem to be important. The first concerns the validity of Bernoulli's Theorem in

[1] Keynes (1921)
[2] Fine (1973)
[3] Keynes (1921).
[4] op. cit. pp325 *et seq.*

practical situations; the second concerns the inexorable limitations on our ability to compute the probability of an hypothesis concerning 'unknown events'. Keynes presents a generalised definition of the problem addressed by Bernoulli, such that if a defined event can occur on a series of occasions, and the probability of occurrence on each occasion is known, we are required to deduce the probable numbers of occasions on which the event is expected within the series. *'Yet',* he says, *'the theorem exhibits algebraical rather than logical insight. .... and requires conditions, before it can be legitimately applied, of which the fulfilment is rather the exception than the rule*[1]*'.* Keynes is particularly concerned with the implicit assumption that, if an event has occurred on every one of *r - 1* occasions, we shall not change our view of its probability of occurrence on the *r'th* occasion, whatever the magnitude of *r*. Interestingly, however, Popper[2], who frequently refers to Keynes, pays no special attention to Keynes' difficulty in this assumed invariance of belief as a trial progresses. Popper's concern is, however, with issues which are very much wider than those raised by Bayes' essay, and are indeed much wider even than those suggested by Richard Price

With regard to the inverse of Bernoulli's problem, Keynes uses a form which is more general than that used by Bayes. It is framed in terms of propositional functions $A(x)$ and $B(x)$ which can both be true of a given argument x. If we are then told that $B(x)$ is true for a certain proportion of the values of x for which $A(x)$ is true, we require to know what is the probability, if $A(x)$ is true for a further value, $x = \alpha$, that $B(x)$ will also be true ? Keynes does not produce a precise answer to this question. He points out that an assumption of a uniform prior for the distribution of the unknown probability[3] P_m, is easily shown to yield *m/n* as the most probable answer, but he does not point out that it is also, in general, a necessary condition[4].

Keynes also refuses to accept the postulate of the uniform prior in the general case of the 'unknown event'. but argues along the following lines: He deduces an expression closely analogous to *(5-18)* above but in which the possible values of P_m are confined to a finite subset of the positive rational fractions, whence, denoting the intervals between the fractions by δx, he makes an assertion[5] equivalent to:

---

[1] op. cit. p341

[2] Popper, (1972).

[3] We largely retain our own notation which differs from that used by Keynes.

[4] There is however an important special case which we discuss in Chapter 11.

[5] We believe there is a mis-print in the original which, in Keynes' notation reads *f(q)/h . f(q')/h . f(q)* but ought to read *f(q)/h . f'(q')/h . f(q)*

Version zh25a.doc at 18:36 on 13 July, 2001

$$\mathcal{P}(\ P_m = x\ ) \quad = \quad \frac{g(x,m,n)\, f_o(x)\delta x}{\int g(x,m,n)\, f_o(x)\ \delta x}$$

Keynes then says that he can see no reason to assume that all values of the term corresponding to $f_o(x)\delta x$ are *a priori* equal[1], but he goes on to conclude that *'as the number of instances is increased the probability, that m/n is in the neighbourhood of P_m, tends towards certainty ..... (and) we know that we can get as near certainty as we choose by a finite number of instances, but what this number is we do not know. This is not very satisfactory, but .... it accords with common sense[2]'.* Keynes does not however point out, nor does he investigate that fact that convergence of *m/n* to P_m will occur even if the *a priori* probability given to the true value of P_m is zero. That is, regardless of the assumed *a priori* distribution $f_o(x)$, as *n* increases, the ratio *m/n* 'converges in probability' on the true value of P_m, *i.e.* as $n \to \infty$, $\mathcal{P}(m/n \to P_m) \to 1$ [3]. This is a fact on which we shall later have cause for quite some reflection.

A little later, Keynes cites an example given by Professor Karl Pearson[4] in which Pearson argues, essentially on the basis of Bayes' result, that if a sample of 100 from a given population shows *10* people afflicted with a certain disease, then the number so afflicted in a second sample of *100* is as likely to fall inside the range of *8* to *14* (approximately), as it is to fall outside that range. Against this, Keynes argues that *'it does not seem reasonable upon general grounds that we should be able on so little evidence to reach so certain a conclusion. ..... The method is much too powerful ..... it invests any positive conclusion, which it is employed to support, with far too high a degree of probability. Indeed, this is so foolish ..... that to entertain it is discreditable'.* Keynes failed however to notice the paradox in the wording of his rebuke: for, if we consider a pair of complimentary intervals, it is impossible that we should produce over-estimates of both the respective probabilities!

We turn now to R.A.Fisher, who, among all Bayes' critics, was pre-eminent. For some *50* years prior to his death in *1962*, Fisher repeatedly put forward claims for his own doctrines, while constantly deploring, and often mis-representing, Bayes' solution to the *'Problem in the Doctrine of Chances'.* Fisher could never bring himself to accept that empirical knowl-

---

[1] *op.cit.* p 387

[2] We have substituted the symbols *m/n* for *q* and P_m for *q'*.

[3] Bernoulli (1713)

[4] Pearson (1907)

Version zh25a.doc at 18:36 on 13 July, 2001

edge might be accessible to us only at the price of an arbitrary assumption concerning the prior distribution of the probability in question. This was a view with which Keynes[1], and many others, had quite some sympathy, and it led Fisher to pursue, over many years, the concepts of fiducial probability and of likelihood. The concept of fiducial probability was directed at achieving a measure of probability, applicable to observable natural phenomena, which could be regarded as being, in some sense, absolute and free from arbitrary assumptions. This was despite the fact that as early as *1911*, Fisher was aware that the analysis of data always requires us to make arbitrary assumptions[2]. Many of those assumptions he chose to ignore, but following a nasty squabble with Pearson and others over the matter of a prior probability[3], Fisher became obsessionally concerned to avoid the assumption of the uniform prior. Thereafter, he made repeated claims for the ability of the fiducial argument to avoid the need for that assumption. Therefore, although nowadays the issue of fiducial probability is often regarded as a matter of only minor interest[4], the fact is that Fisher's claims in this area present a formidable challenge, not only to Bayes' analysis, but also to our own understanding. If Fisher was wrong, we need to know why he was wrong, not just in particular cases[5], but in principle. Merely to ignore the claims does not refute them. Nor can we hide behind the confusion which permeates Fisher's writings on the matter[6]. His reputation is still high; we believe he is widely read, and it is necessary to deal both with his comments on Bayes' achievement, (or, in plainer words than Fisher cared to employ, on Bayes' lack of achievement), and with Fisher's claims for his own success[7]. Further, the fiducial manner of reasoning presents a tempting, ever-present trap[8], into which anyone can fall.

Although we have found no document in which Fisher gives a clear definition of fiducial probability, and, indeed, the concept seems to have varied from paper to paper, the central thrust seems to be the 'logical inver-

---

[1] Keynes (1921) p389. See also Chapter 10 below.

[2] Fisher(1911)

[3] Fisher(1915), Soper *et al* (1916), Fisher (1921)

[4] See *e.g.* Bernardo and Smith, (1994), p458

[5] *e.g.* the 'non-identifiable sub-set'. See Zabell, (1992) .

[6] In our view, however, the lack of detail and explanation in Fisher's writings in this area, coupled to the fact that, in many cases, his arguments seem to be fallacious and his conclusions wrong, would make a detailed, step-by-step exposition impossible. As Keynes says of Laplace's notorious 'Principle of Indifference', *'it is not easy to give a lucid account of so confused a doctrine',* [Keynes (1921) p374].

[7] See Fisher (1911), (1930), (1962), and, in particular, (1956).

[8] Piattelli-Palmarini (1994) Chapters 4 and 5.

sion of a random variable[1]'. That is, if we are told there is a fixed probability underlying the occurrence of a random event[2], and we are told the value of that probability P_m, then Bernoulli's theorem tells us the distribution of probabilities over the various possible outcomes of an *(m,n)* trial. The 'logical inversion' sought by Fisher, attempts to argue from the result of such a trial to a distribution of probabilities over the possible values of P_m, in cases where we have no prior information about P_m, and without invoking Bayes' postulate of a uniform prior distribution. In Fisher's hands, this amounts, by and large, to an attempt to solve Bayes' problem without using Bayes' theorem. In words close to those used by Fisher, the essential argument is that, in certain cases, random samples of a population, can allow us to deduce the probability that a parameter, which governs the distribution of probabilities within that population, lies in any given range. By fiducial reasoning, Fisher asserted, this can be achieved, in certain cases, when nothing is known prior to the observations, and without assuming any prior distribution of probability over the parameter in question. He was, indeed, strongly of the view that, by using his approach it is possible to arrive at *'a complete specification of .... precision[3]'* , and that *'Exactly verifiable probability statements ..... can be assigned when the fiducial argument is available[4]'*. A stark example, to which we return later, concerns the median of a population. If we take a single sample, at random, then clearly the probability that its value is less than the median, is *0.5*. Can we not then assert that there is also a probability of *0.5* that the value of the median is greater than that of our sample ? The *prima facie* relationship of such questions to Bayes' problem is clear, but to produce a fair and thorough assessment of the fiducial argument is a gruelling task. For, while precision and clarity are central in Bayes, this cannot be said of Fisher's writings in this area[5]. Nevertheless, we shall find that the clarification of Fisher's arguments is an important step along our path.

The difficulties with Fisher arise from a number of causes and they are compounded by the manner in which, with the passing of the years, he addresses an ever wider range of aims. Unfortunately, Fisher provides little open guidance to make the reader aware of this fact, nor of which aim is being addressed at any particular point. Given such a complex of motivation and expression, it is impossible to be certain that a correct interpretation of Fisher's meaning has been, or ever can be achieved. For, particularly in the

---

1 Fisher (1945), p131

2 Please see Definitions.

3 Fisher (1956) p60

4 Fisher (1956) p70

5 *'No branch of statistical writing is more mystifying ..'.* Hacking (1965) p133.

Version zh25a.doc at 18:36 on 13 July, 2001

later writings, Fisher rarely explains his mathematical assertions. The reader is, it seems, expected to know, or to accept without question, many sophisticated algebraic results concerning probability distributions[1]. These assertions may, rightly, alarm anyone who is aware of the danger that, in dealing with probability, our algebra can, too easily, depart from the reality it is meant to describe[2]. The later writings abound also with imprecise allusions to authors and teachers, alleged by Fisher to hold views opposed to his own. We are too often, however, denied the references which would allow us to check what these people actually said. We are told that Boole's deprecation of *'supplying by hypothesis what is lacking in the data, points to an abuse very congenial to certain twentieth-century writers*[3]', yet we are given no advice as to where to find such writings, nor indeed, as to how we might avoid them. But two people are repeatedly named: Pearson and Neyman. They are rebuked, yet often when the alleged offence has only minor relevance to the point[4]. Then, with the further passage of time, Fisher seems to have been sadly motivated by an increasing range of 'sore points', as, for example, in the discussion which concludes his paper *'Some Examples of Bayes' Method*[5]*'*. Yet also, and apparently in the belief that he had bettered Bayes, Fisher seemed increasingly concerned to assert the superiority of his own achievements.

Despite these difficulties, we have formed the view that, in his attitude to Bayes' Essay, and to the general matter of inference from observations, Fisher's conscious motive was to assert a philosophical and political point concerning the objectivity of empirical science and the ability of the natural world to speak for itself *via* scientific observations made upon that world. Thus he seeks to achieve an 'objective' or 'absolute' solution to Bayes', and similar, problems. For, Fisher seemingly could not bring himself to entertain the possibility that the acquisition of empirical knowledge by observation of nature, might depend upon a purely personal and arbitrary decision to as-

---

[1] Quite to the contrary, he apologises, or purports to apologise, for *'the reiteration of simple and, it should be, obvious points'*. See Fisher (1962) p123. For comments on Fisher's reluctance, even at an early age, to explain his mathematical conclusions, see Box (1978) pp13-14.

[2] For a similar comment by De Morgan, see Fisher (1956) p29.

[3] Fisher (1956) pp23-4.

[4] But note also the following quotation from the Foreword to the combined volume in which Fisher (1956) was re-issued in 1990. This foreword was written by F.Yates, and tells us that:- *'Fisher ... was remarkably forbearing in his criticisms of Neyman in 'Scientific Inference'. For a fuller account of the conflict .... (see) .... (Yates') .... review of a biography of Neyman by Constance Reid (J.Roy.Stat.Soc.,A, Vol 147, 1984)'*. [See also Bennett (1990) pp. xxvi-vii].

[5] Fisher (1962)

sume a uniform prior. It is also, perhaps, important to note Fisher's remarks on freedom of thought, and on the integrity of the scientific method, which occur in the preface to *Statistical Methods and Scientific Inference*[1], for there can be little doubt that political concepts of 'correct thinking' can present an appalling threat to the integrity of scientific standards[2]. Perhaps, therefore, this threat may have also played a considerable part in evoking from Fisher the repeated claim to show cases in which observation, although subject to uncertainty, could, without *a priori* assumptions or axioms concerning empirical matters, yield precisely quantified probablistic knowledge[3] .

It is therefore in the tangle of these complexities that we have tried, first, to find the essential thrust of the fiducial argument and, second, to evaluate its validity. In this, we were however, further hampered by the many different views of previous authors about the point where the concept of fiducial probability first appears in Fisher's writings. Eventually, we decided that the essential point arises in the title of one of Fisher's earliest papers *'On an absolute criterion for fitting frequency curves*[4]*'*. In that paper, there is no open mention of Bayes, nor of any prior probability distribution, but the essence of Fisher's unease is clear in the opening words:- *'..... we are met at the outset by an arbitrariness which appears to invalidate any results we may obtain'*. The importance of this point in Fisher's writing is that it shows, first, his early concern about making an arbitrary or subjective choice and, second, it allows us to see the point of conception against a philosophical background in a period of still-solid Victorian assurance, before the *1914-18* war, and where mathematics must have appeared to many as a fortress of objectivity. And, if mathematics was such a fortress, then *a fortiori* so was empirical science based simply on observing and counting events. Yet a potent lesson of Bayes' essay is that, if the conclusions to be drawn from his primitive and fundamental experiment are inevitably dependent upon assumptions that we choose to adopt, then it is possible that the whole of quantitative science is, in that sense, subjective and arbitrary. Or is based upon an act of faith to which members of the scientific community generally subscribe. But to Fisher, such implications may well have been anathema. Hence, much though we may resent the

---

[1] Fisher (1956) p7

[2] It is possible that research into the impact, (if any), of Lysenko upon Fisher might be revealing in this connection. See also v.Plato (1994) p206 for comments on related aspects of 'Lysenkoism'.

[3] We have to let the reader decide the conundrum posed by the words *'precisely quantified probablistic knowledge'*.

[4] Fisher (1912)

contortions in his attempts to avoid those implications, we have to credit him with sensing the problem, even if he could not fully face it.

Returning to the 1912 paper, Fisher, having pointed out the lack of any theoretical justification for the form of the equations implied by the 'method of moments', abruptly switches to a generalised discussion in which he declares:- *' If f is an ordinate of the theoretical curve ....., then*

$$log\ P = \sum_{1}^{n} log\ f \qquad (7\text{-}6)$$

(and) ... the most probable set of values ..... will make P a maximum.[1]*'* In conclusion, he writes:- *'We have now obtained an absolute criterion for finding the relative probabilities of different ... values for the elements of a probability system ..... . It would now seem natural to obtain an expression for the probability that the true values .... should lie within any given range. Unfortunately .... the quantity P ...(is a) ... relative probability ..... incapable ..... of giving any estimate of absolute probability'.* However, these claims, even with respect to relative probabilities[2], are invalid. As we have already shown in the case of Bayes' experiment, which is probably the simplest possible case, the assumption of the uniform prior is, generally, a necessary condition if the point where Fisher's criterion achieves its maximum is to be accepted as the most probable value of a parameter. However, we cannot assume that this early paper represents Fisher's mature view and it would be unreasonable to dismiss the whole of his work in this area on the basis of this single invalid claim. Nevertheless, we have here the first signs of crucial weakness in the fiducial argument.

The next paper directly relevant to the fiducial argument is entitled *'On the probable error of a coefficient of correlation deduced from a small sample'* and was published in 1921. In this paper, the arguments swirl and break like eddies in a turbulent stream, but among them we find [3]:- *'The attempt made by Bayes ..... depended upon an arbitrary assumption, so that the whole method has been widely discredited .... two radically distinct concepts have been confused under the name of 'probability' and only by sharply distinguishing these can we state accurately what information a sample does give us respecting the population from which it is drawn'.*

---

1 *op cit.* p.157

2 These probabilities are 'relative' in the sense that each is multiplied by an identical but unknown constant. Fisher seems not to consider the sense in which they are relative to the data.

3 *op. cit. p4*

Unfortunately, here, and as so often in Fisher's writing, he brings us to a point of great expectation and then swings into a massive digression. Thus, it is only when we come to page 16 that the argument is sensibly resumed:- *'.... what I previously termed the 'most likely value', ... I now, for greater precision, term the 'optimum' value ..... . It therefore involves no assumption whatsoever as to the probable distribution . ..... As ... I pointed out in 1912 the optimum is obtained by a criterion which is absolutely independent of any assumption respecting the 'a priori' probability of any particular value. It is therefore the correct value to use when we wish for the best value for the given data, unbiased by any 'a priori' presuppositions'.* There is then an abrupt break in the philosophical argument and it is only eight pages later that, without warning, we find a *'Note on the confusion between Bayes' Rule and my method of the evaluation of the optimum'* which reads:- *'My treatment differs radically from that of Bayes (who) attempted to find, by observing a sample, the actual probability that the population value lay in any given range. ..... Such a problem is indeterminate without knowing the statistical mechanism under which different values ... come into existence: it cannot be solved from .... any number of samples. What we can find from a sample is the likelihood of any particular value ....., if we define the likelihood as a quantity proportional to the probability that, from a population having that particular value .... , a sample having the observed value ..... should be obtained. ..... So defined, probability and likelihood are .... entirely different. ..... Numerically, the likelihood may be measured in terms of its maximum value; the likelihood of the optimum being taken as unity. ..... The concepts of probability and likelihood are applicable to two mutually exclusive categories of quantities. .... We may discuss the probability of occurrence of quantities which can be observed or deduced from observations, in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of hypotheses ..... On the other hand we may ascertain the likelihood of hypotheses ..... by calculation from observation ..... '.*

There is, therefore, across the two papers so far considered, a considerable change of position. Unfortunately, at no point does Fisher acknowledge the change. He merely leaves the reader to make what one can from the shifts of meaning and it is in this manner that we meet Fisher's formalisation of the concept to which he attached the name 'likelihood'. This was, however, rather high-handed, for 'likelihood' was, and is, an ordinary word in the English language which was, and still is, used both by ordinary people and by specialists, in a variety of ways. Serious misunderstandings can therefore follow when the meaning of such a word is arbitrarily confined by a

technical clique to some new and narrow meaning. However, whatever the rights and wrongs of the linguistic issue, it is a plain fact that, in the technical discourse of twentieth century statistics, the term 'likelihood' has become restricted to the sense assigned by Fisher, *i.e.* it is a number proportional to the probability that a defined population, or probablistic process, should yield, from a random sample, a value in a specified range. Hence, if we are considering a pair of alternative values, v_1 and v_2 for a parameter which characterises a given population, and we have an observed value *'r'*, such that $\mathcal{P}(r \mid v_1) = 0.9$ and $\mathcal{P}(r \mid v_2) = 0.1$, then we may say that the likelihood of v_1 is nine times greater than that of v_2, relative to the observed value *'r'*. The existence of such relationships had however long been known and it is interesting to note that De Morgan, had written that *'causes are likely or unlikely, just in the same proportion that it is likely or unlikely that observed events should follow from them. The most probable cause is that from which the observed event could most easily have arisen*[1]*'*. One hundred and fifty years earlier, however, writers, such as Leibniz and Jac. Bernoulli, had been aware of the pitfalls in this mode of reasoning[2] and, as Keynes so aptly comments, *'If a cause is very improbable in itself, the occurrence of an event, which might very easily follow from it, is not necessarily, so long as there are other possible causes, strong evidence in its favour*[3]*'*. Yet, while it must be acknowledged that Fisher's appropriation of 'likelihood' as a verbal tag has been useful in simplifying statistical discourse, it would have been even more helpful had he acknowledged the connection with Bayes' third proposition[4]:-

$$\mathcal{P}(E_1 \wedge E_2) \quad = \quad \mathcal{P}(E_1) \times \mathcal{P}(E_2 \mid E_1) \qquad (7\text{-}7)$$

That is, if E_1 is the event that an hypothesis H relating to an event E_2 is true, then the likelihood of H, which we may denote $\mathcal{L}(H)$, is a quantity proportional to the probability of occurrence of the event E_2, if we are given E_1; that is, $\mathcal{L}(H) = k \times \mathcal{P}(E_2 \mid E_1)$, where k is an arbitrary constant.

Ten years later, Fisher returned to the subject in a 1930 paper to the Cambridge Philosophical Society under the title *Inverse Probability*. This paper was later reprinted in a collection[5], where Fisher added the comment:- *' The importance of the paper lies ..... in setting forth a new mode of rea-*

---

[1] According to Keynes (1921) p178, the quoted view was expressed by De Morgan in *The Cabinet Encyclopoedia*, p27. See also below, Ch 11, p162

[2] See Keynes(1921) pp368-9

[3] Keynes (1921) p 178.

[4] See eqn (3-26) in Ch 3 above

[5] Fisher(1950) p22.527a

soning from observations to .... hypothetical causes'. The paper is, in fact, indispensable for anyone attempting to understand the fiducial argument.[1] It presents a number of impeccable insights, starting with the first paragraph where, in commenting on Bayes' essay, Fisher notes that *'Laplace takes for granted in a highly generalised form what Bayes tentatively wished to postulate in a special case'.* The significance of this remark, in our context, is the open acknowledgement that Bayes was dealing with a special case and that the illicit generalisations were, in fact, due to others. Unfortunately, Fisher seems later to have lost sight of this fact and there is little doubt that the quality of his further work in this area was seriously degraded by his failure to discriminate between what Bayes actually wrote and the misuse of Bayes' work by others. The next item of significance in *'Inverse Probability'* is the paragraph which reads [2]:- *'The ..... development of the subject has reduced the original question of the inverse argument in respect of probabilities to ..... a series of ..... analogous questions [3]; the hypothetical value .... may be a probability, but it may equally be ..... any physical magnitude about which the observations may be expected to supply information'.* Clearly, this paragraph shows a precise understanding by Fisher, at that time, of the strict bounds which had been placed by Bayes on the problem addressed in Bayes' Essay. Later however Fisher fails to remember the contrast between that original, firmly bounded problem, and, on the other hand, the plethora of conceptually related but logically distinct problems. The contrast, however, is crucial to the understanding and analysis of all these matters.

The next eleven paragraphs or so of *Inverse Probability* then present an outstandingly clear exposition of the problem which arises if a uniform prior distribution is assumed in the case of a dimensional variate, *i.e.* a prior distribution which is uniform for one formulation of the unknown may be utterly non-uniform for an alternative, equally valid, formulation. A simple example of this is easily seen in the distributions of times and velocities which result when a given distance is travelled by a population of vehicles moving with various speeds. However, although Fisher also notes that divisions of a population into equally sized groups will preserve that equality under many transformations, he fails to point out that it is precisely this type

---

[1] Hacking (1965) p133 (footnote), however describes Fisher's 1945 paper on *The Logical Inversion of the Notion of the Random Variable* as containing the clearest available statement of Fisher's views on fiducial probability.

[2] Fisher (1930) p529

[3] Fisher actually writes *quite analogous*. The use of the word *quite* in this context seems however peculiar. The phrase ***not*** *quite analogous* would seem to make better sense.

of limited case which Bayes actually addressed by confining his experiment to the counting of events in a specified group. Fisher then concludes this section of *'Inverse Probability'* with the challenging point that although we may be unable to give *a priori* reasons for the assumption of a uniform prior distribution, the method of inverse probability, based on that assumption, could claim the justification that it did at least yield information of a sort on things which were previously unknown, and avoided the hugely greater problem of saying that observation of the unknown could yield no improvement in our knowledge. In Fisher's words:- *' Inverse probability has .... survived so long in spite of its unsatisfactory basis, because its critics have .... put forward nothing to replace it as a rational theory of learning by experience'*. However, it is only after some digressions to mention Gauss and to claim 'supreme value' for the method of maximum likelihood, that, with the last paragraph on *page 532*, we come to the nub of the matter concerning fiducial probability:- *'There are however certain cases in which statements .... of probability can be made with respect to the parameters of the population. ..... In many cases the random sampling distribution of a statistic, T, ... is expressible solely in terms of a single parameter, of which T is the estimate found by the method of maximum likelihood. If T is a statistic of continuous variation, and P the probability that T should be less than any specified value, we have then a relationship of the form* $P = F(T,\theta)'$.

Here, we have to interrupt Fisher's argument in order to point out a simple fact that he takes for granted, but which is necessary for a smooth understanding of what follows: namely, that the above expression assumes a sample of a specific size n from a population which is characterised by a specific value of θ. If, therefore we are told that the sample size is *4* and that θ has a value of *0.5*, P then denotes the probability that the statistic T will have a value less than, say, *0.49*, the magnitude of that probability being returned by evaluation of a function which we would prefer to denote as $F_T(\theta,n)$. With this in mind, we return to Fisher, albeit with slight paraphrasing and modifications to his notation:- *'If now we give to P any particular value such as 0.95, we have a relationship between the statistic T and the parameter* θ *such that we can define* $T_{0.95}$ *as the 95 per cent (probability) value corresponding to a given* θ, ... *'* With that assertion we have no disagreement. Fisher then continues with the statement:- *'... and this relationship implies the perfectly objective fact that in 5 per cent of samples T will exceed the 95 per cent (probability) value'*. This assertion is, however, patently false; for the most that can be said is that the expected proportion of samples in which T will exceed the *95%* probability value is *5%*. He

then says:- *'To any value of T there will moreover be usually a particular value of θ to which it bears this relationship; we may call this the 'fiducial 5 per cent value of θ corresponding to a given T '. ..... (and) ..... If as usually if not always happens, T increases with θ for all possible values, we may express the relationship by saying that the true value of θ will be less than the fiducial 5 per cent value corresponding to the observed value of T in exactly 5 trials in 100 '.* This last assertion, however, is again false: patently so, for the reason given above, concerning the *expected proportion*, and is more subtly fallacious, for reasons that we put aside for a moment while Fisher continues:- *'By constructing a table of corresponding values, we may know as soon as T is calculated what is the fiducial 5 per cent value of θ, and that the true value of θ will be less than this value in just 5 per cent of trials* [1]. *This then is a definite probability statement about the unknown parameter θ, which is true irrespective of any assumption as to its 'a priori' distribution'.* We shall however show presently that such probability statements concerning θ are, in general, fallacious. However, we shall also find that, in certain important cases, Fisher's assertions are indeed true, albeit his reasoning was seriously at fault. In the meantime, after a detailed illustration, using the correlation co-efficient as an example[2], Fisher continues:- *' It is therefore important to realise exactly what such a probability statement, bearing a strong superficial resemblance to an inverse probability statement, really means '.* Unfortunately, having aroused the expectation that we are on the point of being told precisely what such a probability statement really means, the reader is then treated to nothing of the sort, but merely to a few lines describing instances in which the fiducial probability differs from the result achieved by the method of 'inverse probability'.

In 1956, many years after the publication of *'Inverse Probability'*, the fiducial argument received a new twist in Chapter 3 of Fisher's *'Scientific Methods and Statistical Inference*'. Here, he describes an imaginary experiment, concerning the random emission of radio-active particles, to which Fisher refers in numerous other publications. The section in which the experiment with radio-active emission is described, begins with the heading

---

[1] These precise numeric assertions are, of course, totally false and could be extremely misleading for a naïve reader. The truth is that if we take numerous *independent* sample-sets of a given size, we will only *expect* the frequency of cases with which the assertion is found to be correct, to follow, yet again, a distribution of the general binomial form. There is no guarantee that the expected frequency will occur in any sample-set whatsoever, let alone in any *given* sample.

[2] Thus hitting back, as Zabell (1992), points out, at Pearson who, some years earlier had treated Fisher rather shabbily on a closely related question.

'The fiducial argument [1]*'*. Fisher explains the meaning of *fiducial* as follows:- *'In the Bayesian argument the observations are used to convert a random variable having a well defined distribution 'a priori' to a random variable having an equally well-defined distribution 'a posteriori' .... . By contrast, the fiducial argument uses the observations only to change the logical status of the parameter from one in which nothing is known of it, and no probability statement about it can be made, to the status of a random variable having a well-defined distribution'.* Taken at their face value, the assertions contained in this quotation are staggering: that which concerns *the Bayesian argument* ignores, totally, and yet again, the fact that Bayes defined his problem as concerning the case in which 'nothing at all is known antecedently'. We can, however, make more sense of Fisher's assertions if we assume that Fisher's attitude to Bayes was utterly ambivalent, believing with one half of his mind, that Bayes had failed to solve the defined problem, and that Bayes' success was therefore limited to the demonstration of various rather elementary propositions [2]. Yet Fisher could not bring himself to say this openly; the reason was, we suspect, that the other half of his mind could not accept that a scholar as accomplished as Bayes could really have failed so dismally; or, perhaps indeed Fisher could intuitively, but only subconsciously, perceive that Bayes' postulate might be not so arbitrary and unwarranted as it was then fashionable to assert[3]. Whatever the truth, and it is now probably indeterminable[4], it is certainly the case that Fisher's paper of 1962 reveals a marked change of attitude and could well be symptomatic of a subconsciously perceived truth slowly working its way into a verbalised consciousness[5].

If however we view Fisher's primary aim as being to present, somewhat covertly, a contrast between his own (claimed) success and Bayes' (implied) failure, it becomes clear that we have to parse the phrase *'the fiducial argument uses the observations only to change the logical status of the parameter'* in such a way that the term *'only'* qualifies *observations*, rather than the infinitive, *to change.* Hence, re-arranging the words to eliminate the ambiguity gives us:- *'the fiducial argument uses only the observations'*. That is, we have a contrast with Bayes' solution where we

---

[1] Fisher (1956), p54

[2] It is remarkable that although Fisher quite often damns the work of others, such as Pearson and Neyman, in his writings, he appears to treat Bayes with an awed reverence, despite the repeated implications that Bayes' essay utterly failed to achieve the stated objective.

[3] *i.e.* in contrast to a pragmatic *convenience*.

[4] *i.e.* failing discovery of previously unknown papers by Fisher.

[5] The rôle of the intuitive part of the mind in mathematics is strongly argued in Penrose (1989), p538.

require both the observations and an assumption about the prior distribution. We also find, in the above quotation, a stark admission of the claim, previously only implicit, that *'.... the fiducial argument change(s) the logical status of the parameter ..... to (that) ..... of a random variable having a well-defined distribution',* albeit we are not told whence we obtain the specification of this allegedly well-defined distribution.

However, on p55 of *'Scientific methods and Statistical Inference',* Fisher asserts:- *'If direct and exact observations could be made on the parameter itself, a similar change of logical status would be effected by the observation of its value, from one in which it was wholly unknown, ... to one in which it could be assigned a definite value. It is, therefore, perhaps not surprising that similar exact observations, though not on the parameter itself yet on variates having distributions known in terms of the parameter, should be able in favourable cases to effect, at a lower level, a similar change of status'.* Thus, here, in the words *variates having distributions known in terms of the parameter*, Fisher is yet again invoking prior knowledge of a kind which may not be identical to Bayes' assumption of the uniform prior, but is remarkably similar in a general sense. It is again staggering that Fisher could not see this for himself. He then gives as an example of his own mode of reasoning, the radio-active source emitting particles at instants which are 'completely independent' of each other but such that the time-interval between any two successive emissions is randomly distributed with a probability governed by an exponential distribution. This is followed by several pages of appallingly obscure argument where, without the help from other sources, we would have found it utterly impossible to see at what Fisher was driving. Fortunately, there is a clue on *page 60*, in the Section headed *Accurate statements of precision*. The section begins *'The possibility of making exact statements of probability about unknown constants of Nature supplies a need long felt of making a complete specification of the precision with which such constants are estimated.....'..*

Later, a paper in the *Journal of the Royal Statistical Society* for 1962, provides a further clue[1], leading us to a more lucid, and perhaps more sympathetic, understanding of Fisher's aims. The paper begins:- *'It has become realized in recent years (Fisher, 1958)[2] that although Bayes considered the special axiom associated with his name for assigning probabilities 'a priori', and devoted a 'scholium' to its discussion, in his actual mathematics he*

[1] Fisher (1962)

[2] The date *1958* may be a misprint. The references in the paper contain no publication with that date.

avoided this axiomatic approach as open to dispute, but showed that its purpose could be served by an auxiliary experiment, so that probability statements 'a posteriori' at which he arrived were freed from any reliance on the axiom, and shown to be demonstrable on the basis of observations only[1]*, such as are the source of new knowledge in the natural sciences '.* Putting aside the fact that Bayes actually used the experiment in order to support, in the *Scholium*, the assumption of the uniform prior, it remains interesting that Fisher then outlines parallels which, he asserts, can be drawn between his thought-experiment with the radio-active source and Bayes' experiment. He indicates four problems which can be considered in relation to such experiments, and in discussion of these experiments he displays the slow change in his attitude to Bayes on which we have already remarked. Unfortunately, he does not finalise the change but remains fixed on the superficially opposite objective of ensuring *that no probability statement need be made axiomatically*[2]. This, he seems to believe he demonstrates in his discussion of 'Problem A' of the paper, in the conclusion to which he writes:- *'In fact, as by Bayes, the unknown p is evaluated as a random variable, and .... The dogma that direct observations cannot serve to express an unknown parameter of Nature as a random variable cannot be sustained.*[3] *'* Although Fisher gives no indication as to where we might find any assertion of the alleged dogma, we have, perhaps, come as close as we shall to a clear and concise expression of the motive that actually seems to dominate Fisher's writing in this area, *i.e.* to demonstrate that *direct observations can serve to express an unknown parameter of Nature as a random variable.*

With this motivation established, many other parts of the puzzle fall into place, but we should not pass without mention the attempt to claim Bayes' support for the evaluation of *'the unknown p ..... as a random variable'*: this being yet another appalling distortion of Bayes' argument. For Bayes unswervingly evaluates *p* as a governing parameter in a statistical process. However, it is worthwhile considering a variant on Bayes' experiment in relation to the emission of radioactive particles, in terms of a trial which tests whether a particle is emitted in a time-interval of a fixed duration, starting at an arbitrary instant. Clearly, a set of *n* such trials has the same logical form as Bayes' original experiment, and the emission of particles in *m* of these trials can be analysed in terms of, (an assumed constant),

---

[1]This is, in our view, a gross and utterly reprehensible misrepresentation of Bayes approach. It is totally at variance with what Bayes wrote in the *Scholium*, which we have examined in detail in Chapter 4 above.

[2] Fisher (1962) p.119

[3] Fisher (1962) p.120.

Version zh25a.doc at 18:36 on 13 July, 2001

underlying probability. This does not quite meet the objective which Fisher would like to achieve, *i.e.* determination of the underlying *average* rate of emissions to within definable probability-limits. To achieve this, we make the duration of each trial short enough to make the probability of more than one emission in a single trial, vanishingly small. By this means, we effectively eliminate the time-dimension from the problem and we are able to treat it, as in Bayes' experiment, as a problem in the estimation of a pure, non-dimensional, probability, *i.e.* we have a set of *n* identical trials in which a given event occurs on *m* occasions, and the chance that the underlying probability of that event lies between any two given values can be computed exactly as shown by Bayes.

The case of the radio-active experiment is however a complex example which depends critically upon taking for granted the exponential model of the underlying process, and it seems totally to escape Fisher's notice that the magnitude of the epistemological and experimental problem of justifying such an assumption, may well seem to render Bayes' assumption of the uniform prior trivial in comparison. It is better therefore to take the much simpler example of Fisher's contention which is given in his *1945* paper entitled *The Logical Inversion of the Notion of the Random Variable*. In this paper, after dealing with a somewhat complex case involving the mean values of normally distributed variates, he continues:- *' It is instructive to compare the general form of the fiducial argument set out above with a special case of great simplicity, suitable for examining its logical cogency. Let* μ *be the median of a distribution of which nothing is known save that its probability integral is continuous ..... Let* x_1 *and* x_2 *be two observations of the variate; then for any given value of* μ *it will be true that: (1) in one case out of 4 both* x_1 *and* x_2 *will exceed the median, (2) in two cases out of 4 , one value will exceed and the other be less than the median, (3) in one case out of 4, both will be less than the median,*[1] *(Hence) ..... we may argue from two given observations, now regarded as fixed parameters that the probability is 1/4 that* μ *is less than both* x_1 *and* x_2*, that the probability is 1/2 that* μ *lies between* x_1 *and* x_2*, and that the probability is 1/4 that* μ *exceeds both* x_1 *and* x_2*. The argument thus leads to a frequency distribution of* μ*, now regarded as a random variate. The idea that probability statements about unknown parameters cannot be derived from data consisting of observations can only be upheld by those willing to reject this simple argument*'. Thus we have here an example of the fiducial argument in what is probably its most plain and simple form. Although the argument received a

[1] Taken at face value, these numerically precise assertions are again simply false.

further twist in terms of a non-recognisable subset which was later shown to be indeed recognisable[1], it is our view that the essence of the argument is adequately represented above and it is now appropriate to evaluate in more detail its validity.

We start by noting that we are concerned with the probabilities of three different types of event:-

Type 1 : A defined event, randomly occurring within a distribution which is governed by one or more parameters, occurs in a specified trial.

Type 2 : A defined, deterministic, but uncertain event occurs in a specified trial.

Type 3 : An inference concerning a governing parameter, based upon observation of a *Type 1* or *Type 2* event[2], in a specified trial, is correct.

That is simple. It is also often simple, at least in principle, to define the rules by which we compute the probability if the assertion concerns, for example, the number of occasions on which a *Type 1* event occurs in *n* independent trials, and we are given, either as an assumption, or as evidence, the probability of its occurrence in a single trial. In such cases, the mapping operation which we symbolise by $\mathcal{P}(\,.\,)$ proceeds mechanically. In deterministic *Type 2* situations, as we saw in chapter 6, there are fixed causal factors but we are unsure of their magnitudes and relationships, and the aim is to compute the probability of an event in a specific situation. In such cases there is, however, no random independence between trials. A given situation always produces the same result. *Types 2* and *3* differ, however, from *Type 1*, in a number of ways, starting with the fact that, whereas *Type 1* events have a fairly straightforward interpretation in terms of relative frequencies, no such interpretation can be placed on probabilities of *Type 2* or *Type 3*. With *Type 3*, however, the fact that an event, which may belong to *Type 1* or to *Type 2*, has or has not happened, comprises the evidence, from which, in conjunction with certain assumptions, we are required to compute the probability that a defined hypothesis concerning a governing parameter, is true. (As an aside, it is worth noting that a great deal of nonsense about the probability of sunrise stemmed from a failure to perceive that the rising of the sun is a deterministic *Type 2* event. Commentators who treated it as a random *Type 1* event were therefore liable to some ridicule. Price, however,

[1] See Buehler and Feddersen (1963)

[2] Or set of events,

Version zh25a.doc at 18:36 on 13 July, 2001

in his addendum to Bayes' essay[1], is punctilious in his analysis of these issues, albeit he unfortunately fails to differentiate formally between *Type 1* events and *Type 2* events).

However, for each type of case, we have hitherto used a symbolic assertion of the single form $\mathcal{P}(E) = x$ to denote all these different kinds of probability and we have often overlooked the need, in every case, to define the evidence and assumptions on which the proposition is based. In every assertion of probability, however, the argument has three parts:- *(i)* the proposition, *(ii)* the evidence, *(iii)* the assumptions. Given the marked differences between the three types of assertion, it seems misleading and unjustifiable, that we should continue to use the same symbolic expression in each case, without distinction. We therefore propose, tentatively, the symbols $\mathcal{P}_R$, $\mathcal{P}_D$ and $\mathcal{P}_H$ to denote the assignments of probabilities in the three cases[2]. Thus we use the symbolic term $\mathcal{P}_R(E \mid k, A)$ to denote the probability, given evidence k and assumptions A, that a random *Type 1* event, E, will occur in a defined trial. Similarly, we let the symbolic term $\mathcal{P}_D(D \mid k, A)$ denote the probability that a deterministic *Type 2* event, D, will occur, given k, A. Finally, we let $\mathcal{P}_H(H \mid k, A)$ denote the probability that a defined hypothesis, H, concerning the governing parameter of a trial in which an event of *Type 1*, or of *Type 2* occurred, is true. These symbols, however, still do not differentiate sufficiently between the types, for the nature of the support, k,A, which is required by a *Type 1* assertion, is quite different from that required by a *Type 3* assertion. Specifically, a *Type 1* assertion requires that the term k,A shall provide information about the governing distribution over the possible values of the event E. In contrast, a *Type 3* assertion requires that k,A shall provide information about the occurrence of a *Type 1* event E. (We are here putting aside further consideration of *Type 2*, which is not part of our main enquiry and will be better investigated elsewhere). Hence, in relation to *Type 1* and *Type 3*, we have to elaborate the symbolic expressions to denote the specific structure which is required in each case. That is, for a *Type 1* assertion we require a symbolic term equivalent to $'\mathcal{P}_R(E \mid G(k,A))'$, where $G(k,A)$ denotes evidence and assumptions relating to the governing distribution over the event E. Correspondingly, for *Type 3*, we require a symbolic term equivalent to $'\mathcal{P}_H(H \mid J(k,A)),'$ where $J(k,A)$ denotes evidence and assumptions concerning the occurrence of a *Type 1* event.

---

[1] Bayes (1763) pp.409-410.

[2] Lindley (1970, Section 5.1) uses $'p'$ where we use $\mathcal{P}_R$ and $'\pi'$ where we use $\mathcal{P}_H$

Returning to the fiducial argument, we can now see that, where in a *Type 1* assertion, the random event *E*, is an algebraic proposition, defining a relationship between a random *Type 1* variable and one or more given parameters, a true assertion will be true for all other propositional values of *E* which are algebraically equivalent[1]. For example, if *E* has the propositional value *'e > 0.1',* where *e* is a random variable having a distribution given by *G(k,A)*, and if it is also given that

$$\mathcal{P}_R(e > 0.1 \mid G(k,A)\ ) = 0.5 \qquad (7\text{-}8)$$

then all other assertions of the form $\mathcal{P}_R(E \mid G(k,A)\ ) = 0.5$, where *E* can take on any propositional value which is algebraically equivalent to *'e > 0.1'* will also be true. For example,

$$\mathcal{P}_R(0.1 < e \mid G(k,A)\ ) = 0.5 \qquad (7\text{-}9)$$

will be true, as will

$$\mathcal{P}_R(1.1 < 1+e \mid G(k,A)\ ) = 0.5 \qquad (7\text{-}10)$$

In the fiducial argument, a fundamental problem, which is concealed by the ambiguous notation, now becomes clear. That is, the fiducial argument asserts that if a statistic *T* is derived from a set of *n* events of *Type 1*, drawn from a population governed by a parameter θ, and we are able to make an assertion of the form[2]

$$\mathcal{P}(\ T < \theta - a\ ) = F(n,a,\theta) \qquad (7\text{-}11)$$

(where *a* is an arbitrary offset from θ), then, Fisher asserts, we are conversely able to assert:-

$$\mathcal{P}(\ \theta > T + a\ ) = F(n,a,\theta) \qquad (7\text{-}12)$$

which has the superficial appearance of an assertion concerning a probability distribution over the possible values of θ. If however we use the unambiguous notation presented above, we get, instead of *(7-11)*:-

$$\mathcal{P}_R(\ T < \theta - a \mid G(\ a,\theta,\ A)\ ) = F\ (n,a,\theta,A) \qquad (7\text{-}13)$$

and, although the equivalent expression

$$\mathcal{P}_R(\ \theta > T + a \mid G(\ a,\theta,\ A)\ ) = F\ (n,a,\theta,A) \qquad (7\text{-}14)$$

---

[1] The equivalence must, of course, be unique. A relationship such as *x^2 = 4* is not a unique equivalent of *x=2*.

[2] The expressions $F(n,x,\theta)$ and $F_R(n,x,\eta,A)$ simply denote functions which return values which we can equate to probabilities.

has the superficial appearance of an assertion concerning the distribution of θ, a glance at the conditioning term $G(a,\theta, A)$ shows immediately that the values of both θ and a are given and that the assertion actually concerns the distribution of the statistic T. In these situations, θ is a governing parameter which does not vary randomly from trial to trial, but is fixed for the set. It is therefore contradictory to treat a random variable, such as T, derived from the outcomes of a trial-set as if it could also be a governing parameter in that same trial. Being based on this contradiction, it follows that the fiducial argument is clearly fallacious. The relevance of the *'statistical mechanism'* which brings θ into existence is however a question to which we return in *Chapter 10*.

A further, and important illustration of fallacious reasoning in the fiducial argument is provided by the case of a median, symbolised by μ, which is known to have a value, say *0.3*, in a given population, and an object, selected at random from that population, the value of which is unknown, but is symbolised by *t*. By the definition of the median, this allows us to assert:-

$$\mathcal{P}_R (t > 0.3 \mid \mu = 0.3) = 0.5; \qquad (7\text{-}15)$$

or indeed to assert

$$\mathcal{P}_R (0.3 < t \mid \mu = 0.3) = 0.5 \qquad (7\text{-}16)$$

However, if we examine the sample and find that $t = 0.4$, an expression of the form

$$\mathcal{P}_R (\mu < 0.4 \mid t = 0.4) = 0.5 \qquad (7\text{-}17)$$

which appears to be, and indeed is, an assertion that μ is a random *Type 1* variable, is yet again self-contradictory and therefore fallacious, in that μ is already defined as being a governing parameter, fixed for the entire trial-set, while *t* varies randomly from trial to trial within the set.

The approach which we actually require, if we wish to make a probablistic assertion concerning the value of μ, based upon a trial in which we have measured *t,* is to evaluate an expression of the form

$$x = \mathcal{P}_H(\mu < t \mid t) \qquad (7\text{-}18)$$

i.e. we are required to assess the probability that an inference concerning the value of a governing parameter, based upon observation of an event, is correct. This question is addressed by Bayes, quite generally[1], in Proposi-

[1] *i.e.* Bayes' treatment is not tied to the discreet, binomial problem at this point in his essay.

tions 4 and 5, and in the *Scholium*. Also, we shall later find, by following Bayes, that the assertion, so similar to *(7-17)*

$$\mathcal{P}_H(\mu < 0.4 \mid t = 0.4) = 0.5 \qquad (7\text{-}19)$$

is indeed both logically valid and numerically correct. Fisher's conclusion, in the case of the median, was correct; his reasoning was flawed.

A less fundamental, but also self-contradictory, aspect of the fiducial argument, is that the knowledge of the distribution governing the occurrence of Type 1 events, upon which Fisher bases the fiducial argument, is totally dependent upon the pre-existence of a governing parameter. And it is indeed central to the fiducial argument that the value of a governing parameter is a valid object of a probability assertion. Yet, while Fisher asserts that cases exist in which the prior distribution of probabilities attaching to different values of governing parameters is known, and to which the method of Bayes can be validly applied[1], Fisher also claims that if we lack this knowledge, the fiducial argument will allow us to compute the required probability without reference to the missing parameter. Yet, if we lack prior information concerning that parameter, it is inconceivable that an abstract argument can supply the missing data. The fact is that in cases where we have no prior information, we have three options. We can:- *(1)* refrain from any inference concerning the value of the probability in question, or, *(2)* we can state likelihood ratios based on the observations and various hypotheses, or *(3)* we can assume the uniform prior, and accordingly give our estimate of the probability in question. However, in later chapters we shall see that, despite the flaws in the fiducial argument, Fisher's concerns were indeed valid and are capable of resolution, albeit not on his terms.

We must also sound a note of caution concerning the use of likelihood. For there can be a temptation to avoid the difficulty and tedium of considering prior probabilities by choosing to work only in terms of likelihood, and implicitly pleading that admission of the fact provides a justification for the deed. Careful consideration shows, however, that one should always, before using likelihood as a deciding factor, consider the effects that 'prior' information might have on one's views. A relatively minor problem can arise if one is using likelihood because, when dealing with 'unknown events' one refuses to accept Bayes' postulate of the uniform prior, but an unexpected discovery of prior information then requires, at least conceptually, a major revision of views. A more serious danger is illustrated by Fisher's mice, discussed in the Appendix, where the likelihoods of a pair of mutually exclusive hypotheses are in the ratio of *128 : 1*, yet the probability of either hypothesis can be raised to certainty if certain facts become known, or if

[1] See *e.g.* below, Appendix A, *'Fisher's Mice'*

certain events were to happen. The likelihood ratios, however, give no indication that these crucial sensitivities are present and it is a plain fact that decisions based upon likelihood in such situations, could lead to mistakes and catastrophes which would be avoided by paying proper attention to the probabilities and, wherever possible, to the acquisition of decisive evidence.

Postscript to Chapter 7

It is important to note that, according to the circumstances, Bayes' theorem can be expressed in various forms e.g. :-

$$\mathcal{P}_H(E_1 | k, E_2) = \frac{\mathcal{P}_H(E_1 | k) \times \mathcal{P}_R(E_2 | k, E_1)}{\mathcal{P}_R(E_2 | k)} \qquad (7\text{-}20)$$

or

$$\mathcal{P}_R(E_1 | k, E_2) = \frac{\mathcal{P}_R(E_1 | k) \times \mathcal{P}_R(E_2 | k, E_1)}{\mathcal{P}_R(E_2 | k)}$$

In (7-20), $\mathcal{P}_H$ applies to a value that is fixed for a set of trials in which the event, or set of events denoted by E_2 has occurred. In (7-21), $\mathcal{P}_R(E_1.....)$ applies to an event which can vary at random within a set of trials and may be only loosely correlated with the occurrence of E_2. The essential points are, first, that E_1 and E_2 can be events of any kind - provided they are events to which probabilities can be assigned - and, second, that the symbols must correctly represent the physics of the situation.

Chapter 8

A Critical Case

Chapter 31 of Keynes' *'Treatise on Probability'* is entitled *'The Inversion of Bernoulli's Theorem'*. Keynes sums up the works of many predecessors in the area as *'the children of loose thinking and the parents of charlatanry.[1]'* , but he also says:- *'It is reasonable to presume that, subject to suitable conditions .... an inversion of Bernoulli's Theorem must have validity*'. He then deals in an elegant and thorough manner with a general version of the problem addressed by Bayes, concluding[2]:- *'therefore, ... as the number of instances is increased the probability that the true value[3] is in the neighbourhood of m/n ..... tends towards certainty ..... But we are left with vagueness ..... respecting the number of instances we require. We know that we can get as near certainty as we choose by a finite number of instances, but what this number is we do not know. ..... It would be very surprising, in fact, if logic could tell us exactly how many instances we want, to yield us a given degree of certainty in empirical arguments. ..... Yet many persons seem to believe that ..... we can attribute a definite measure to our future expectations and can claim practical certainty for the results of predictions which lie within relatively narrow limits. Coolly considered, this is a preposterous claim, which would have been universally rejected long ago, if those who made it had not so successfully concealed themselves ..... in a maze of mathematics'*. To a sensitive ear, these words may sound like the thuds of heavy nails fixing the lid on the coffin of Bayes' theory.

However, all assertions of probabilities concerning phenomena in the physical world, are empirical in nature**;** they are never analytically, necessarily or tautologically true**;** they are always to some degree uncertain, they can therefore never be asserted with absolute confidence, but are dependent upon a basis of knowledge, or assumptions, about the physical world. They can be made only relative to that basis. Any assertion of a probability concerning a specific situation can be wrong in the sense that the totality of the assertions and the assumptions does not match the actual situation. Hence, every assertion of probability must always be, itself, subject to a governing

---

[1] Keynes (1921) p384

[2] *op.cit.* p389, but with little direct reference to Bayes.

[3] The true value is denoted by the letter *q* in the original. The ratio *m/n* is denoted by *q'*.

probability. Superficially, this leads to an infinite regress, which, in the limit, would drain every assertion of an empirical probability of its meaningful content. But the infinite regress can be avoided by, at any stage, stating the facts and assumptions which, if themselves true, will be sufficient to make the statement of probability also true.

Statements of prior distributions concerning phenomena in the physical world are likewise empirical assertions, subject to uncertainty, and to statements, or questions, of probability concerning their correctness. If therefore we have, on the one hand, an assertion of a prior distribution of probabilities over the possible outcomes of a trial, and, on the other hand we have an outcome which, on the basis of the prior, had an extremely low probability of occurrence, we can take the prior as given, and compute a posterior distribution in which the intensity of our confidence that the truth is in the region of the experimental result is attenuated by the prior, or we can allow the experimental result to attenuate our confidence in the correctness of the prior distribution. The latter course is equally reasonable and suggests that, in real-life situations, it will often be rational to require a fair degree of correspondence between the prior and the trial, for the pair to be jointly credible: in which connection one can use the denominator in Bayes' equation:-

$$ {}^{n}C_{m} \int_{0}^{1} x^{(m)}\ (1-x)^{(n\ -\ m)} P_{0}(x)\, dx $$

as a measure of the joint credibility of the prior and the observed data. But there are, of course, situations which involve intensely concentrated prior probabilities, surrounded by bands where the prior probabilities are very low, but the correspondingly rare events are by no means impossible.

Yet it may be thought that because there is only a very low probability - arguably infinitesimally small - that the assumption of a uniform prior, in a trial of an unknown event, is correct - there is a correspondingly low probability that the solution to Bayes' problem as in equation *(5-18)* will be correct. This is however fallacious. If a statement of probability concerning an hypothesis is of the form $\mathcal{P}_H(.) = f(d, a_1, a_2 \ldots\ldots a_n)$, where d denotes the observed data and $a_1 \ldots a_n$ denote the assumptions, there is a temptation to assume that if the truth of each assumption is governed by a probability $\mathcal{P}_H(a_i)$ then the resultant value of $\mathcal{P}_H(.)$ is attenuated by a proportional uncertainty. This is not, however, true.

The true situation is in general complicated and delicate, for the impact of the probability concerning each assumption can only be assessed by

careful consideration of the influence which each has on the resultant value of $\mathcal{P}_H(\,.\,)$, taken in the context of all other assumptions, their own probabilities and, not least, the observed data. Thus the impact of assuming a uniform prior in a Bayes' trial of an unknown event needs to be examined quite carefully in relation to possible forms of the true, but unknown, prior and also in relation to the magnitudes of m and n. More generally, we have the fact that, although an assertion may be qualitatively wrong, a quantitative assessment may show the error to be of no practical consequence. In this regard, authors[1] who felt that Bayes' equation was invalidated by the (allegedly arbitrary) assumption of the uniform prior, were, perhaps understandably, failing to discriminate between, on the one hand, a conclusion to an argument which is deemed qualitatively invalid because it can be shown not to follow rigorously from the premisses, and, on the other hand, a conclusion which may be qualitatively invalid but is, quantitatively, of acceptable accuracy. In practice, the latter may be much better than an assumption of total ignorance, to which we might be led by purely qualitative reasoning. No physicist, setting out to measure the speed of sound, can show that there is no possibility of error in the experimental set-up. Hence, by purely qualitative reasoning we can easily show that a decision to accept the results from such an experiment is 'arbitrary', and therefore, by facile extension, 'worthless'. From a practical point of view, however, it may be essential to accept such a result as the best available, and to bear in mind that it may be not totally dependable in all situations. It is, in this context of quantitative reasoning, and considering the connection between morality and probability, intriguing to note Kant's view that *'Empirical principles are wholly incapable of serving as a foundation for moral laws.*[2]*'* In our view however, quantitative reasoning, which is profoundly empirical in its essence, together with Bayes' concept of probability are, unless and until some better method be found, fundamental to the rationality of moral behaviour in the conditions of uncertainty under which many people feel that they have to live their lives. Some, however, blessed with feelings of total certainty in all situations, may feel no need of Bayes, but may little realise that their feelings of blessed certainty may manifest themselves as appalling inflictions on others. From a quantitative and pragmatic point of view, therefore, the assumption of the uniform prior may be a matter, not of theoretical necessity, but of practical necessity for those who have to balance uncertainties concerning empirical facts against the probabilities, penalties and rewards of alternative actions.

---

[1] *e.g.* Boole, Keynes, Fisher, Fine,

[2] Kant (1785). See p60 of Abbott (1959).

The following scenario illustrates, the crux of the matter. Suppose that a sample of blood from a critically ill patient, shows that, in a total of *n* cells, a number *m* of those cells are of *type-M*. The scenario permits no possibility of acquiring further information and we have to decide, from the evidence provided by the sample, on a course of action according to a rule that, if the true proportion of *type-M* cells, P_m, is less than a threshold *T*, the correct treatment is to administer drugs, otherwise do nothing. If we take the wrong action, the patient dies. The scenario also requires us to assume:- *(i)* the validity of Bernoulli's theorem, and *(ii)* that Bayes' solution is valid in cases where a prior distribution is known. So, we consider the possible courses of action, which appear to be:-

(1) do nothing, or
(2) administer the drug, regardless of the evidence *(m,n)*, or
(3) decide by the flip of a coin whether to administer the drug , or
(4) accept *m/n* as our best estimate of P_m, such that, if it is less than *T*, we administer the drug, otherwise we do nothing, or
(5) we adopt Bayes' solution, we assume the uniform prior, and we compute the probability that P_m is below the threshold, *i.e.* $\mathcal{P}_H\{(0<P_m<T)|m,n,P_u)\}$. If the result is a probability greater than *0.5*, we administer the drug**;** otherwise we do nothing.

We now consider each possible course of action and thence determine the values of the parameters which we would have to use in Bayes' equation in order to get a result which would (just) justify the given action.

Regarding *(1)* above, defences for that course could be based on one of two claims:- either *(a)* that, if the true prior distribution is unknown, then Bayes' equation is irrelevant, or *(b)* that, in the scenario situation, Bayes' equation has no solution; hence, there is no rational basis on which to make a choice. These defences are however vitiated both by the existence of the information *(m,n)*, which has been, apparently, deemed to be of no relevance in the formation of the decision, and by the fact that, if Bayes' equation is accepted in situations where the prior is known, then the decision to do nothing implies the existence of a prior distribution $P_0(x)$, such that, whatever the values *(m,n)*, the probability that the true value is less than the threshold is always less than *0.5*. The assumption of such a prior distribution amounts, however, to an *a priori* decision to ignore the data *(m,n)*. We conclude that the course of action *(1)* is irrational and indefensible. By the converse reasoning, the course of action *(2)* is likewise irrational and indefensible.

The third course of action, *i.e.* to flip a coin, is, in its relevant logical features, pretty well identical with *(1)* and *(2)*, and we therefore move on to *(4)*, where, if m/n is less than T, we administer the drug; otherwise we do nothing. Here, instead of ignoring the value of m/n as in the previous cases, we use it, by implication in the following way:-

$$[m/n > T] \Rightarrow \mathcal{P}_H\{(T<P_m<1)|m,n\} > 0.5 \quad \therefore \rightarrow \textit{Do not administer}^1$$

$$[m/n < T] \Rightarrow \mathcal{P}_H\{(0<P_m<T)|m,n\} > 0.5 \quad \therefore \rightarrow \textit{Administer the drug}$$

whence, by principles of continuity *etc.*, there is an implicit assertion that if $m/n = T$, then the probability that $P_m < T$ is exactly equal to the probability that $P_m > T$ and therefore

$$\mathcal{P}_H\{(0<P_m<m/n)|m,n\}$$

$$= \mathcal{P}_H\{(m/n<P_m<1)|m,n\} = 0.5 \qquad (8\text{-}1)$$

However, Bernoulli's theorem tells us that the result (m,n) could be due to any value of P_m in the interval $(0 \rightarrow 1)$. If, therefore we consider the probability that each possible value of P_m should produce the result (m,n) and integrate these probabilities, with uniform weighting over all the possible values of P_m, the probability of the result (m,n) is given by:-

$$\mathcal{P}_R(m|\, n, 0<P_m<x\,) = {}^nC_m \int_0^1 x^{(m)}\,(1\text{-}x)^{(n\,-\,m)}\,dx \qquad (8\text{-}2)$$

Hence, it is only when $m = n/2$, giving a symmetrical binomial distribution about m, that the integrated probabilities are equally divided on each side of the point m/n as is needed to satisfy *(8-1)*. Therefore, if the posterior probabilities are to be equally balanced about m/n, whatever the value of that ratio may be, this can only be achieved if the posterior distribution has been reached *via* a prior distribution which has the property of ensuring that m/n is always the median of the posterior probabilities. This would imply that certain values of P_m had a greater prior probability than others and would constitute a significant and unwarranted assumption of additional knowledge above anything justified in the definition of the scenario. Furthermore, whenever the value of T can be equated to a rational fraction a/b, that same fraction can be generated by any trial in which $m = Na$ and $n = Nb$, where N is any positive integer. However, although the resultant value of m/n will be a/b in each case, the curve on which the putative prior distribution has to operate, acquires a progressively narrower peak as N increases, which would

[1] *i.e.* If m/n is above the threshold value, we infer that the probability of the true value (P_m) also being above the threshold, is over *50%*. Therefore we administer the drug.

imply that the prior distribution itself must change as a function of *N:* this is utterly irrational and is incompatible with the nature of a prior distribution which is, by definition, fixed before the trials take place. Such notions are therefore, at least *qualitatively*, untenable.

However, having rebuked others for failing to remember that qualitative errors may be quantitatively trivial, we must not ourselves fall into that same trap; for, in practice, on countless occasions, in practical situations, people do commonly compare a ratio *m/n* against a threshold *T*, when deciding upon a course of action. They do not stop to compute the median of the posterior probabilities according to Bayes' equation. The justifications for this simple approach are, first that it is practical, without requiring the use of calculators *etc.* in complex and physically arduous situations. Second, over much of the central range of possible values of P_m, the point defined by the ratio *m/n* is numerically 'close' to the median of the binomial distribution. Towards the edges however, *i.e.* where $P_m \to 0$ or $P_m \to 1$, the approximation breaks down. Thus, it may be defensible to use a simple comparison of *m/n* against *T* when we are working with well conditioned values of *m* and *n*, and we can assume a uniform prior distribution over the possible values of the unknown parameter P_m. Our conclusion is, therefore, that, while qualitatively speaking, the course of action defined in *(4)* entails contradictions and is qualitatively irrational, if we adopt a quantitative point of view, it is reasonable and acceptable when appropriate conditions are fulfilled.

In action *(5)*, we accept Bayes' theory and we assume the uniform prior distribution. The reasoning which allows us to do this is simply that, all the available information being contained in the numbers *(m,n),* the only rational point-estimate which we can give of P_m is the ratio *m/n***.** This is, as we saw in Chapter 5, our best estimate, even though we may have little or no idea as to how good or bad it may be. Having therefore allowed the assumption of the uniform prior in deciding that *m/n* is our best estimate, we may further continue that assumption into the assessment of the bounds on P_m. Indeed we must do so, for not to do so would be to create a contradiction within our reasoning. Yet, we have here a case where the rational decision has to be made, not upon the 'most probable' value of P_m, but upon our estimate of the point about which the posterior odds are evenly balanced *i.e.* the posterior median, μ. Therefore, we use the assumption of the uniform prior together with the equation :-

$$\mathcal{P}_H\{(0<P_m<\mu)|m,n\} = \mathcal{P}_H\{(\mu<P_m<1)|m,n\} = 0.5 \qquad (8\text{-}3)$$

in order to reach a decision. However, while this is the best we can do, it seems to be a best of extremely dubious quality. The essence of the matter is that, in the situation of ignorance, as defined in the scenario, when we are faced with, in Bayes' terms, 'unknown events', a procedure based on Bayes' postulate is the only non-irrational means of deciding whether to administer the drug or to refrain. To do nothing, even on the basis that we are totally ignorant of the true prior, is not a neutral option because, as shown above, the practical implication is equivalent to knowing a non-uniform prior distribution, which gives higher probabilities to some values than to others. In a purely mathematical context, we normally deem a conclusion which does not follow rigorously from the premisses to be invalid and we may therefore deem any course of action based upon such a conclusion to be likewise invalid. But, as Keynes pointed out, we do not expect reasoning which is purely mathematical, to determine questions of an empirical nature, nor indeed to determine questions of a moral nature. Our reasoning is not, therefore, purely mathematical, for we have extended it to moral and other practical dimensions where the rules of reason hold good as rigorously as in mathematics, but are subject to additional constraints and considerations. It is however the practical element rather than the moral element which is fundamental to the issue for, clearly, one can construct fundamentally similar scenarios around the distillation of perfume or the brewing of tea, from which one may perceive that the crucial elements in such situations are the combination of the binary choice and the prior distribution which is implicit in the decision.

Putting aside therefore the fact that, in a practical situation, acceptance of Bernoulli's theorem entails [1] strong assumptions, none of which can be justified from the simple data *(m,n)* defined in the scenario, the aspect of Bayes' solution which is conventionally regarded as the least tenable is the value computed for $\mathcal{P}_H\{(x_1 < P_m < x_2) \mid m, n, P_0(x)\}$ when we assume the uniform prior distribution. However, denoting the assumed prior distribution by $P_o(x)$, we can express a fairly general form of Bayes' equation as:-

[1] Keynes (1921) Chapter XXIX

Version zh25a.doc at 18:36 on 13 July, 2001

$$\mathcal{P}_H\{(x_1 < P_m < x_2) \mid m, n, P_o(x)\}$$

$$= \frac{\int_{x_1}^{x_2} x^m (1 - x)^{(n - m)} P_0(x)\, dx}{\int_0^1 x^m (1 - x)^{(n - m)} P_0(x)\, dx} \qquad (8 - 4)$$

However, if we set $P_o(x)$ equal to the uniform distribution $P_u(x)$, then *(8-4)* reduces to:-

$$\mathcal{P}_H\{(x_1 < P_m < x_2) \mid m,n,P_u(x)\}$$

$$= \frac{\int_{x_1}^{x_2} x^m (1 - x)^{(n - m)}\, dx}{\int_0^1 x^m (1 - x)^{(n - m)}\, dx} \qquad (8\text{-}5)$$

For, when m and n are the totality of our information, it seems intuitively clear that the ratio m/n is the only rational estimate and is therefore the most probable value of P_m, relative to the only information available. However, as it would be perfectly rational and acceptable for a person who was totally ignorant of Bayes' equation to form such a conclusion, it seems that the deeming of m/n as the most probable value is not strictly dependent upon the assumption of a uniform prior, in the sense that, even without Bayes' equation, we would be capable of intuitively reaching that same conclusion. Hence, it seems that the result m/n and the assumption of $P_u(x)$ are interdependent in the sense that the uniform prior is necessary for Bayes' equation to produce the only acceptable answer in the defined situation. It still seems however, as Keynes protests, that it is stretching our credulity too far to ask our assent to a bald and naïve assertion that the probability that P_m lies between x_1 and x_2 is given absolutely by the expression:-

$$\mathcal{P}_H((x_1 < P_m < x_2) \mid m,n) = \mathcal{B}(x_1, x_2, m, n)$$

where, as in *(8-5)* above,

$$\mathcal{B}(x_1, x_2, m,n) = \frac{\int_{x_1}^{x_2} x^m (1 - x)^{(n - m)}\, dx}{\int_0^1 x^m (1 - x)^{(n - m)}\, dx} \qquad (8\text{-}6)$$

However, the reality, which Keynes and many others unfortunately failed to perceive, is that *(8-6)* gives us, not an absolute value, but an estimated value based on the data *(m,n)* and certain assumptions. It should therefore be expressed as, at least:-

$$\mathcal{P}_H\left((x_1 < P_m < x_2) \mid m,n, P_u(x) \right) = \mathcal{B}(x_1, x_2, m, n) \qquad (8\text{-}7)$$

Yet, even if we are prepared to assert that, on the information available in *m* and *n*, the value returned by $\mathcal{B}(.)$ is the 'best estimate' of $\mathcal{P}_H(.)$, we seem at present to have no means of quantifying just how good this estimate actually is, nor indeed of quantifying just how awful it might be. For, in any situation where Bernoulli's theorem is valid, it is mathematically and physically possible that, if P_m has any value other than *0* or *1*, then *m* may have any value in the closed interval $[0 \rightarrow n]$. It is therefore mathematically quite conceivable that, whenever *m* is greater than zero, $P_o(x)$ could be a delta function at any point in the open interval $(0 \rightarrow 1)$. On the other hand, we also know that if $P_o(x)$ were a delta function very close to *0*, the probability, of observing *100* events in a set of *100* trials, would be extremely small. This fact suggests that we might, therefore, integrate the probabilities $\mathcal{P}_R(m|n)$ over all the possible values in the range $0 \rightarrow x_1$, and over the range $x_2 \rightarrow 1$, the sum of which, let us call it Σ_q, could easily appear to be the probability of observing the event *(m,n)* if the true value of P_m were in the interval $0 \rightarrow x_1$ or in the interval $x_2 \rightarrow 1$, *i.e.*

$$\Sigma_q = \mathcal{P}_R\{(m,n)|((0<P_m<x_1) \vee (x_2<P_m<1))\} \qquad (8\text{-}8)$$

whence one could be led to believe that the probability of P_m being within the interval $x_1 \rightarrow x_2$ is given by:-

$$1-\Sigma_q = \mathcal{P}_R\{(m,n)|(x_1<P_m<x_2)\} = \mathcal{B}(x_1,x_2,m,n) \qquad (8\text{-}9)$$

Therefore, it could be argued, the probability of getting the result *{m,n}* from a value of P_m which is outside the interval $x_1 \rightarrow x_2$ is equal to $1 - \mathcal{B}(.)$. Correspondingly, it can also be argued that $\mathcal{B}(.)$ gives us the probability, when the value of P_m is in the range $x_1 \rightarrow x_2$, that we will obtain the trial result *{m,n}*. However, seductive though such reasoning may be, there is within the integral denoted by $\mathcal{B}(.)$, an implicit assumption of a uniform prior probability over the possible values of P_m, and, although complaints about such assumptions in the calculation of direct probabilities, $\mathcal{P}_R(.)$, are rare, their arbitrary nature is just as problematic as in calculating inverse probabilities.

Yet, it is clearly informative to consider the implications of various possible values of P_m. If, for instance, $P_m = 0$, the probability $\mathcal{P}_R(m|n,P_m)$ for any value of $m > 0$ is also zero. If we then progressively increase the possible maximum value of P_m and plot the integrated probability of m events in a trial of size n, *(i.e.* $\mathcal{P}_R(m \mid n, P_m < x)$ *)*, we get the type of result illustrated in figure 8.1:-

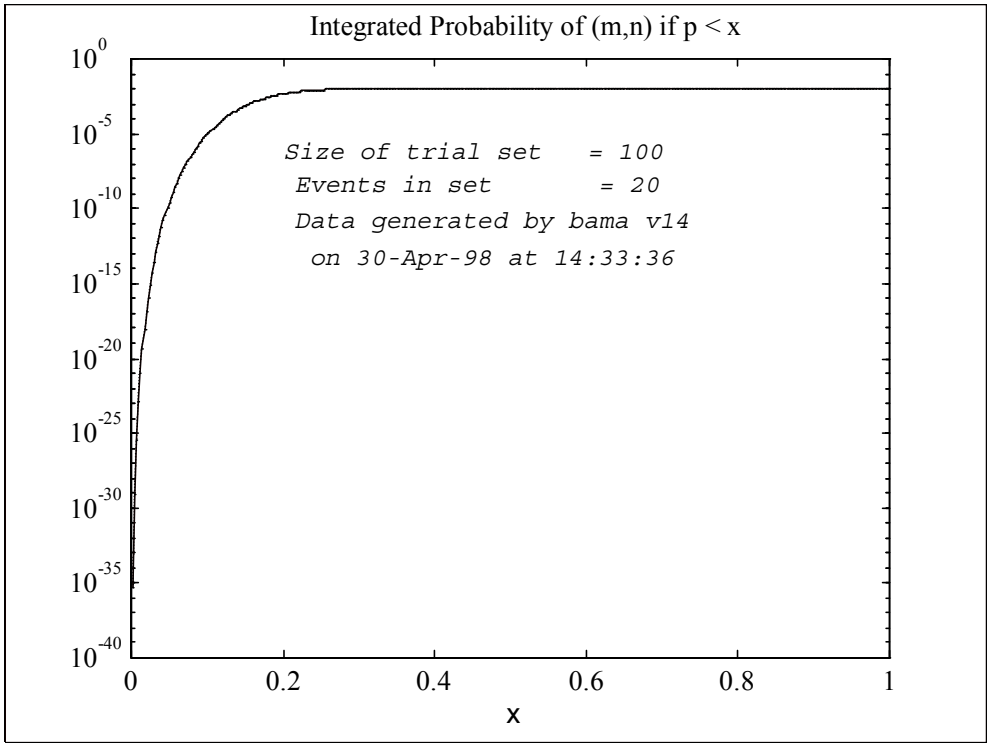


Figure 8.1

On this basis we could imagine a technique by which to measure the 'intrinsic' quality of m/n as an estimate of P_m. For example, we could find a value x_1, such that $\mathcal{P}_R(m \mid n, P_m < x_1) = ½ \times 10^{-6}$ and on this basis, assert that the chance is less than one in two million that an event *(m,n)* would be produced by a value of P_m which is less than x_1. Likewise, further to the right of $x = m/n$, we can find the point x_2, such that the chance is less than one in two million that an event *(m,n)* would be produced by a value of P_m which is greater than x_2. We could then assert that the chance is less than one in a million that an event *(m,n)* would be produced by a value of P_m which is outside the range $x_1 \rightarrow x_2$. Hence we could use the width of the interval $x_1 \rightarrow x_2$ as a measure of the quality of the ratio m/n as an estimate for the value of P_m.

However, while such an approach seems to avoid any direct entanglement with the postulate of the uniform prior, and might seem to have much in common with both Fisher's method of likelihood and the concept of 'confidence limits', closer consideration shows that the process of integration, as suggested above, accidentally introduces an implicit assumption of a uniform prior distribution. This can be seen by considering a case where we know that the different possible values of P_m are governed by a weighting function $P_o(x)$. In such a case, the probability that $P_m = x$ and that there will be m events in a trial of size n, is the product of the 'Bernoulli' probability of m, given x and n, and the probability that $P_m = x$, given $P_o(x)$, *i.e.*:-

$$\mathcal{P}_R (m \mid n, x, P_o(x)) \quad = \quad {}^nC_m P_o(x) \, x^{(m)} \, (1\text{-}x)^{(n - m)} \qquad (8\text{-}10)$$

Hence the probability that P_m will fall in the interval $0 \rightarrow x_1$ and that m events will occur in a trial of size n is:-

$$\mathcal{P}_R (m \mid n, x, P_o (x), P_m < x_a)$$

$$= \quad {}^nC_m \int_0^{x_a} x^m \, (1 - x)^{(n - m)} \, P_o(x) \, dx \qquad (8\text{-}11)$$

If, however, we omit the term $P_o(x)$ from *(8-10)*, this is equivalent to imposing a uniform, unit value on $P_o(x)$, and therefore imposing a uniform prior on the possible values of P_m. Further, as the approach we have just considered has quite a lot in common with the methods of both likelihood and confidence limits, it would appear that those methods also may be not entirely free from a hidden assumption of a uniform prior when considering the possibility that the value of P_m may be in a defined interval. However, the crucial difference between those methods and Bayes' solution is that the numerical value returned by the function $\mathcal{B}(.)$ is, explicitly and directly, an estimate of a probability concerning the true value of P_m. The numerical value returned by $\mathcal{B}(.)$ can therefore be multiplied by a monetary value and used, quite rationally, to compare the probable values or probable costs of alternative courses of action, given the specified assumptions.

In sum, therefore, in real-life situations, probabilities can be computed only relative to our own perceptions and assumptions. Absolute probabilities, if they 'exist', (whatever that may mean), are not accessible to us. The critical case described above is realistic and requires a rational decision**;** that is, a decision which is at least internally consistent and does not contradict the scenario definition. To make such a decision, given the extremely limited information, requires us to make assumptions. Clearly, we should assume the minimum which is needed to reach a rational, self-consistent choice of action. And, although it is conventional to assume without ques-

tion the validity and relevance of Bernoulli's theorem, it is arguable that to do so involves far more in terms of additional structure *etc.* than is involved in the assumption of the uniform prior. We therefore conclude that although Bayes' solution may not be mathematically justified, in the sense that it cannot be rigorously deduced from the axioms normally accepted as fundamental to mathematics, it is certainly justified in situations which require that our decisions and conduct shall be rationally justifiable and consistent.

Chapter 9

The Ruler

Although some readers may find the arguments and conclusions of Chapter *8* fairly reasonable, there remains, as with Bayes' Scholium, an unsatisfying lack of deductive rigour. It brings to mind yet again Keynes' *'children of loose thinking and parents of charlatanry'*. We therefore, in this chapter, make a stern effort 'to do better', taking as our starting blocks the blatant conflict which seems to exist between the parametric demands of Bayes' theorem and the everyday practice of mankind. While we can little doubt or dispute the rigour with which Bayes derived *Proposition 5*, and even less the achievements of Kolmogorov, Jeffreys or de Finnetti in corresponding regards, we equally cannot doubt our ability to weigh a bag without needing to know what it contains, or to measure the length of a stick found on the beach, without knowing the characteristics of the population from which it was taken. The Egyptians built pyramids with astronomical precision, thousands of years before Bayes.

Since there is here *prima facie* evidence that, in a situation of direct measurement, Bayes' theorem is not relevant - for it is certainly not often used, (at least consciously), by carpenters, doctors, nor indeed physicists - it is necessary for us to ask how such irrelevance, or at least neglect, could come about, given that Bayes' analysis seems to be quite general ? We are forced to consider that although Bayes' theorem is undoubtedly a true representation of valid reasoning in certain problems concerning probabilities, it may not be, in its conventional interpretation, the whole truth and there may well be other forms of equally valid, and not-incompatible, probablistic reasoning. Fisher clearly suspected this to be the case in his pursuit of 'fiducial probability', and we may find this to be implicitly the case in our use of measuring instruments such as kitchen scales and builder's rulers. Such devices also raise the difficult point that, if a uniform prior probability has been assumed when measuring, say, a resistance in an electrical circuit, the corresponding prior if we chose instead to measure the potential difference,

would be decidedly non-uniform[1]. However, if we let E_A denote the event that a certain stick has a true length l and let E_B denote the event that its measured length is m, then the algebraic form of Bayes' theorem, *(3-51)*, can be expressed as:-

$$\mathcal{P}_H(E_A \mid k, E_B) = \frac{\mathcal{P}_H(E_A \mid k)}{\mathcal{P}_R(E_B \mid k)} \times \mathcal{P}_R(E_B \mid k, E_A) \qquad (9\text{-}1)$$

where k denotes the supporting knowledge and assumptions. However, both the term $\mathcal{P}_R(E_A \mid k)$ and the term $\mathcal{P}_R(E_B \mid k)$ represent probabilities in which the 'other event' does not appear and, as with their ratio $\mathcal{P}_R(E_A \mid k) / \mathcal{P}_R(E_B \mid k)$, it is rare for any of these terms to play any explicit or acknowledged part in human activities which are concerned with measurement. Furthermore, although probability is, it seems, a vital part of rational behaviour in a world which combines great regularities with great uncertainties, there are many situations where perfectly rational people feel totally unable to supply any value for those terms, as they seem to be conventionally understood. Equally, there is no doubt as to the ability of people to act rationally and in accordance with sensible estimates of the probabilities which they consider relevant. Even more disturbing is the fact that, in many situations, it seems that whatever effort we might make to determine, by observation, and in the best traditions of science, the appropriate values for $\mathcal{P}_R(E_A \mid k)$ and $\mathcal{P}_R(E_B \mid k)$, if these terms are to be understood as prior probabilities over the true length of a stick and its measured length, respectively, we shall be forever unable to do so. We could therefore, almost in desperation, consider the possibility of taking the line of Chapter 8, where we analysed various decisions in terms of their implied assumptions regarding $\mathcal{P}_R(E_A \mid k)$ and $\mathcal{P}_R(E_B \mid k)$. However, in the cases we are here considering, that analysis may not be acceptable to the person making the measurement as a true representation of their position. For, although our analysis may produce a valid analog of their position, we cannot validly assert that the analog actually *is* that position.

An alternative approach is therefore to challenge the common assumptions, ('common' that is, among philosophers and statisticians), that:- *(i)* Bayes' 'first event' equates to a true value being in a given interval, and *(ii)* that Bayes' 'second event' equates to our measuring a value also in an interval belonging to the same set of values as the first event. In contrast to these assumptions, if we actually consider how builders, doctors, physicists and engineers go about the making a measurement, we see that they *(i)* place

[1] Jeffreys, H. (1983), Ch 3, discusses in great detail various techniques for overcoming problems of this type in specific, practical cases but seems to have no general solution.

an object against a ruler, or *vice versa* they place a ruler against an object, and *(ii)* they call out, or write down, a value. Relating this process to Bayes' analysis which we discussed in Chapter 3, the general problem is to evaluate the probability that event E_A has happened, when it is known that event E_B has happened, the answer being given in terms similar to *(3-51)* by:-

$$\mathcal{P}_H(E_A|E_B, ...) = \frac{\mathcal{P}_R(E_A \wedge E_B|.....)}{\mathcal{P}_R(E_B|.....)}$$

$$= \frac{\mathcal{P}_H(E_A|....) \times \mathcal{P}_R(E_B|E_A, ...)}{\mathcal{P}_R(E_B|.....)} \qquad (9-2)$$

In the context of measuring the length of a stick, and given that E_B is an event of the form *'the length of the stick, as measured on the ruler, is x inches'*, and that E_A is an event of the form *'the true length of the stick is in the interval $x \pm \delta x$ inches',* there is an immediate temptation to ask for a prior population distribution, $\mathcal{P}_R(E_A|....)$, to describe the lengths of the sticks to be found in the world at large, or, perhaps, on our part of the beach. Fortunately, the question is so ludicrous as to arouse our immediate suspicion; but, had we formulated the question in terms of the heights of people, it would be all too easy to avoid any hint of suspicion, such is the degree to which the ideas of Quetelet[1] and his followers have saturated our minds. However, it also yields little sense to let the laying of a pencil alongside a ruler constitute the 'first event', and the calling out of a measurement the 'second event', thus creating a situation in which Bayes' theorem allows us to determine the posterior probability that the pencil was laid against the ruler.

However, we have seen several times in previous chapters that, in Bayes' analysis, the order of the events is not fundamental and that, the terms 'first event' and 'second event' are merely names used to denote different events. These names must be taken neither to imply that either event has happened nor that the events happen in any particular order. In principle, a 'second event' can occur without the occurrence of a 'first event'. This possibility is an integral part of the question which is addressed by Bayes in proposition 5:- '*if it is discovered that the second event has happened, and I then guess that the first event has also happened, the probability that I am right is B/p_2*'. However, while there are in these different situations, subtle differences of meaning in the terms used, none of these differences removes the fact that a second event can occur independently of a first event and that,

[1] Hacking (1990) p108

by this fact, Bayes' analysis acquires a generality concerning the correlation of events which it might otherwise lack. Bayes' analysis would therefore allow for my calling out a number before the pencil is laid alongside the ruler, or indeed without there being any object to be measured. In the case of measuring the pencil and other 'given objects', however, we do not require the generality which Bayes allows, and we are not required to consider a situation in which I sit at my desk, I call out a number, and an observer is required to guess whether I was measuring a pencil. The fact that I am measuring a pencil is defined in the scenario. Indeed, the scenario tends to include an axiomatic assumption that the declaration of a measurement is always and only made in response to the input of an object, albeit, as we see later, it is highly desirable, and is arguably essential, that the scenario shall allow for mistakes and mechanical failures. But equally, in the measuring process, there is a sense in which both events are assumed, by the definition of the scenario, to have happened, *i.e.* we accept that we have loaded an object into a measuring device and that the measuring device has subsequently declared a measured value.

Thus, we accept that, in such cases, the ordering of the events is essential. Otherwise the declaration of a value prior to the loading of the object would betoken a non-operational machine. These things therefore suggest that we should investigate the implications of assuming an equal and reciprocal relationship in probability between the length of the object we are measuring and our reading of the ruler. In abstract terms, we are interested in arguments of the form $'\mathcal{P}_R(E_A \mid k, E_B\ ) = \mathcal{P}_R(E_B \mid k,\ E_A)'$. For example, if we are asked to measure the length of a pencil on a regular schoolroom ruler, we shall have no hesitation in answering that the pencil is approximately, say, *155mm* in length and that our answer is accurate to, say, *+/- 2mm*. Thus, leaving ourselves a small allowance for extraordinary errors, we might confidently assert an ability to confine our errors to less than *+/- 2mm* in *95%* of cases. The choice of words is however important, for, bearing in mind the constraints of Bayes' theorem, we must not immediately assert that, in *95%* of cases, the true length of the pencil will be within *+/- 2mm* of the measured value.

Having however claimed that we can achieve a certain performance in measuring the lengths of pencils, we may be challenged to show supporting evidence. One way of answering this challenge is to show that the error limits are well-supported by consideration of the physics involved, as is the

case in radar and similar devices [1]. Another, and not incompatible answer is to put the matter to a practical test in which we use a machine as a judge, and we then measure *1000* pencils. If all our answers are judged to be correct, we can use Bernoulli's theorem and assert, as discussed in Chapter 8, that the probability of this result being produced by an underlying performance worse than our claim is not greater than $1 : 10^{25}$, which would rather seem to support the claim. We may however be asked to justify the assertion that the claimed performance of *0.95* is indeed a 'probability'. Conventionally this requires us to show, for example, that *(a)* it has a non-negative value in the range *0 - 1*; *(b)* that the probabilities of the mutually exclusive outcomes sum to *1*; *(c)* assuming that the probability of error is in each case unaffected by the outcome in every other case, that the joint probability of measuring *'n'* pencils correctly is not less than $(0.95)^n$ *etc.*. The answers to all of which questions are clearly in the affirmative.

This approach can therefore be extended to give a calibrated performance for a measurement process, which we can define in terms of an object which provides a true or 'input' value, followed by the 'addition' of errors in the measurement process, followed by the output of a reading. In the terminology of electronic systems, we have an input signal ℓ a randomly additive error [2] ε, and an observed value m, so that $m = \ell + \varepsilon$. If we also postulate that the variations and uncertainties in the magnitude of the error ε shall have the characteristics which are essential in a variable which is to be manipulated by the conventional probability calculus, we can assert that, given the true value ℓ, the probability of observing a value m is precisely the probability of occurrence of an error $\varepsilon = m - \ell$, that is:-

$$\mathcal{P}_R \{m \mid k, \ell, C(.)\} = \mathcal{P}_R \{\varepsilon \approx m - \ell \mid k, C(.)\} \qquad (9\text{-}3)$$

where $C(.)$ denotes the calibration which has given us the distribution of probabilities over the possible magnitudes of the error ε.

Thus, because the random errors in measuring devices are, over the ranges of objects encountered, either fully independent in magnitude of the true value or vary only slowly with the true value, they can be assumed

[1] See Woodward (1964), Ch 5. The assumption of a 'uniform prior' which Woodward mentions on line 10 of p84, concerns the positions of radar targets and is a necessary condition for the production of a bias-free measurement. This is required to deal with unpredictable target manoeuvres which could otherwise be obscured by a prior expectation.

[2] Often called 'noise' in electronic systems. Originally this was a reference to errors produced by phenomena associated with the 'hissing' noise in a radio receiver.

independent in any given part of the scale[1]. Quantisation errors, particularly in automatic measuring devices, are often of this kind. The assumption of additive independence is also made by many common methods in statistical inference, *e.g.* in the use of the arithmetic mean, the method of least squares with its many variants, and the 'normal' or 'Gaussian' error distribution[2]. The assumption is not, therefore, especially presumptive and is often justified both by the physics of the measuring device and by the approach taken to the overall modelling of errors. That is, provided that the amplitude of the randomly additive error is small compared with the magnitude being measured, even those errors which are a function of the magnitude can be handled by other terms. In other cases, even where errors on individual observations are large compared with the underlying true value, additive independence allows filtering and smoothing techniques to be applied and the true value to be derived to high degrees of accuracy. Such conditions are common to probably all measuring devices and it is likewise probable that a device which does not conform to these conditions will be unusable as a measuring instrument. There is here, therefore, no assumption of any arbitrary or privileged position.

Thus, if we define events such that, in measuring a property of an object, we let E denote the event that the random error has a given value, or falls in a defined range, and D denotes the event that the measuring device declares a measured value, Bayes' theorem gives:-

$$\mathcal{P}_H(E \mid D, k) = \mathcal{P}_H(E \mid k) \times \frac{\mathcal{P}_R(D \mid E,k)}{\mathcal{P}_R(D \mid k)} \qquad (9\text{-}4)$$

Therefore, assuming that the calibration provides a true model of the randomly additive error probabilities, it follows that the calibration provides an objective basis on which to assert a value for the prior probability $\mathcal{P}_R(E \mid k)$. That is, the calibration gives us the probability that the error on any measurement is of a defined magnitude.

Taking the term $\mathcal{P}_R(D \mid k)$, we see that D, the event that the machine declares a measured value, only has meaning in response to the input of an object which is to be measured. We may, for example, have a machine, which is activated by a 'go button' indicating that we have loaded an object, and which 'bleeps' when the measured value is declared. We therefore

[1] Bias and scale errors are covered by other techniques, albeit bias can be included in a probablistic calibration table.
[2] Keynes (1921) Ch. XVII

define our underlying assumptions and data *etc.*, denoted by *k*, to include the calibration data and the fact that an object, which is to be measured, has been loaded into the machine and that the 'go button' has been pressed. Hence, the term $\mathcal{P}_R(D \mid k, E)$ denotes the probability that a measured value will be declared, given that an object has been loaded into the machine, and that the 'go button' has been pressed, and that an additive error defined by *E* has occurred. The term $\mathcal{P}_R(D \mid k)$ simply denotes the probability that a measured value will be declared if it is given that an object has been loaded into the machine, and that the 'go button' has been pressed.

The physical meaning of the ratio $\mathcal{P}_R(D \mid k, E) / \mathcal{P}_R(D \mid k)$ can then be clarified by taking, as a general model, any process which comprises a device for the generation of errors and a device which reports the values observed. Normally, we will expect all values to be reported and therefore both numerator and denominator to have the value *1*. If, however, there is a bias in the reporting mechanism such that, when the error has the value defined by *E*, some, or all, reports are suppressed, the numerator $\mathcal{P}_R(D \mid k, E)$ will have a value less than *1* to reflect this fact. The suppression also affects the denominator, as can be seen if, as in *(3-51d)*, it is expanded to the form:-

$$\mathcal{P}(D \mid k) = \sum_i \mathcal{P}(D \mid k, E_i).\mathcal{P}(E_i \mid k)$$

where the set $\{E_i\}$ represents the possible values of the error.

It is however important to consider the place of the calibration in relation to this model of the meaurement process. As it will generally be necessary to have the reporting device in place in order to carry out the calibration, the calibration will include the bias at that time. The ratio $\mathcal{P}_R(D \mid k, E) / \mathcal{P}_R(D \mid k)$ thus represents the probability that the process is, in this regard, working according to the calibration.

We must however point out that a bias of the form just described is an extremely theoretical notion. That is not to say that such bias could not occur in any realisable mechanism, for it is a trivial matter to write an algorithm which displays such bias. The practical issue is that the biasing process needs to be supplied with the value of the error and it is unusual for the value of an error to be available within a measuring system and yet not used to correct the output. However, in practice, this sort of effect could be caused by anomalous coupling between electronic circuits such that an interfering signal both introduces an error in the measuring device and, coincidentally, causes a blocking of the reporting mechanism. While phenomena of this kind are rarely experienced by the majority of users of electronic devices, they are quite often encountered by engineers who are called

upon to diagnose the reasons for 'inexplicable' random failures in complex equipment.

The phenomenon of 'drift' is however common, especially where digital devices are not available and 'analogue' or continuosly variable devices have to be used as 'transducers' to convert a representation of a magnitude from one physical form to a different form, such as the conversion of a pressure to a voltage. Such devices are often subject to many different physical influences, *e.g.* microscopic changes in shape due to the effects of steady forces over a period of time, and, as a result, the performance of the device changes. When using such devices, it may therefore be necessary to adjust the calibrated error-distribution to allow for drift since the time of calibration, or to allow for operation at a different temperature, *etc*.

Thus, summarising the symbols, we have:-

$C(\,.\,)$	the calibration data
$C(e)$	the calibrated probability of an error e
$C'(e)$	the adjusted probability of an error e at the time of measurement
E	the error has a magnitude e
D	the observing device declares an output value
k	other assumptions and data, including the calibration $C(\,.\,)$
ℓ	the true value
m	the measured value
$\mathcal{P}_R(W \mid k)$	the probability that the measuring system is working as calibrated,

and the probability that an assertion E is true, is given by:-

$$\mathcal{P}_R(E \mid k, D)$$

$$= \mathcal{P}_R(E \mid k) \times \mathcal{P}_R(D \mid k, E) \,/\, \mathcal{P}_R(D \mid k)$$

$$= C(e) \times \mathcal{P}_R(W \mid k) = C'(e) \qquad (9\text{-}5)$$

We therefore have a situation in which the measured value m is a random variable, formed by the addition of a fixed, but unknown, governing value ℓ and the random error e so that $m = \ell + e$. Hence, as a fact of arithmetic, $\ell = m - e$. If, therefore, we are told the value of m, and we are told the probability $C'(e)$, then the probability of the hypothesis that ℓ has the value $m - e$ is:-

$$\mathcal{P}_H(\ell \approx m - e \mid k, m) = C'(e) \qquad (9\text{-}6)$$

which accords precisely with common practice.

That is, where we are dealing with random additive errors, which are independent of the sample value, we can, by using Bayes' theorem, and without assuming any prior probability over the true value, achieve the result for which Fisher strove so hard and so long [1] *i.e. 'The possibility of making exact statements of probability about unknown constants of Nature ..... (and) ..... making a complete specification of the precision with which such constants are estimated .....'.*. It is also important to note that the challenging case of the median [2] can be viewed as a form of measurement, subject to an additive error, or uncertainty, which is independent in its magnitude from the sample value. In its bare bones, the definition of the median μ constitutes an *a priori* calibration which tells us that, if we observe a value $m = \ell + \varepsilon$, then there is a probability of *0.5* that $\mu \geq m$ and an equal probability that $\mu \leq m$. Thus, provided that the errors on each observation are randomly independent, if we are given a pair of observed values m_1 and m_2, then, exactly as Fisher asserted, there is a probability of *0.5* that μ lies between m_1 and m_2. This is, however, achieved without any change of logical status between fixed quantities and random variables: once ℓ is given, it is fixed and does not vary; the independent random variable remains ε, and the essential statement of probability remains that of *(9-3)*. Given *(9-3)*, the assertion *(9-5)* is a simple statement of arithmetic. If we are told there is a probability of *0.95* that a random variable ε has a value e, then there is correspondingly a probability of *0.95*, given a value z, that the random variable $z+\varepsilon$ has a value $z+e$. It also follows that if a statistic $T(m_1 ..... m_n)$, derived from observed values, is distributed about the true value ℓ in a manner which is independent of ℓ, then our knowledge of that distribution constitutes a calibration. This allows us both to assert a probability concerning the true value and to avoid the provision of any prior distribution of probability over that true value. Hence, while there are here superficial similarities with Fisher's fiducial argument, there are also profound differences. It is sad that, although Fisher was pursuing a thoroughly valid objective with the fiducial argument, he failed to find a valid way of achieving that objective.

We conclude, therefore, that, in using Bayes' theorem, we have a remarkable freedom to define 'first event' and 'second event' to suit our purpose and that the chosen definitions can have marked effects upon our ability to supply values for the required parameters. That is, always provided that the chosen events and the 'trials' by which they are determined are such

---

[1] Fisher (1956) p60. See also Ch 7 above

[2] See Ch 7 above.

that valid probabilities can be assigned to all the possible outcomes of the trials[1]. As we show below, examination of alternative outcomes can be important and revealing. It must also be emphasised that the independence of the calibration *C(.)* from the value of ℓ is critical in the above argument and cannot be assumed to be true in general. It is, in particular, not true in the case of Bayes' experiment[2]. There are however many cases of measurement in everyday life where this assumption is sufficiently valid for practical purposes and allows people to operate very successfully in accordance with, but without ever referring to, or knowing of, Bayes' theorem.

Further, it should be noted that, in the above approach, where we are given independent additive errors, the prior distribution relates only to the uncertainties and errors in the measuring process, and is in no way related to the object being measured. Thus, we have disposed of any need to consider 'populations of sticks' when we wish to measure the length of the stick which the dog brought back from the beach. Likewise, we have eliminated from the process of measurement the need for invariance of the prior distribution when the parameter being measured is subject to a non-linear transformation, as when, for example, we measure potential difference where we have previously measured resistance. That is, the only prior information we require is that which concerns the performance of the measuring device at the point of measurement. Transformations of the input signal prior to the act of measurement are, by virtue of additive independence, irrelevant to the probability of error in the measurement. This dissolves a problem in many important applications for Bayes' method which, though long-ignored in practice, has been deeply disturbing to those more concerned with matters of principle. Jeffreys[3], for example, was strongly concerned to solve this problem and was delighted with his own discovery of certain distributions which avoided the problem. More recent authors[4], aware of the difficulty and aware also of the force in successful common practice, have had to satisfy themselves, (if not perhaps all their readers), with arbitrary rules such as accepting Bayes' approach where the unknown parameter is a 'random variable with a prior probability density function', or, where 'there is an unknown true value', of ignoring the issue of probability and simply using the 'maximum likelihood' estimate.

---

[1] For an understandable treatment of these requirements, see Howson, C. and Urbach, P., (1993) esp. Ch.2.

[2] See Ch 13

[3] Jeffreys, H. (1983).

[4] *e.g.* Van Trees (1968); Bar-Shalom and Fortmann (1988)

It is therefore satisfying to find that, in the case of direct measurement, the use of calibrated instruments allows us to assign probabilities to the possible values of unknown parameters and thus reconcile common practice with the fundamental principles in Bayes' theory of probability.

Further and more complex aspects of the dimensional measurement problem arise in radars and other devices which may have to make measurements of weak signals which may be moving through a noisy environment. Such issues are often addressed by repeating and integrating the observations, in order to improve the relevant probabilities. The application of Bayes' methods to these situations is the subject of Chapter 13.

One *caveat* remains to be addressed. If the calibration of a ruler, or analogous device, is determined by empirical tests and histograms of errors, we will often have a situation where the calibration is effectively based on a 'Bayes trial' from which we have to determine the probability that the underlying frequency of errors in a defined band, falls within defined limits. Although this means that our investigation has turned a full circle, we are not exactly back at our starting point, for we are at a significantly greater depth and the resolution of this problem forms the central issue of the next two chapters.

Chapter 10

The Individual

Although we have, in previous chapters, made some useful progress in the matter of ruler-like devices, where we can reasonably assume that the error-probabilities are effectively independent of the magnitude we are measuring, we have seen also that Bayes' experiment[1] falls not into that category, yet is often fundamental to the process of calibration. We are therefore forced, yet again, to confront the issue of the prior distribution: albeit, having shown the population prior to be of no relevance in the case of a direct measurement, we may, not unreasonably, wonder just how relevant, if at all, it may be in the case of measuring P_m, *i.e.* an unknown probablistic frequency[2]. Also, we are sceptical as to whether, in the case of measuring an individual value of P_m, it can ever be valid to introduce information derived from other individuals. However, that doubt seems immediately to be dispelled if we simply consider a population within which there are just two values of P_m, taking *0* and *1* as an extreme case. If a member is selected at random from such a population, a single trial immediately determines the P_m value of the selected member. The objective relevance of the population, considered as a prior distribution, appears to be indisputable. To take an even more extreme case, we can consider a population which is defined and selected to consist entirely of people called Smith. Within such a population, the probability of finding anyone called Jones - the possibility of mistakes being excluded - is zero.

We also have to question the means by which we can acquire knowledge of a prior distribution. If we were of a doctrinaire disposition, it would be tempting to lay down an over-arching rule that valid knowledge of a prior distribution can be acquired only by knowing the process by which the prior is created, *e.g.* by knowing that a ball is thrown randomly onto a smooth and level table. Indeed, in his *1921* paper, Fisher argued[3] that:- *'Such a problem is indeterminate without knowing the statistical mechanism under which different values ... come into existence : it cannot be solved from .... any*

---

[1] Ch 4 above.

[2] See *Notation* and also Ch.5 above.

[3] Fisher (1921) in a *'Note on the confusion between Bayes' Rule and my method of the evaluation of the optimum'*

number of samples. There seems however to be neither proof nor justification for such a strong assertion. Perfect knowledge of the distribution can be acquired by measuring every individual within the population. But where we are dealing with hypothetically infinite populations or we are constrained, as in Bayes' experiment, to observe only a sample sub-set, then there would seem to be serious doubt of our ability to deduce, from samples, valid limits for the true distribution.

To progress the matter, we consider a scenario where, in samples, say of blood, taken from randomly selected members of a population, the proportion, P_m, of cells of *type-M*, falls outside limits $x_1 \rightarrow x_2$ only rarely. The actual distribution of values outside the known concentration in the band $x_1 \rightarrow x_2$ is assumed to be uniform. For example, we may have a population, where the normal limits of P_m are between *0.29* and *0.31*, and the total probability of a person, (selected at random *etc.*), being outside these limits is only one in a thousand, *(fig 10.1)*:-

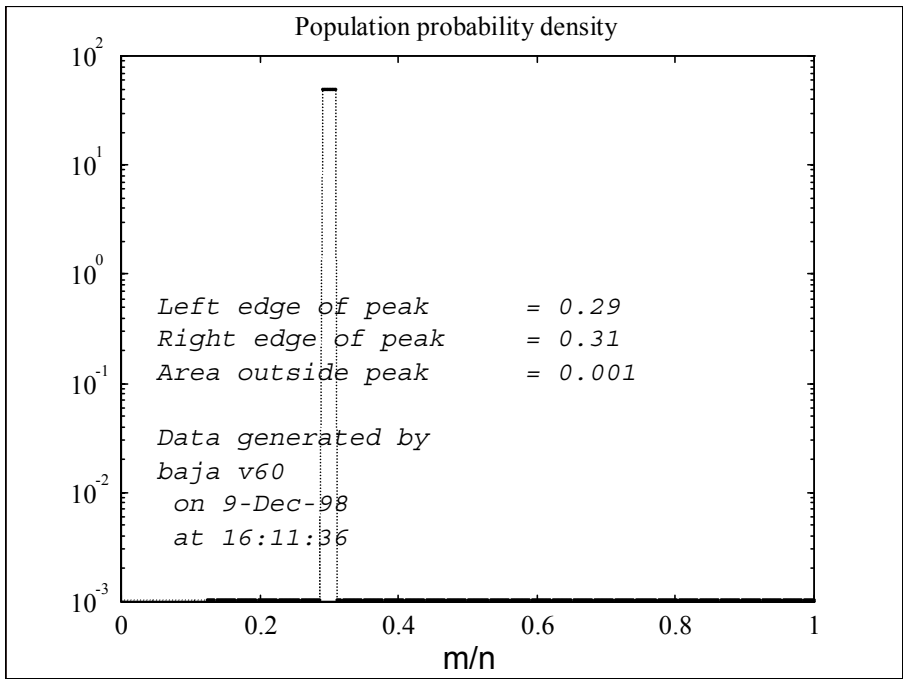


Figure 10.1

Suppose then that a sample taken from a patient shows that, of *100* cells which have been tested, *50* cells are of *type-M* and we wish to compute the resultant probability distribution with respect to the possible values of P_m for this patient. If we use Bayes' theorem to combine the data from the sample taken from the patient with the prior knowledge of the population, we find the post-trial distribution shown in Figure 10.2. In such a

situation, we may feel well-advised to seek a larger sample and Figure 10.3 shows the result of increasing the size of the sample to *500* cells with a corresponding attenuation of the prior so that the probability of P_m in the region of the population norm is some ten orders less than that of a value in the region of *0.5*.

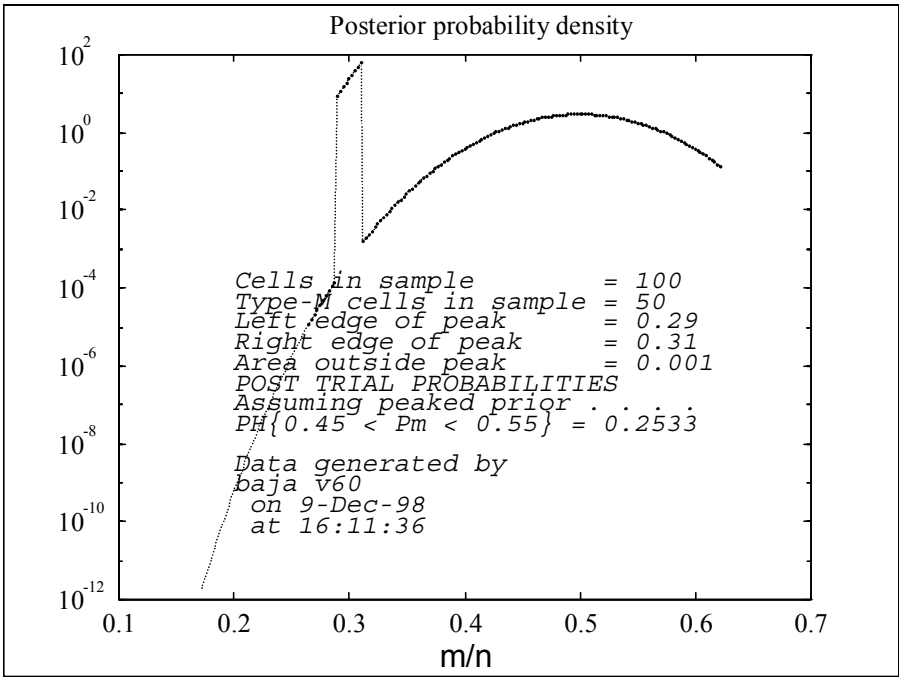


Figure 10.2

But the issues go much deeper than merely whether we need a bigger sample, for assuredly, if we make the sample big enough and the ratio *m/n* remains at *0.5*, the peak of the posterior will also converge ever more closely on that value, albeit without ever quite getting there. The deeper point concerns, as Keynes warns[1], the loss of contact with common sense in a maze of mathematics. For, with scant justification, we are applying a prior probability distribution, derived from a sample of a population, to measurements, taken on a specific person who was possibly not even a member of the sample on which the population statistics were based: even worse, we are treating the measurements made on this individual as if they were part of a single statistical continuum with the rest of the population, and in which the population takes on the rôle of a governing process over the individual.

[1] Keynes (1921), p389

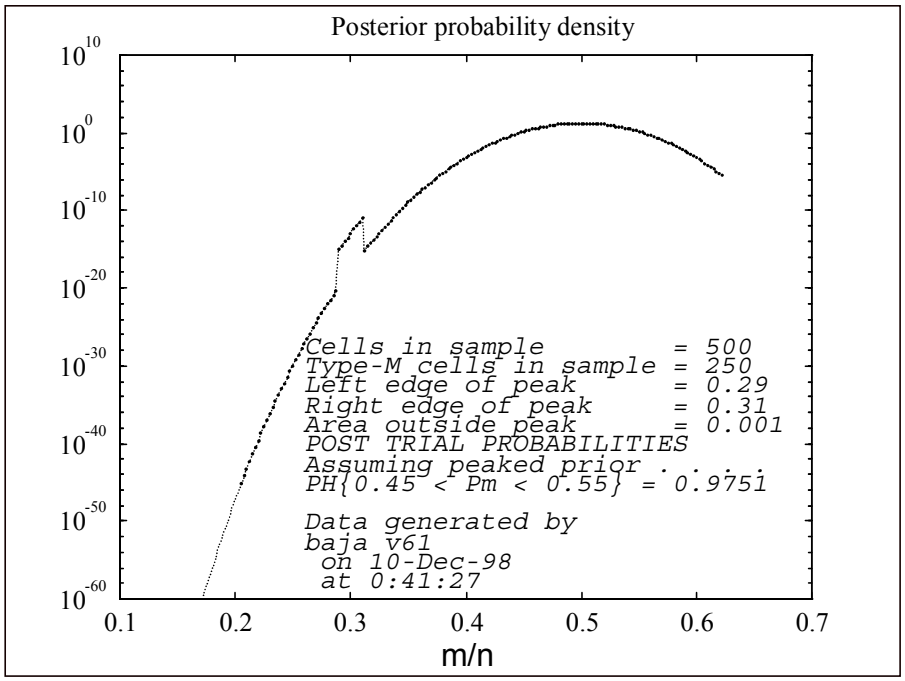


Figure 10.3

These are ghastly errors. The statistics of a population do not affect our ability to count accurately, nor do they affect the variability of *m/n* in samples taken from a given person: the taking of successive samples from that person is a process utterly distinct from that of taking successive samples from different members of the population. The fact that we may have before us a person whose *M-count* is measured as being well above the norm will never give us reason to assess the true count as being somewhere between the norm and the value we actually counted. The statistics of the population are of massive irrelevance. If we need a better answer, we increase the size of the sample, or we look for other information relevant to the object under examination; we do not need the epistemological lunacy of looking at other people in order to improve our estimate of the level of *M-cells* in the person before us[1].

It is important, therefore, that we seek an explanation for the clash between, on the one hand, the views of authorities such as Boole, Keynes, Fisher, Fine; and, on the other hand, the common practice of objective measurement. For the great authorities, differ though they may on many matters, are united in their outright rejection of the assumption of the uniform prior: conversely, they are united in affirming relevance of the prior distribution,

[1] See e.g. Wonnacott (1972), p 367. This is an excellent introductory text which clearly *describes,* without attempting to justify, the basic techniques of contemporary statistics.

when it is indeed known. Yet, an important attribute of the uniform prior is that it has no bias against the unexpected and, at every point, gives the maximum weight to the observed data which is rationally consistent with *a priori* ignorance as to the actual value of the parameter in question. This can be particularly important when we need to be alert for the occurrence of events which are rare, but not impossible. Furthermore, in the common act of measurement, it is widely regarded as serious malpractice to allow external prior probabilities to impede the objectivity of the metric process. Multitudes of engineers, builders, scientists, doctors and navigators, in every minute of every day, use counters, and similar instruments to estimate probable values of attributes and, indeed, probability limits on those values, with never a glance in the direction of an external prior distribution. Counting heads does not, decide such arguments: but it gives us good reason to pause and reflect.

In the context of Bayes' essay, which seems to have exercised enormous influence on perhaps every author who has subsequently examined the issue, the fault seems, metaphorically speaking, more geological than moral, in that all seem to have failed to perceive the discontinuity between the population and the individual [1]. For, even in the case of Bayes' experiment, once the first ball has been thrown and has come to rest, we are in a situation, essentially, of measuring the distance of a particular ball from the edge of the table. To that process of measurement, the prior distribution of probabilities over the possible positions in which the first ball can come to rest, is irrelevant. The root of the matter is that, once an individual has been chosen from a population, the distribution over the population, of the values of the attribute in which we are interested, is of no relevance to the measurement of the value of the attribute within the selected individual. Within the individual, the value of the attribute is now fixed [2], and any statistical variations will be induced by the metric process. The act of selecting a specific individual causes the question of probability distributions to shift from the population from which the individual was selected, to the process which is used to measure the value of the attribute within the individual.

Yet we have the case, illustrated above, where a prior distribution over a population of two members with values *0* and *1* respectively, enables a single trial immediately to determine the value of the selected member. Something is paradoxical, or simply wrong in our thinking. A clue towards

---

[1] Clear acknowledgement, in conversation, of this independence is however common among many educated folk in circles such as teaching, engineering and medicine.

[2] *i.e.* putting aside the matter of a time-varying attribute, which is not here at issue.

the resolution is however obtained if we also consider another population of two members, but this time where the values of P_m are extremely close to each other, say, *0.4999* and *0.5000*. If one of these is selected at random, the size of the trial needed to determine to any reasonably high degree of confidence which was selected, will clearly be large. The relevance of the population as a prior in the determination of the value of an individual is now small and, as the difference between the values of P_m diminishes, so does the significance of the prior. Indeed, if we consider members of a population which are arbitrarily close, then we immediately see that only the assumption of a bias-free uniform prior will be capable of discriminating between those members. Which suggests that, in our conventional approach to such issues, we may be taking too much for granted and it seems right to wonder whether, or under what conditions, a 'population prior' is ever objectively relevant to an individual case.

This, however, raises a further serious issue in connection with populations, namely that of defining the population to which an individual shall be deemed to belong. For there is generally nothing essential about an individual being a member of a given population, and there is no limit to the number of different populations we can define, of which a particular individual is a member[1]. It may therefore be a matter of arbitrary choice as to which population we use on a particular occasion: a fact which casts yet further doubt on the use of population data as a prior distribution within a metric process which is directed at individuals.

These questions force us, therefore, to ask where is it, and how, that prior knowledge of a population can be so powerful in providing knowledge of an individual ? The answer, we would suggest, is that the power arises when the population prior identifies classes, within which certain characteristics of members are confined to highly specific values, or to narrow bands of such values. The process of measurement then allows us to identify an individual as being, to some degree of probability, a member of a certain class. Given such an identification, we can then use the precise prior definition of class membership to determine any number of attributes and values. Similar considerations apply when we know that objects in a given set are limited by well-defined bounds, *e.g.* we may know the lengths of the longest and shortest pencils in a given bundle. In such cases, the limits clearly apply to any pencil selected from the bundle.

In relation to Bayes' experiment we therefore consider a scenario where prior information tells us that a certain population comprises two

[1] *cf.* Popper (1972), p.210, who makes a similar point.

classes of objects *X* and *Y*. In a certain type of trial, all members of *X* have a probability of success, $P_m(X) = x$, and all members of *Y* have a corresponding probability of success $P_m(Y) = y$. Suppose then that objects are picked randomly from this population and we are required to judge, on the basis of an *'m/n'* trial in each case, whether the object is from *X* or from *Y*. This situation is, in many ways, similar to that of the critical case discussed in Chapter 9 and we can adopt an appropriate strategy. The point special to this case is however that, having reached a decision, the *a priori* information tells us, if our decision is correct, the exact value of P_m. This is, of course, a generalisation of the cases illustrated above, where we considered populations of just two members with P_m values of *0* and *1* in one case and of *0.4999* and *0.5000* in the other case. Superficially, very superficially, therefore, all these cases appear to show that *a priori* information concerning values of P_m within a population can help us achieve better estimates of the P_m values than could be achieved entirely on the basis of the observed *m/n* results. At a deeper level however, the truth is that these 'better values' are not achieved by combining according to Bayes' theorem, the observed *m/n* data with an *a priori* probability distribution, for the distribution scarcely enters the picture. Rather, we are here using the observed values of *m/n* in order to make a probablistic identification between the object being observed and membership of a class which is know to us *a priori*. Indeed, many radar surveillance systems use precisely this type of procedure when they exploit the relationships between:- *(i)* the observed 'blip/scan ratio' on a particular radar target, a parameter which equates exactly to an observed *m/n*, and *(ii)* the radar performance parameter denoted by P_d, which corresponds exactly with our P_m, and *(iii)* the aspect angle and 'echoing area' of the target, and, finally, *(iv)* pre-defined libraries of characteristics for known types of aircraft, ships, *etc.*. On this basis it is possible, in certain cases, to proceed by entirely valid inference from an observed blip/scan ratio to a classification which will have a very high probability of being correct.

Similarly, in many other real-life situations, previous experience and memories provide libraries of classification data such that measurements on an individual allow us to identify, with some degree of probability, membership of a class. We move from measurement to identification: to diagnosis. The *a priori* data which are used to support the identification are not used in the measurement process. Were the data to be so used, it could cause serious errors which would not otherwise occur. Prior to an individual being selected, randomly, from a population, the relative sizes of classes within the population clearly determines the probability that an individual will be selected from any given class. After a random selection has been made, the

probability, relative to the population data, that the individual will be found to be a member of a given class is again determined by the relative sizes of the classes. But the size of a class does not influence, the probability, relative to the measurement data, that an individual, having been selected, is a member of that class. We do not balance the size of the class against the improbability of a possible magnitude in deciding the probability of membership. To do so is to make the size of the class implicitly equivalent to a measurement made upon the individual, which is decidedly not true. Indeed, the use of population data as prior information on an individual, must, almost always in a measurement process, degrade the accuracy of the process and cause the resultant almost never to converge on the true value.

This line of argument leads to yet another point. That is, the metric process which does best in each individual case must necessarily do best in the aggregate of such cases. This view however seems flatly to contradict the common experience that actions decided by reference to a population prior will often perform better in a real-life situation than will actions based on attempts to measure precisely each individual. For example, if we have a population in which one million objects are of *Type-X* and only one object is of *Type-Y* it will often be a waste of time to even acknowledge the existence of the *Type-Y* object. But this is by no means always true: it will not always be a waste of time to look for a single diamond which has been lost among a million fragments of glass: all depends upon the value of the diamond, and the probable cost of finding it. We have, here, a door to a new world. Bayes' theory of probability, stemming largely from his choosing to define not 'probability in itself', but rather a rule for the determination of its magnitude in any situation, *i.e. the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening*[1], leads automatically to worlds in which we shall try to 'optimise' our conduct in accordance with our scheme of values and the limited resources which are available[2]. Often, but not always, we shall be operating over a whole population: the population, that is, becomes an individual object and it becomes the convolution of our scheme of values with the probability distribution over the population which determines the overall degree of success. Yet, fundamentally, our knowledge of a population depends upon our ability to measure the distribution of probability over the possible magnitudes in each individual case, and to converge on certainty wherever necessary.

---

[1] Bayes' *Definition 5*.
[2] Polak, (1997)

Even so, this view still seems to conflict with Bayes' theorem regarding the joint occurrence of events in that, if we consider direct probabilities, the expected frequency with which combined events such as 'E_1 = *individual picked at random' and '*E_2 = *M-count in range* $x_1 \rightarrow x_2$' will occur, is given correctly by Bayes' third proposition [1]. Therefore, in that Bayes' *Proposition 5* is crucially dependent upon *Proposition 3*, it seems clear that there must be a sense, or a context, in which the population validly enters the picture when we reason inversely from a finite set of observations to the governing value of probability. This is certainly true when the population itself is the individual object with which we are concerned and we can validly equate 'probability' with the expected frequency with which we will find an hypothesis concerning a 'first event' to be true, taken over the whole population. This is however in stark contrast with an equally valid view of a probability as a 'degree of reason to believe' an hypothesis to be true in an individual case; for there is a wide difference in meaning between 'believing' in the case of the individual and 'expecting' in the aggregate. Yet there is no fundamental conflict, for even in the case of the rational financiers [2], if they make the effort to get the answer right in each individual case, they will inexorably get it right in the aggregate [3]. In contrast, procedures which base decisions about individual cases on population statistics have no such guarantee. In many practical situations, however, where time and resources are limited, financiers and others may be coerced, openly or implicitly, into the adoption of constrained optimisation procedures which will entail operating on the basis of population statistics rather than on individual assessment. This will often produce acceptable overall results at an acceptable cost, for a time, but will also, with a probability approaching certainty, cause a number of failures, and, given enough time, a set of exceptional cases which is large enough to make a catastrophe. In contrast, procedures which treat each individual as a unique case, can make the probability of catastrophe vanishingly small, and can be applied equally where there is available to us no information whatsoever concerning any relevant population.

In this context, it is however, extremely disturbing to contemplate the procedures adopted by certain British 'Health Authorities' who arbitrarily exclude people, on the basis simply of their age, from certain forms of medical treatment. Certain authorities arbitrarily exclude, for example, physio-

---

[1] See above, Chs.2 and 3

[2] See Ch 6 above.

[3] These assertions are, strictly, subject to the assignment of appropriate degrees of probability. We can however, in principle, achieve any required degree of certainty by raising the probability threshold in each individual case.

therapy for people who have suffered a stroke and are over the age of 65. This is, of course, equivalent to assigning a dogmatic prior value of precisely zero to the probability of improvement in such persons who may be, and often are, restored if they can afford private treatment, or move to an area where no such policy is in force. The moral irony can however be further illuminated by considering the fact that, under such policies, a 65 year old nurse or doctor would be excluded from treatment while, in contrast, a 64 year old criminal would qualify. It is therefore perhaps fortunate that advances in our understanding of the mechanisms of disease may progressively cause such decisions to focus more on the individual - including the knowledge and judgement of the individual patient's own doctor - and less on some arbitrary subset of the population to which patients are assigned. We must, however, emphasise 'perhaps', for it is terrifyingly easy to substitute a disguised, but politically acceptable population for one that is obvious and politically unacceptable. As Ridley points out in the context of the human genome project[1]: *'To test somebody for a disease that is incurable is dubious at best ..... the prediction might be wrong'.* But, Ridley also goes on to say: *'The genome will revolutionise medicine by forcing doctors to treat the individual, not the population'.*

Our own criticisms should not, however, be taken to imply that decisions to exclude certain people from treatment, simply because they belong to an easily identified subset of society, are taken lightly. Nevertheless it is a fact of political and legal reality that, while discrimination against certain subsets of the population is apparently deemed acceptable by British, and probably many other governments, the identification for such purposes of certain other subsets would be blatantly unacceptable and decidedly illegal.

That the issues here raised are difficult, is plain to see; but, there are no grounds to believe that they would remain insoluble if addressed by persons of adequate education, intellect and political acumen. Yet, even from our own, narrowly theoretical point of view, the questions raised by such practice are profoundly important because of the value of the expectation which they imply, for the implicit use of the dogmatic prior and for the exclusion of individuals - doctors and patients - from the decision making process.

Nevertheless, putting aside all considerations of morality, political expedience, and legality, our analysis shows that, as a matter of simple logic, the distribution of an attribute within a population does not provide a valid basis for the assignment of a prior probability in the case of an individual selected from that population. When we need a prior probability for the

[1] Ridley (1999a). See also Ridley (1999b).

individual, it is elsewhere we must look and to that matter we shall turn in Chapter 11.

Before doing so, however, we have to face that fact that our discussion of the individual has raised some important questions concerning the calibrated ruler, for that technique clearly involves a transition from a frequentist view of random error-probabilities to a probability concerning the true value of the individual object which has been measured. The calibration may tell us that *90%* of errors are within given limits, *e.g. +/-2mm*, but it is not clear how this allows us to assert a *"90% confidence"* or *"degree of reason to believe"* that a specific object has a true length within *+/-2mm* of the measured value. Conceptually it seems that when we use a calibrated device under properly controlled conditions, the calibration is passed-on to any appropriate object[1] which is subsequently measured, precisely as a degree of confidence or of 'reason to believe'. This appears to be the implicit assumption that supports the use of calibrated devices but neither its justification nor its implications are immediately obvious. If we ask the meaning of an assertion such as:- *"There is a 90% probability that the true length of the pencil is within +/-2mm of the measured value"*, one answer is given by Bayes' definition of probability, namely that *90%* of the prize-value is the amount at which one ought to value the expectation in a trial where the pencil is measured against the calibration standard [2]. This may seem a vacuous response but, translated into an operational context, it can easily be given effect in terms of decisions about the allocation of resources and the prudent disposition of reserves.

It is also useful to consider that, while random selection of a pencil, even from a known set, can, only in special cases, tell us anything about the length of that pencil, a calibrated measurement despite its random error, can often tell us a great deal. That is, when we make a measurement, we generally expect that a reduction of uncertainty will occur. Thus, if a pencil has been selected from a set, all of which are between *170mm* and *180mm* long, knowing the distribution of lengths within the set gives us no reason to believe anything about the length of the selected pencil, other than that it must be between the known limits. However, if we now measure the length of the pencil, using a ruler which is known by calibration to be accurate to *+/-2mm*, we shall effect considerable reduction in uncertainty. This is true to an even greater degree in cases where we have no knowledge of the

[1] A piece of string might not be appropriate in a device designed to measure the length of rigid, pencil-like objects.

[2] Subject to the wager being within a prudent limit. *cf.* the St Petersburg problem in Ch.6

population or limits on the object which has been selected. This is, in effect, the reasoning which supports the use of calibrated devices in a great deal of medical, engineering and navigational practice, where, after calibration, we assert that a properly-working device is accurate to $+/-\varepsilon_{max}$, and therefore any actual error ε is governed by:-

$$\mathcal{P}_R(\varepsilon > \varepsilon_{max} \mid k, C(.)) = 0$$

whence the measured value m and the true value ℓ, must satisfy:-

$$\mathcal{P}_H((m - \varepsilon_{max}) \leq \ell \leq (m + \varepsilon_{max}) \mid k, C(.), m) = 1 \qquad (11\text{-}1)$$

Even so, we have to face the inductive chasm: *i.e.* however large the calibration trial, it can never 'prove' the impossibility of errors greater than those observed. The practical answer to this problem, *pace* Boole and Fisher[1], is, first to declare the assumption and, second, to allow for the probability that the device may not be working as calibrated. If we do not make these, or equivalent declarations, rational decisions and actions become impossible and we open the gates to anarchy and wild guesswork. There may indeed be no possible analytic justification for these assumptions. The justification is plainly pragmatic: given the assumptions and the corresponding disciplines, we can design and build things that work. Such 'things' range from wheelbarrows to models of the observable universe. Given calibrated rulers, these things can be made to work reliably; otherwise they do not.

The justification of the transition from a frequentist calibration to a probability concerning the true value of an individual object remains a more difficult issue. If we consider just a single observation, it is, at first sight, hard to say exactly why it is appropriate to take the object as a fixed quantity and the error as a probablistic variable. If the object has been selected at random from a set with a known distribution, its standing might seem to correspond exactly with that of the error on a single observation - also selected at random from a known distribution.

However, if we think of the observations rather than the errors, we see that the measurement process directed at an individual object comprises, in principle, a set of measurements and our prime interest is in the properties of that set. When we make just a single measurement, and infer probabilities concerning the true value of the object, we are using the single measurement, together with the calibration, to estimate the properties that we would expect to find in a larger set of observations. ***The inference here is***

[1] Fisher (1956) pp23-4.

from the individual to the population. We know from the calibration the general properties of the error-set and we are using a single observation to estimate its location.

Thus each calibrated observation is specifically informative about the object being measured and can be integrated with other observations of that same object. It follows that the calibration is specifically informative in each individual case. When using a calibrated instrument, independent measurements of a given object are cross-informative. Each measurement both improves our knowledge of the object and provides information about the magnitudes of the errors on the other observations. That is, we use a computable function of the observations - a smoothing or filtering function - which closes progressively on the true value as the size of the set increases[1]. In the process of measurement, the error-set is a population-object about which the calibration provides valid prior information.

This is in marked contrast to the objects which are being measured, where, in general, an error in the measurement of a selected object cannot be in any way improved by involving random selections of different objects[2]. It is also in marked contrast to the population distribution, which, as we have seen earlier, is generally misleading in any individual case and actually prevents the metric process from ever converging absolutely on the true value.

Thus, while the distribution of a population from which an individual object has been selected, gives us degrees of reason to ***expect*** the object to have a property of a given magnitude, it gives us no reason to ***believe*** that the magnitude actually has that value. In contrast, the calibrated distribution of errors gives us no reason to expect anything about the magnitude of an object prior to an observation, but, following an observation, it gives us a great deal of reason to believe, or not to believe, that it is of a given magnitude.

In sum, the transition from a frequency-based calibration to a degree of confidence in a measurement is achieved by equating our degree of confidence to the frequency with which the measuring device has been shown, by calibration, to perform to the specified level. It seems that, in practice, we

---

[1] It may be possible to show that there are theoretical distributions for which no smoothing function can exist. It may, equally, be possible to show the opposite. We are not however aware of any such work.

[2] There are however situations where the deliberate introduction of randomly additive errors of a known distribution can provide, *via* a filtering function, significant improvements in accuracy.

regard calibration of an instrument as akin to assessing the credibility of a witness. It is a characterisation of the device, which then gives us "degrees of reason to believe" in the probabilities of different errors. This assumes either stability in the characteristics of the measuring system, or that we can know the probability of drift. Such assumptions are however widespread in the gathering and application of scientific knowledge and it is not unreasonable, therefore, to make them here.

Chapter 11

The Valid Prior

Although there are, as we have seen, many instances where a population can be regarded as an individual object, there are also many cases where our interest is in an attribute of an individual, independent of any population to which it may belong, and on which we may have no prior data. Where the attribute of interest is a dimensional parameter, measurable on a conventional scale, and the uncertainties *etc.* are independent of the magnitude of the attribute, we can assess the probability distribution over the possible magnitudes of that attribute by the logic of the calibrated ruler which we discussed in Chapter 9. However, when the attribute of interest is a probablistic frequency, as with the magnitude of P_m in Bayes' experiment, the uncertainties are not independent of the value of P_m and we again have to face the unresolved issue of the prior probability. Further, it is unfortunate that, in Bayes' essay, the unacceptability of the population data as a prior distribution over an individual is obscured by Bayes' use of the table, upon which, the value which is to be measured, is determined by the throw of the first ball, followed by the repeated use of an effectively identical ball as the measuring device. The 'first event' is that the first ball generates a value, or position, P_m relative to the frame of the table on which the later throws take place and the 'second event' in Bayes' analysis is the result that, in n further throws of the ball, it comes to rest to the right of the position corresponding to P_m on m occasions. However, the fact that our ability to determine the position of the first ball depends in no way upon our knowing the prior distribution of probability over the possible positions of that ball can be seen in a scenario where a person, who can be selected by any process whatsoever, is asked to place the first ball on the table. The second ball is now thrown n times and comes to rest to the right of the first ball on m occasions. Assuming that the table is level *etc.* the position of the first ball can be estimated with progressively improving accuracy, from the ratio m/n, provided that the resting points of the second ball are randomly and uniformly distributed across the table. If we bias the result by any undeclared, non-

uniform distribution of the second ball, this must, in general, degrade the accuracy of the result. Similarly, the result will be degraded if we assume any non-uniform prior probability over the possible values of P_m unless we can be assured that the peak of the prior will always coincide with the true value of P_m. While one can be tempted into believing that such a coincidence will generally have a vanishingly small probability of occurrence, we shall see below that there are indeed cases where this coincidence is the rule, rather than the exception.

We can, however, make Bayes' experiment even more general if we consider a case where the first ball is thrown onto a large dance floor. We then place, wherever we wish, but enclosing the first ball, a rectangular frame, into which the second ball is thrown n times. With the frame in position, we count m and we estimate the position of the first ball relative to a given corner of the frame. We can then choose any other position for the frame and repeat the experiment. Not only shall we get a different answer for each different position of the frame, there actually will be a different, and true, answer for each of the infinitely many possible positions of the frame. Clearly therefore, the prior distribution over the positions on the dance floor has no relevance within the metric process, but there can equally be no doubt that in every case the first ball has a true position relative to the frame. That position can be given the logical force of Bayes' 'first event', to the probability of which there is one prior which is compatible with, and invariant for, any position of the frame, and will lead to an unbiased rate of convergence in every case: that is, the uniform prior. It seems therefore that a uniform prior distribution has special and powerful properties within the metric process which is invoked by Bayes' experiment and by logically equivalent cases, and is a matter quite distinct from the process which positions the first ball. Therefore, although several points mentioned in earlier chapters[1] concerning the uniform distribution of P_m stand as stated, they are a separate matter, and do not undermine the validity of the rôle of the uniform distribution within the metric process on Bayes' table. Furthermore, although, in the above illustration, the order in which the balls are thrown seems fundamental to the metric process, a little reflection shows that the essential features are present simply in the positions in which the balls come to rest. For, after all the balls have been thrown, their positions can be marked and we can arbitrarily designate one such position as being that of

[1] See esp Ch 5 and the references there to Murray, Molina, Stigler, Edwards *et al*.

the 'first' ball. We can then place the frame, at will, and proceed as described above [1].

We therefore decided to put these views to the test by conducting a set of experimental trials [2], similar to Bayes' experiment, but where the true value of P_m is determined arbitrarily and there is no knowable prior probability distribution over the possible magnitudes. Following each trial, we were told that the event *M* happened *m* times in *n* throws and we were required to decide whether the value of P_m probably was, or probably was not within limits $x_1 \rightarrow x_2$ and to invest accordingly an amount determined by our estimate of that probability. Modelling such a set of games and allowing the band defined by x_1 and x_2 to be chosen at random, we found a remarkably high number of games in which Bayes' method correctly determined whether P_m was within the limits, and therefore achieved a remarkably high average gain *e.g.*:-

True value of	$P_m = 0.5$	*RESULTS*
Throws in a game	$nt = 5$	*Won 96, Lost 4 ;*
Games in a set	$Ss = 100$	*Mean gain per game = 0.8407*
Lower bound	$x_1 = 0.04704$	
Upper bound	$x_2 = 0.219$	

Examination of such results shows however that, if we allow x_1 and x_2 to be chosen at random with uniform probability, there is a relatively high probability that they will bracket a section of the posterior distribution where the probability of its containing the true value of P_m is so very low as to make the probability of Bayes' method giving the right answer correspondingly high. A more stringent test was to force x_1 and x_2 to bracket the true value of P_m quite closely *e.g.*:-

True value of	$P_m = 0.5$	*RESULTS*
Throws in a game	$nt = 5$	*Won 0, Lost 100 ;*
Games in a set	$Ss = 100$	*Mean loss per game = 0.7192*
Lower bound	$x1 = 0.4$	
Upper bound	$x2 = 0.6$	

and we then found that we had to increase *nt*, the number of throws in each game, to *21*, before we reached a state of generally winning:-

---

[1] For comments on surface effects *etc.*, see Chapter 5.

[2] The trials were conducted using a *MATLAB* simulation.

True value of	P_m *= 0.5*	*RESULTS*
Throws in a game	*nt = 21*	*Won 66, Lost 34 ;*
Games in a set	*Ss = 100*	*Mean gain per game = 0.1718*
Lower bound	*x1 = 0.4*	
Upper bound	*x2 = 0.6*	

However, when we increased P_m to *0.9*, we found:-

True value of	P_m *= 0.9*	*RESULTS*
Throws in a game	*nt = 2*	*Won 0, Lost 100 ;*
Games in a set	*Ss = 100*	*Mean loss per game = 0.5912*
Lower bound	*x1 = 0.8*	
Upper bound	*x2 = 0.999*	

and an abrupt transition from losing to winning as *nt* was increased from *2* to 3:-

True value of	P_m *= 0.9*	*RESULTS*
Throws in a game	*nt = 3*	*Won 74, Lost 26 ;*
Games in a set	*Ss = 100*	*Mean gain per game = 0.2179*
Lower bound	*x1 = 0.8*	
Upper bound	*x2 = 0.999*	

This led us to consider values of P_m spaced evenly across the unit interval, for each of which we iteratively computed the number of throws per game which were necessary, at each value of P_m, in order to make a net gain from following Bayes' rule. The results are illustrated in Fig 11.1:-

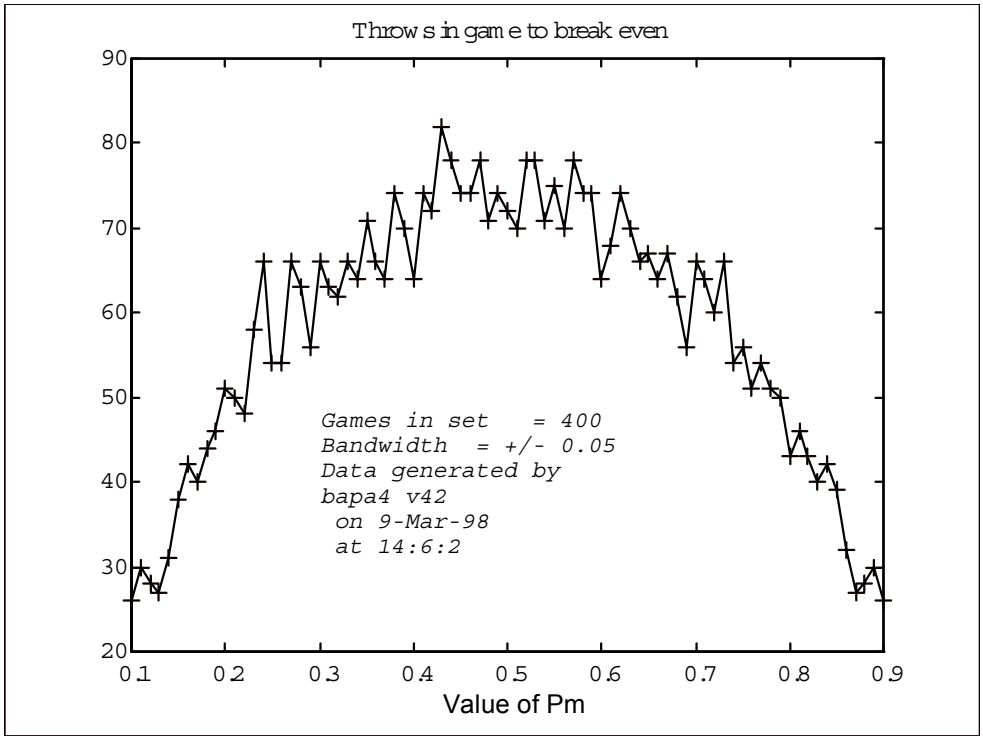


Figure 11.1

Now, although these results are few, experimental and crude, almost to the point of being merely anecdotal, they are sufficient to cause suspicion that Keynes may have been not entirely right when he wrote:- *' ..... as the number of instances is increased the probability that [the true value]*[1] *is in the neighbourhood of [m/n] ..... tends towards certainty ..... But we are left with vagueness ..... respecting the number of instances we require. We know that we can get as near certainty as we choose by a finite number of instances, but what this number is we do not know.* For, what Keynes failed to point out is that, once a value for P_m is determined, the ratio *m/n* will tend to that value at a rate and with a probability which are determined entirely by the binomial distribution. Indeed, a conventional prior probability may be thoroughly misleading as to the true value of P_m whereas even a very short trial may provide a highly informative result. The situation therefore seriously demands an explanation in that our experiments seem to show that Bayes' rule is highly effective, despite the massive criticisms levelled against the postulate of the uniform prior by so many authorities.

[1] Keynes denotes the true value by the letter *q*. The ratio *m/n* is denoted by *q'*.

We therefore need to explore more deeply the reasons why Bayes' rule works so well in the cases illustrated. This takes us back to Bayes' fundamental question, posed in the opening words of his essay, before the introduction of the experiment with the table, concerning the case where we simply count *n* throws and *m* events and, as ever, wish to know the probability that the value of P_m lies between any two defined values. To answer this question we have to consider fully the nature of this type of metric situation, which comprises three elements:- *(i)* the identification of events, *(ii)* the counting of *n* throws and *m* events, and *(iii)* the random variations in *m* induced by the binomial process. Assuming there are no errors in *(i)* and *(ii)*, it remains to look quite closely at *(iii)*, where the process which predetermines the value of P_m, is irrelevant to the subsequent problem: that is, we are now concerned only with the numerical value of P_m and with the variations in *(m,n)* which are induced by the binomial process.

The core of the problem is the fact that, for any value of $n > 0$, and for any value of P_m which is greater than zero and is less than unity, any value of *m* which is not greater than *n* can occur. Conversely, therefore any given result *(m,n)* can be produced[1] by any value of P_m which is greater than zero and is less than unity. Further, the probability of any such result in this type of case, unlike in cases which are analogous to that of the calibrated ruler[2], can be quite strongly dependent upon the true value of P_m. At this point, we may feel that we have turned yet another full circle and are back to where we began. It is, however, more appropriate to think of descending a spiral towards the roots of the problem, one such root being the fact that Bayes, followed by countless other authors, fell into the trap of using language which implies that we are dealing in 'absolute probabilities' and that they are, in principle, knowable. Indeed, in the analysis by Keynes quoted above[3], and despite his use of the symbol *'a/h'*, Keynes fell into this same trap. This is sadly ironic, for as Betrand Russell pointed out in his review of Keynes' *Treatise:- ' ... a proposition does not have a probability in itself, but only in relation to certain data. It may have different probability relations to different sets of data ..... The probability of a proposition 'a' relative to data 'h' is represented* .. [in Keynes' Treatise] .. *by 'a/h'. ..... It is remarkable how often the logic and philosophy of mathematical concepts has gone astray through the employment of symbols which did not contain explicitly all the variables upon which their value depended. If one were to employ such a symbol as (say) 'P(a)' for "the probability of 'a' ", adding a*

[1] There are of course certain restrictions at the extremes where $P_m = 1$ *or* $P_m = 0$.

[2] See Ch 9 above.

[3] Keynes (1921) p388

Version zh25a.doc at 18:36 on 13 July, 2001

proviso that this should be taken relatively to certain data 'h', it would be impossible .... (always) to remember ..... the relevance of 'h', and one's analysis would be certain to suffer ..... The introduction of the symbol 'a/h' is therefore of great importance. This symbol means the degree to which 'h' makes 'a' probable[1]*'.*

Unfortunately, Keynes did not consistently bear in mind, that although probabilities may be 'objective', they are defined only in relation to certain data and assumptions: in any process of inference from observation to an underlying truth, which we attempt to express by images, be they verbal, mathematical or graphic, the mapping from the underlying realities onto the image-field requires us to make assumptions and adopt conventions. All assertions of probability concerning empirical matters are therefore derived by operations which combine observational data with assumptions and conventions, all of which are needed to formulate the inference. Every assertion of a probability concerning an empirical matter ought therefore to state or designate the data and the assumptions, relative to which it is made. As we saw earlier[2], the values we compute for empirical probabilities can themselves be only probable to some degree, and again relative to data, albeit this fact is not explicit in Bayes' essay and is overlooked by many others. Thus, when Bayes defines his problem as being to compute *the* probability that P_m lies between any two given values, the computation can only be relative to the data and the assumptions. Unfortunately, Bayes' language and that of many others, frequently gives the impression that they are dealing in absolute values[3]. Yet, if we refer back to Chapter 3, we find that, in Bayes' *Corollary to Proposition 4* the value of the expectation (and therefore of the corresponding probability) depends, not upon the full truth about the situation, but rather upon what we believe, assume, or think that we know.

In the metric situation represented by Bayes' experiment, and by logically equivalent situations, a declaration of the prior distribution which is assumed is, therefore, a vital factor. Yet we are not free to assume at whim any arbitrary distribution. For, we know that the ratio m/n tends to P_m as n increases and we know[4] that only two priors yield m/n as being, relative to the data, the most probable value and most accurate estimate of P_m: one such prior is the uniform, the other is that which has, in each case, its peak at P_m. Hence, if we lack a prior with a peak at P_m, wherever that may be, there is

[1] Russell (1922)
[2] Chapter 8 above
[3] See also Edwards (1992) p3 on the rejection of 'absolute belief'.
[4] See Ch 5 above.

no alternative to accepting the uniform prior as the basis for estimating the probability that P_m lies in any given interval: for we cannot rationally assume a different prior distribution in estimating the probability that P_m lies within defined limits, *i.e.* $\mathcal{P}_H\{x_1 < P_m < x_2 \mid m, n, \ldots\ldots\}$, from the prior which we assume in estimating P_m itself. We therefore have to assert that, in the absence of any other valid prior distribution over the possible values of P_m, Bayes' rule that we should assume a uniform prior in computing the probability, relative to the experimental *m,n* data, that P_m lies in a defined interval is valid. There is, however, no knowable measure of the probability that the result is, in any absolute sense, true. The result is valid and correct in relation to the information and assumptions. As is true of every valid and correct assessment of probability.

Nothing we have said, however, implies that there can never be a valid non-uniform prior in a Bayes trial. An illustration of such a prior can be provided by a stick which has an unknown length ℓ, which we wish to determine. To measure ℓ by means of a common ruler, we first align one end of the stick with the zero mark on the ruler and we then examine the other end of the stick in relation to the gradations in that area of the ruler, following which we report, for example that the length of the stick is *'About 90.5 millimetres'*. Answers of that type will be about as good as we can get by this approach: no matter how many times we measure the stick, we shall, within the limits of patience, give the same answer. If, however, we have calibration data, we can use that data to give a distribution of probability over the possible values of ℓ. If, however, we need to go further and arrive at a more refined answer than is possible simply by means of the ruler, we can devise a procedure in which we create a direct connection between Bayes' experiment and the problem of measuring a stick. That is, in Bayes' experiment:- *the event that the red ball comes to rest in the rectangle ADso, between the line os and side AD in a single trial is called the happening of the event M ..... (and) ..... the white ball having been thrown, the probability that the event M will occur in a single trial is equal to the ratio of the line segment Ao to the length of the whole side AB*[1]. Conversely therefore, one may envisage fairly simple conditions under which the length *Ao* is equal to the unknown probability P_m multiplied by the length of the side *AB*. Hence, as Bayes' experiment provides us with a means of estimating P_m to whatever accuracy we require[2], simply by making the number of trials, *n*, sufficiently

[1] Lemma 2 to Postulate 2.

[2] This accuracy can be measured as the probability of getting the observed results if the true value were outside given limits. See also Ch. 9 above.

large, it likewise provides us with a means of estimating the length *Ao*, in relation to the length *AB*, to whatever precision we require.

Thus, to refine our estimate of the length of a stick, we may construct a table, level and uniformly flat, in which the side *AB* is formed by the ruler. We then lay the stick alongside the edge *AB*, with one end of the stick aligned to the point *A*, and the other end of the stick we equate to the point *o* in Bayes' experiment. To refine the distribution of probabilities over the length of the stick ℓ, we throw a suitable ball *n* times and we count the number of occasions on which it comes to rest to the right of the line *os*[1]. Given then that the result is *m* successes out of *n* trials, we can apply Bayes' theorem in its full form[2]:-

$$\mathcal{P}_H\{(x_1 < \ell < x_2) \mid m,n,P_0(x)\} = \frac{\int_{x_1}^{x_2} x^m (1-x)^{(n-m)} P_0(x)\, dx}{\int_0^1 x^m (1-x)^{(n-m)} P_0(x)\, dx} \qquad (11\text{-}1)$$

where $P_0(x)$ denotes the prior distribution derived by combining a conventional measurement and the calibration data for the ruler. Such a prior will generally peak at a probability of zero error and thus lead to a distribution of $P_0(x)$ which correspondingly has its peak at the true value ℓ.

Hence, although the distribution of an attribute within a population cannot be used as a prior probability distribution over the possible values of that attribute in an individual, perfectly valid prior information can be derived from previous[3] observations which have been made upon that same individual. The essence of the matter is that, if the object to which we wish to apply Bayes' theorem is an individual, then both event E_1 and event E_2 in equations *(3-26)*, *(3-51) etc.* must relate specifically to that same individual. That is not to assert any blanket logical prohibition against the use of population statistics in applications of Bayes' theorem. We do, however, assert that the result of doing so is a probability in the sense of a proportional frequency among the population and does not provide evidence about an

[1] Being severely practical, we accept that individual trials will occur when we cannot decide on which side of the line the ball has come to rest. Various techniques can however be envisaged for dealing with this problem.

[2] See also equation *(5-11)*

[3] The temporal order is not necessarily significant; it is the logical order which determines the prior.

individual drawn from that population. Thus, where we know that a probability P_m is selected randomly by a known statistical mechanism, we can make valid statements of the form $\mathcal{P}_R(\mathcal{P}_m \approx x|\ k) = y$ but, once P_m is selected, it becomes a fixed governing parameter and, on the basis of experimental results $\{m,n\}$ derived under $\mathcal{P}_m$ we can only make statements of the form $\mathcal{P}_H(\mathcal{P}_m \approx x|\ k,m,n) = y$

We return now to the fundamental question of this chapter which concerns the validity of assuming uniform prior distributions in cases which are logically equivalent to Bayes' 'unknown event' and against which so many authors [1] have strongly argued, essentially because the assumption is, or seems to them, so major and so lacking in warrant. To some of those authors, a further and conclusive objection is that, in a dimensional case, a prior distribution which is uniform in relation to a given form of measurement, may be seriously non-uniform if we choose instead to measure some other function of the parameter. As we mentioned in Chapter 9, if a uniform prior probability has been assumed when measuring, say, a resistance in an electrical circuit, the distribution of the reciprocal if we chose instead to measure the potential drop under a constant current, would be decidedly non-uniform[2]. We have however shown in Chapter 9, that when dealing with dimensional quantities, the logic of the calibrated scale removes the need to provide an initial prior for the object being measured, for we may indeed not know what it is that we are measuring and we may, even less, understand what is meant by a 'prior probability' in such cases. Instead, the calibration of the measuring device provides the prior distribution of probabilities and uncertainties for the first observation, and that result becomes, in turn, the prior distribution over the possible values of E_1 as further observations are made.

We are left, therefore, with the central and fundamental objection that, in cases which are logically equivalent to Bayes' experiment, there is no warrant for the assumption that all possible values of P_m have a uniform prior probability. Yet, in view of our own previous discussion, we are forced to wonder whether the problem may not stem more from the name 'uniform' than from the substance which it represents[3]. For the essence of the matter is not that we have to assume all the probabilities to be equal, it is that we have no information to rate any probability more highly, or less

---

[1] *e.g.* Boole, Keynes, Fisher, Hacking, Fine

[2] Jeffreys, H. (1983), Ch 3, discusses in great detail various techniques for overcoming problems of this type in specific, practical cases but seems to have no general solution.

[3] *cf. 'much confusion may lie in the use of symbols and the notion of variables in probability'*; Keynes (1921) p58

highly, than any other. 'No information' is the crucial point and we therefore require a means of representing this fact in Bayes' equation. Tentatively, therefore, we abandon the name 'uniform prior' and call it instead the 'Information zero' distribution, symbolised $I_Z(\ .\ )$.

It is now worth noting that, if we were to assume that zero information about the prior probability of an event required us to set that probability itself to zero, then Bayes' equation would collapse to *0/0* and become imponderable: which might seem, to some people, a fair representation of the situation. On the other hand, it is also arguable that setting the prior probability to zero would be not only equivalent to claiming a great deal of information where we have specified there to be none, it would also deem the occurrence of the event M to be impossible, which would be immediately contradicted by any (m,n) trial in which m has any non-zero value. It therefore seems much better to maintain that probabilities are always relative to the information available; and that, for the reasons just given, we cannot, in the case of an unknown event, set the prior probabilities to zero; nor can we make the prior probability of any possible value of P_m greater than that of any other possible value; and therefore, considering algebraically a set of possible, but exhaustive and mutually exclusive values $x_1, x_2, \ldots. x_n$ for P_m, we have the following very simple equations. First, as we can, in no case set $\mathcal{P}_H(P_m = x_i \mid I_Z(x)) = 0$, all must be greater than zero:-

$$\mathcal{P}_H(P_m = x_1 \mid I_Z(x)) > 0;\quad \mathcal{P}_H(P_m = x_2 \mid I_Z(x)) > 0; \ldots\ldots$$

$$\mathcal{P}_H(P_m = x_{n-1} \mid I_Z(x)) > 0;\quad \mathcal{P}_H(P_m = x_n \mid I_Z(x)) > 0; \qquad (11\text{-}2)$$

Then, as we can set no value greater than, nor less than any other value:-

$$\mathcal{P}_H(P_m = x_i \mid I_Z(x)) \not< \mathcal{P}_H(P_m = x_j \mid I_Z(x));\ \forall(i,j) \qquad (11\text{-}3)$$

$$\mathcal{P}_H(P_m = x_i \mid I_Z(x)) \not> \mathcal{P}_H(P_m = x_j \mid I_Z(x));\ \forall(i,j) \qquad (11\text{-}4)$$

from which it follows that:-

$$\mathcal{P}_H(P_m = x_1 \mid I_Z(x)) = \mathcal{P}_H(P_m = x_i \mid I_Z(x)) \ldots\ldots$$

$$= \mathcal{P}_H(P_m = x_n \mid I_Z(x)); \qquad (11\text{-}5)$$

That is, if we define the probability of a proposition, relative to the known data *etc.*, as the degree of reason to believe the proposition to be true, then, clearly the *a priori* degrees of reason that we have to believe each such proposition $P_m = (x_1 \mid I_Z(x))$, and $P_m = (x_i \mid I_Z(x))$, and $P_m = (x_n \mid I_Z(x))$, are the same in every case. Hence, it seems clearly to follow that the probabilities, ***relative to the data***, must be equal.

This inference, was, however, fiercely attacked by Boole[1], who castigated it as an unwarranted postulate of *"the equal distribution of our ignorance."* Boole was, however, as was Karl Pearson many years later[2], under the disadvantage that, at the time of their writing, the concept of 'information' as a measurable quantity had not been formalised[3]. They could not, therefore, see clearly that a situation of zero information can be defined as a situation in which all probabilities are equal - relative to the available data. More fundamental, therefore, was their failure to remember that probabilities in real life can only be assessed in relation to the available evidence.

Keynes[4] also devotes considerable attention to a closely related topic under the heading of the Principle of Indifference, and points out a number of difficulties. For example, nothing in equations *(11-2)* thru *(11-5)* asserts or implies any uniformity in the distribution of the possible values $x_1, x_2, \ldots x_{n-1}, x_n$, each of which can represent a point, or a range of values. This may seem, at best paradoxical or self-contradictory, and at the worst, ludicrous. For, taken in isolation, it implies that, into whatever arbitrary set we may partition the totality of possibilities, each will have an equal probability. We may look at a group of tourists and ask, *'Are they Australians, Muscovites or New Englanders ?'* and it is clearly not acceptable to apply a uniform prior across such a partition. Hence, we would suggest that whenever we are dealing with non-numeric partitions, an acceptable procedure may be to form binary pairs from each attribute and its opposite and, where feasible, to evaluate each pair separately, *e.g. 'Australian' or 'not-Australian' etc.*.

In the case of a Bayes' experiment, however, we are not free to use any arbitrary partition, for we require that, whatever the true value of P_m may be, the peak of the posterior distribution shall be able to converge on that value 'absolutely', at least in the sense that we can make the value of any rational fraction *m/n* converge as closely as we wish on the value of any given real number in the *0-1* interval. Hence, the need for a uniform partition of the unit interval, within which we assign the priors equally, is forced upon us by the requirement that the posterior shall be allowed to converge without hindrance or bias on the true value of P_m.

Thus, when we then apply *(11-5)* to a uniform partition of alternatives among the real numbers in the $0 \rightarrow 1$ interval, and apply the result in Bayes' equation, we find that the prior, being a constant in both numerator and

---

[1] Boole (1854) p.370
[2] Pearson (1920)
[3] See Hartley (1928) and Shannon (1948)
[4] Keynes, (1921), Ch 4.

denominator, graciously disappears of its own accord [1]. This saves us from the embarassment which would be caused by its presence in the expression which is actually evaluated. It can therefore be argued, and not unreasonably, that this disappearance is not a meaningless algebraic fluke, but actually provides welcome support for the belief that the algebra is a valid image of the information present in the physical reality [2]. It also follows that, if we are using a Bayes table to refine the measured position of a pointer, then it does not matter whether the pointer designates volts or ohms, for, in either case, the uniform partition and the assignment of equal probabilities to the segments of that partition, simply denote zero prior information and permit unbiased convergence on the true value, regardless of the units in which that value is measured.

As an aside, it is also worth noting that the value of specific evidence, in relation to a set of competing hypotheses, can be measured by the difference which that evidence makes to the relative probabilities of the hypotheses. If the evidence makes no difference to those probabilities, its information value is zero in relation to that set of hypotheses. Thus, given evidence k and a set of competing hypotheses, h_1,..... h_n, we can take the hypotheses in pairs and use the sum of the logarithms [3] of their probability ratios, relative to the evidence, as a measure of the information-value relative to those hypotheses [4]. On this basis, we can define an Information-value function $I(h_1,.... h_n \mid k)$ for evidence k in relation to hypotheses $h_1 .... h_n$ such that:-

$$I(h_1, ... h_n \mid k) \quad = \quad \frac{1}{2} \sum_{i,j} \left| \ln \frac{\mathcal{P}_H(h_i \mid k)}{\mathcal{P}_H(h_j \mid k)} \right| \qquad (11\text{-}6)$$

and therefore, if k conveys no information [5] regarding the given set of hypotheses, the magnitude of the index in relation to that set will be zero. Finally, in this connection, it is worth noting that Jaynes[6] and many others have shown that the density function which maximises entropy is uniform if the density is constrained within a known interval. Such works claim that

---

[1] *i.e.* it multiplies both the numerator and denominator in *eqn (5-18)* and therefore disappears.

[2] The precise nature of the 'physical reality' which we call 'probability' is a nice question.

[3] The arguments for the use of logarithms are given succinctly in the Introduction to Shannon (1948). See also Hartley, R.V.L. (1928).

[4] Summing over all pairs, including $i = j$, we take the modulus of the logarithm of each probability ratio and halve the result to allow for the duplication.

[5] See also Medawar (1963). His 'law' of the conservation of information is a useful adjunct to this argument.

[6] Jaynes (1982).

entropy, defined as $\int P(x)\log P(x)dx$ is a good measure of 'freedom', and therefore that maximum entropy corresponds minimising assumptions. Hence, the uniform prior can be claimed to minimise the assumptions when we only know that the probability is constrained to the unit interval.

Resuming the main theme, if we now look back to equation *(5-18)* and re-write the left hand term as $\mathcal{P}_H\{ x_1 < P_m < x_2 \mid m,n,I_Z(x) \}$ it clearly follows, because the items denoted $m,n,I_Z(x)$ constitute the totality of the evidence, that the most probable value of P_m in relation to that evidence can only be the ratio *m/n*. This, in the case of an 'unknown event', where a prior observation is excluded by definition, is only the case when we assign the same value to the prior probability of each possible value of P_m.

We conclude therefore that, far from our having no reason to assume *a priori* equal probabilities, the reasons for doing so are, in fact, compelling. To the objections of Boole, Keynes, Fisher and Fine that, by assuming these probabilities to be equal, we are implicitly claiming that empirical knowledge can be known *a priori*, we would answer that, precisely as we have demonstrated, it is by making these probabilities equal that we actually express the absence of prior information. In sum, where we are trying to measure the underlying probability of an event, about which we have no previous knowledge, Bayes' rule, with the explicit assertion of the 'information zero' prior, leads to a result which is valid, objective, and corresponds with a great deal of everyday practice. We infer therefore, that, in constructing a calibration table, and excepting valid reasons to do otherwise, we can, and we must, assume a uniform, 'Information-zero' prior.

It is therefore important to note that calibration histograms which are constructed in this way, do not cause a non-informative prior to become informative if we change the units of measurement. This is because the error-rate in a band is simply the ratio of two positive, finite, non-dimensional integers. The denominator is the number of trials and the numerator is the number of occasions on which the error falls into the given band. The uniform prior is therefore also non-dimensional. Thus, if the units of measurement are changed, say from ampères to watts, the widths of the error-bands are automatically scaled. There is no contradiction of the assumed ignorance and there is no assumption of a privileged position.

In academic circles, there has been an awareness of, and embarassment by the question mark which has hung over Bayes' rule for the past *150* years and, in consequence, there has been a marked reluctance to speak clearly and openly of probabilities derived by inference from observation. This embarassment has helped no one: it has led to confusion, to bitter inter-

personal conflicts [1] and a failure to deal with questions which have seemed clear and simple enough in themselves and to warrant clear answers. Yet, time and again, where we seek clarity, we find ourselves thrown into confusion by problems over the nature of the prior probability. Although we would suggest that this matter can be resolved, as shown above, in the case of a Bayes' experiment, there remain, in other scenarios, difficulties which we have not yet resolved .

These difficulties are particularly acute when we wish to establish probable explanations, probable causes, and probabilities of hypotheses. Putting aside the philosophical difficulties of defining precisely what we mean by 'cause' *etc.*[2], the essence of our difficulty is nicely summarised by Keynes, who first quotes De Morgan:- *Causes are likely or unlikely, just in the same proportion that it is likely or unlikely that that observed events should follow from them. The most probable cause is that from which the observed event could most easily have arisen*[3]. Keynes however then argues against De Morgan:- *If this were true the principle of Inverse Probability would certainly be a most powerful weapon of proof, even equal, perhaps to the heavy burdens which have been laid on it. But the proof given in Chapter XIV* [4]*, makes plain the necessity in general of taking into account the 'a priori' probabilities of the possible causes. Apart from formal proof this necessity commends itself to careful reflection. If a cause is very improbable in itself, the occurrence of an event, which might very easily follow from it, is not necessarily, so long as there are other possible causes, strong evidence in its favour.*

However, although we have shown one case of a valid non-uniform prior, *i.e.* where we have a previous measurement of the attribute in question, we have not, so far, shown any general basis on which we can distinguish a valid prior from one which is not valid. We can, however, illustrate the problem, and propose a solution by considering an air traffic control highway, an 'air lane', for which we have available full statistics concerning types of aircraft, departure times, heights, speeds *etc.* going back many years. Suppose, then, that we have a radar which is surveying this air lane and, at a certain time we get an initial detection of a new aircraft. This detection gives us the geographic position of the aircraft, but no further

---

[1] See e.g. Fisher (1937), Neyman (1961).

[2] See e.g. the article on *Causation* in the Cambridge Dictionary of Philosophy. The author is not listed.

[3] Keynes (1921) p 178; quoted from De Morgan's *Essay On Probabilities* in *The Cabinet Encyclopaedia,* (no date).

[4] *i.e.* in Keynes (1921)

information, and we would like to know, without waiting for a further detection, the probabilities attaching to various hypotheses as to the heading, height and speed of the new aircraft. Given, therefore, all the statistics which are available to us, it is easy to assert, that, for example, *'90% of all aircraft which use this air lane and are initially detected in this position, have a heading of* $090^o \pm 2^o$*, a height of 20,000ft± 1000ft, and a speed of 420 kts± 50kts'.* Further, it is easy and entirely valid to regard these data as providing 'probabilities' in the sense of 'rational expectations of proportions, based on previous history, within the population of aircraft using this air lane'.

It is then tempting, but quite invalid, to use the statistics to provide a distribution of probabilities over the initial parameters for a particular aircraft. The depth of this invalidity is easily shown by considering the implications if the pilots of aircraft were to follow similar reasoning in the exercise of their own duties and were to assume, for example, that the speed of the aircraft could be gauged to some degree by considering the speeds of other aircraft on other occasions. The results would quickly be disastrous because such reasoning is totally fallacious, notably in that information about a population of previous flights conveys no information whatsoever about the actual state of any specific aircraft [1]. If however we consider a flight plan for a specific aircraft and we know, for instance, that the plan is used by an on-board computer to control engine settings *etc.*, then we may be in a totally different kind of situation, where the provision of the flight plan may be logically equivalent to separate evidence directly relevant to the specific case. Yet, as with the ruler, we must allow some probability that the system is not working as intended.

The essence of the matter thus appears to be that, for data to provide a valid prior probability about a specific situation, the data has to be informative about the physics of that specific situation. The fact that a given flight [2] has always, on previous days, been above a certain height at a certain point in the air lane, tells us absolutely nothing about its height at that point on a new occasion. It may tell us what it is reasonable to expect on a new occasion, especially if we know no reason to expect anything unusual, but it tells us nothing about what we should believe to be the case.

---

[1] In cases where the population statistics are complete and include the specific aircraft, they may however provide valid information about the maximum and minimum possible values of parameters.

[2] We are here using the term 'flight' in the special sense used for airline timetables.

The distinction between statistical expectation and probablistic belief can also be important in deciding the true performance of, say, a radar, when flight-trial results appear to be in conflict with the design intent [1]. For, in the case of a radar, the detection performance is generally determined by means of a Bayes trial against a known aircraft and it is arguable that the design intent constitutes a valid prior, with a logical status akin to that of a previous measurement. The logic of this situation is again closely related to that of the calibrated ruler [2], and it is therefore arguable that, if the design intent is to be introduced as a prior, then a term denoting the probability that the radar is actually operating according to the design intent, should also be included.

Different, but related examples are provided by Fisher's mice, which are discussed in the appendix, and which actually create several logically distinct types of situation. For example, if the parents of the mother mouse were known to be genetically $RR \wedge RR$, then the physics relevant to the mother are absolutely specific and we know with certainty, without reference to her offspring, that the mother must also be of genetic type *RR*. Conversely, if the mother's parents were $RR \wedge gg$, we know without reference to her offspring, that the mother must also be of genetic type *Rg*. However, a litter of seven red offspring and an assumption of the 'information zero' prior would have given the *Rg* hypothesis a probability rating of only *0.0078*; a fact which serves to re-emphasise just how important it is to declare one's assumptions and to beware of inferences which may be dramatically sensitive to errors in those assumptions. Yet, as shown in Chapter 8, qualitative errors in our assumptions may have quantitatively trivial implications and it is therefore important to consider each case on its merits. In contrast, however, to the above cases where the genetic facts of the mother's parents are such that they specifically determine the genetic constitution of the mother, there are other cases where the mother's parents determine only the probable expectation of genes and provide no specific information to support any degree of reasonable belief about the mother's actual genetic constitution.

It therefore appears that we may here have some further problems of notation and of semantics in relation to the differences between 'probability' meaning 'a reasonable expectation concerning the frequency of an event' and

---

[1] The difference does not always reflect unfavourably on the true performance. The authors know of one radar where the true performance greatly exceeded the design intent. This was eventually traced to the omission of a scaling factor in the specification of the antenna.

[2] See equation *(9-4)*.

'probability' in the sense of 'the degree of reasonable belief about a specific situation'. However, the notational problem is not severe, if we always remember that the symbol $\mathcal{P}_R$ denotes an expectation concerning the frequency of an event within a population, and the symbol $\mathcal{P}_H$ denotes an hypothesis or belief concerning a fixed value in a specific case. In operational systems, however, it is probably desirable to further emphasise cases where a displayed value is merely an expectation derived from statistics in contrast with those cases where the displayed value is derived from evidence concerning the specific object. For example, in radar surveillance systems, there are many statistical correlations between parameters such as the positions, heights, speeds and identities of aircraft. Quite commonly, the geometry of the earth's curvature is used to give an estimate of the height of an object at the first detection. This 'guesstimate' of height can then be used to produce a further 'guesstimate' of speed on the basis that, especially in airways, the heights and speeds of aircraft are strongly correlated. Hence, quite often the figures thus produced will be fairly accurate. Yet, occasionally, the object detected will turn out to be a stray meteorological balloon and the 'guesstimated' speed will be badly in error. Experienced operators, however, often draw upon regular correlations in the interpretation of radar returns and it is not unusual to find that such correlations are assumed in the formulation of operational procedures, *e.g.* there may be an axiomatic assumption that all aircraft within the coverage of the system are flying according to known rules. Equally, however, many systems provide for the measurement of specific parameters and it will often be the case that a value which is measured specifically will be displayed to a controller in preference to a value which is inferred from procedural rules or statistics. It therefore seems clear that where the values displayed to the users of a system can be of these different kinds, the display should make clear what type of value is being shown and that the users should be given an adequate understanding of the differences in origins and dependability.

These issues raise a further question when a probability based on observation of an individual is combined with a population-based probability. For example, radar data, R may show that a specific aircraft, observed in a known airlane, is one of two types T_1 and T_2 such that $\mathcal{P}_H(T_1 \mid R) = p_1$ and $\mathcal{P}_H(T_2 \mid R) = p_2$. Statistical data S may also give the relative frequencies with which aircraft of those types use the airlane such that $\mathcal{P}_R(T_1 \mid S) = f_1$ and $\mathcal{P}_R(T_2 \mid S) = f_2$. Hence there may easily arise a strong desire to combine these two sources of data. However, while we may feel uneasy about such combinations, we can see no grounds upon which to deem them invalid - provided there is a clear understanding that the resulting

probability relates to a random event in the sampling of a population and not to an hypothesis about the specific aircraft. That is, the population comprises the aircraft on which the statistics are based and the combined probability $\mathcal{P}_R(T_1 \mid R,S)$ represents the relative frequency with which aircraft of type T_1 occur in that population and are expected to give radar returns R. Conversely for T_2. The addition of the statistical data tells us nothing beyond that which we know from the specific observation R about the individual case.

There is also a problem with the use of the term 'prior probability' in relation to belief about a situation, for this term may heavily pre-condition and distort our understanding of its meaning, especially if it is denoted symbolically by, for example, $\mathcal{P}_H(E_1)$ with the impression of an absolute probability being conveyed by the omission of the conditioning term. It is therefore important to understand that, in relation to belief about a situation, the term 'prior probability' denotes a probability based on prior evidence, (or assumptions), about that specific situation. Even so, the term 'prior evidence' is better understood as 'separate evidence'; for, as we have seen many times, the temporal order is not fundamental [1] and we can often exchange the values assigned to the terms E_1 and E_2 without changing the reasoning or the result.

We now return to the questions of probabilities of hypotheses and causes and to Keynes' assertion, against De Morgan, that there is:- *the necessity in general of taking into account the 'a priori' probabilities of the possible causes. If a cause is very improbable in itself, the occurrence of an event, which might very easily follow from it, is not necessarily, so long as there are other possible causes, strong evidence in its favour.* In this context, Keynes seems to be thinking not so much of the fairly simple quantitative *a priori* probabilities which we have considered above, but much more of the dogmatic priors which, in contemporary scientific culture provide massive defences against a plethora of individual and *ad hoc* explanations being deemed 'most probable' in relation to anything and everything which is observed. Here, the dogmatic priors channel our thinking in generalising, and often 'reductionist' directions. The dogmatic prior is however an instrument very different in nature both from a prior expectation over a population and from the prior belief stemming from direct observation of an individual. For it is *via* the dogmatic prior, even though often hidden, that we give effect to the principles of simplicity and generality, enjoined upon us by Empedocles [2] and Occam[1]: the focal axes in the 'modernist' scientific

[1] In the case of independent additive noise, as discussed in Ch 9, the events are effectively simultaneous.

[2] See Ross (1936) p487

view of the natural world. Dogmatic priors are therefore a means of creating high, often insuperable, barriers against explanations *etc.* which are viewed as being *a priori* unwanted, absurd, or perhaps impossible[2].

For example, one explanation of the apple which reputedly hit Newton's head, is that a demon pulled it from the tree and threw it at Newton. Indeed, everything that happens can be explained as the activity of demons, which is a massively reductionist form of explanation, but is dogmatically outside the contemporary scientific ethos and is arguably not so much 'wrong' as 'unwanted' by the scientific community. It is therefore excluded from our thinking when we consider the inference of probable causes and explanations. Dogmatic priors are however dangerous instruments; for the fantastic, surprising or outlandish explanation is not *ipso facto* improbable in a specific case: rather, the probability in the specific case has to be determined by reference to the specific evidence[3].

We therefore conclude that, when we are concerned to evaluate the distribution of probabilities over the possible magnitudes of an attribute of an individual, and the techniques applicable to the calibrated ruler are not available, the only valid prior distributions are, in one case, those derived from prior observation of the specific individual, or, in the other case, a distribution which represents zero information in the specific context of the attribute and the measure being employed. In the next Chapter we look at some practical applications for the procedures which follow from these principles.

---

[1] Although the famous 'razor' has never, we understand, been found in the known writings of William of Occam, the version to be found on pottery at the parish church of Ockham, in the county of Surrey in England, reads *'Frustra fit per plura quod potest fieri per pauciora' i.e. 'it is vain to do by more that which can be done by less'.*

[2] See also Richard Jeffrey's discussion of Dogmatism in Jeffrey, R. (1992), p.45.

[3] This can be illustrated by the case of a penny which, flipped 20 times, gives 'heads' on every occasion. If the penny came from the bank, we are surprised. If however the penny was purchased at a joke shop, there is very little surprise.

Version zh25a.doc at 18:36 on 13 July, 2001

Chapter 12

The Trajectory

From the seventeenth century onwards, increasing importance, commercial and military, and therefore also political, became attached to the accurate computation of the trajectories of stars and planets across the sky and of ships across the oceans[1]. The aim was to improve the safety and efficiency of maritime navigation. There were many aspects to this problem, involving navigators, surveyors, astronomers, together with many scientists and mathematicians who joined in the search for solutions. The search led to the realisation that, by combining a series of observations along a trajectory, we can often improve the quality of information over that provided by treating the observations separately.

Much later, the *20th* century saw a great widening of the field to which people, particularly communications and control systems engineers, found that they could apply the modern versions of the methods which, *300* years earlier, began their steady development to improve navigation and astronomy. The *20th* century expansion was directed at many different devices, systems and processes which change in a progressive manner but are subject to probablistic uncertainties, both in their observability and in their evolution. Navigation, communications and aerospace surveillance systems still provide prime examples of such phenomena, all over the world and in the inter-planetary space, where large numbers of micro-processors are, implicitly, using Bayes' theorem thousands of times every second.

These are serious matters, for such technologies not only provide vital support for the economic, cultural, sporting and military activities of wealthy nations, but also provide means which are vital to the effective execution of international missions of mercy and protection. Serious also are the evil uses to which such technologies can be applied. To debate that balance is not our purpose; rather it is, in this chapter, to explore the rea-

[1] See *e.g.* Adrewes (1996); Sobel (1996).

soning which supports the designs of these devices which have, as their essential purpose, the progressive integration and refinement of observational data obtained at successive points on the trajectory of an evolving process.

We shall therefore discuss means by which Bayes' reasoning, albeit often un-perceived and un-acknowledged, is applied to such situations. We must however point out that, in many of these situations, apprehensions about prior probabilities, particularly those affecting the first observation of a series, have caused many people to avoid the issue of probability altogether and to work rather in terms of likelihood. The consequence has been a widespread inability to pursue questions to their logical conclusions in terms of 'expectation', or 'probable cost'. This seems to be especially the case in aerospace systems and may well be also true in economics and social studies. However, as observations are often made by means of a calibrated ruler, or some logically equivalent device, the difficulties with priors, which have previously prevented the assignment of probabilities, can, as we saw in Chapter 9, be overcome. This should enable many deeper operational aims to be pursued through conceptual thickets previously thought to be impenetrable.

To illustrate a simple process of the relevant kind, we imagine a group of children who amuse themselves on the school bus by taking turns to guess the speed of the vehicle. We suppose that the guessing errors have a known distribution and we suppose that the speed of the vehicle is determined by numerous factors such as the speed limit, traffic density, road conditions *etc.*, plus other random elements which are individual to the vehicle and its driver. In this situation, if one child guesses the speed of the vehicle as, say, *30mph*, this guess conveys prior information as to the probability that the next guess will be, say, *35mph* and the true speed, say *36mph*. Another example is where we wish to model the acceleration of an aircraft. Here, the thrust of the engines will vary randomly by small amounts and our measurements of the speed will be subject to random errors.

In such situations, Bayes' reasoning, often combined with simplifying models of the physics, allows us to make a series of observations and to improve progressively the quality of the resultant information. While we must be careful not to over-generalise from specific examples, it is useful to regard the observations as being related so that if the attribute at the *n'th* observation has a true value X_n then the probability that the attribute at the next observation will have any 'guessed' value x_g, is given by a forecasting, or transitional probability function symbolised by the expression

$\mathcal{P}_F(X_{n+1} \approx x_g \mid X_n)$. By this means we can model a great variety of processes and linked activities where we know, or can reasonably assume, that the prior probabilities over the possible values of X_{n+1} are entirely determined when the probabilities over X_n are known, or are assumed to be known. This is known as the Markov[1] property and, coupled to Bayes' theorem, it allows us to construct computable models for a wide range of serially connected events.

As another simple example, we consider a series of measurements, taken with a calibrated ruler, where the object of interest is a pencil and the attribute of interest is its length. The first measurement gives, say *98.95mm*, and we wish to know the probability that the true length is, say, *100mm +/- 0.01mm*. In the case of additive errors which are independent of the length of the pencil[2], this corresponds with wanting to know the probability that the error was between *1.04mm* and *1.06mm*, which we can obtain from the calibration table. That is, we can, on the basis of such a calibrated measurement, make a plain statement of probability about the true length of the pencil without the inhibitions concerning priors which have for so long precluded such statements. Further, such probabilities can be multiplied by monetary values in order to achieve the magnitudes of probable costs.

If however we are to improve our knowledge by making further 'independent' measurements, we have to consider quite carefully the logical basis upon which this can be done. Conceptually, the task can be viewed as one of estimating and progressively refining the posterior distribution of probabilities over the possible values of a fixed attribute, which we can observe only via measurements, y_1 ..... y_n, each of which is independently subject to an additive error such that $y_i = X + \varepsilon_i$, where X is the true value and ε_t has a probability distribution, known by calibration. We include this fact in a term k_0 which denotes our relevant knowledge prior to the first observation. The first observation, y_1, then expands the information so that k_1 includes k_0 and y_1.

Thus, having observed a value y_1 we wish to know the probability distribution over the possible values of the attribute. We therefore apply the logic of a calibrated ruler, as described in Chapter 9, such that, in terms of Bayes' analysis[3], E_1 is the event that the error ε_1 has the value $y_1 - x_g$ and E_2

---

[1] Markov (1906).

[2] Provided that the calibration table is sufficiently detailed, it is necessary only that the errors shall be sufficiently small, in proportion to the true value, to support the assumption of independent additivity.

[3] See equations *(3-51) - (3-51f)*.

is the event that the measuring device is working according to the calibration. Hence, because the calibration, *C(.)*, tells us the probabilities appropriate to different errors, we know that the probability of the true value being equal to the guessed value x_g is :-

$$\mathcal{P}_H(X \approx x_g \mid k_1) = \mathcal{P}_R(\varepsilon_1 \approx (y_1 - x_g) \mid C(.)) \qquad (12\text{-}1)$$

The second observation further expands the information so that k_2 includes k_0 and y_1 and y_2. However, equation *(12-1)* now provides a prior probability, based on k_1, that X has the guessed value x_g. This allows the observed value y_2 to have the logical force of E_2, whence Bayes' theorem gives us:-

$$\begin{aligned} &\mathcal{P}_H(X \approx x_g \mid k_2) \\ &= \mathcal{P}_H(X \approx x_g \mid k_0, y_1, y_2) \\ &= \frac{\mathcal{P}_H(X \approx x_g \mid k_1) \cdot \mathcal{P}_R(y_2 \mid k_1, x_g)}{\mathcal{P}_R(y_2 \mid k_1)} \end{aligned} \qquad (12\text{-}2)$$

An important detail in *(12-2)* concerns the term $\mathcal{P}_R(y_2 \mid k_1, x_g)$ which expands, first, to $\mathcal{P}_R(y_2 \mid k_0, y_1, x_g)$ and further, with a small rearrangement to $\mathcal{P}_R(\varepsilon_2 = y_2 - x_g \mid k_0, X \approx x_g, \varepsilon_1 \approx y_1 - x_g)$. This denotes the probability that we will observe y_2, if:- *(i)* we are given k_0, and *(ii)* $X \approx x_g$ is true, and *(iii)* the error on the first observation was $\varepsilon_1 \approx y_1 - x_g$. One of our main assumptions is, however, that observational errors are independent of each other, whence it follows that the probability of an error $\varepsilon_2 \approx y_2 - x_g$ is independent of ε_1 and therefore the probability of observing y_2 at the second observation, if the true value of X is x_g, *i.e.* $\mathcal{P}_R(y_2 \mid k_0, x_g)$, is independent of the previously observed value y_1. At first sight, this may be very surprising and may seem to contradict relationships which are implicit and important in, for instance, the guesses of the children on the school bus. The essence of the matter is however that the *errors* on the guesses of the children are independent while all the guesses convey strongly correlated information about the true value. And it is precisely our purpose to extract that strongly correlated component. We therefore reduce the term $\mathcal{P}_R(y_2 \mid k_1, x_g)$ to:-

$$\mathcal{P}_R(y_2 \mid k_1, x_g) = \mathcal{P}_R(y_2 \mid k_0, x_g) \qquad (12\text{-}3)$$

whence:-

$$\mathcal{P}_H(X \approx x_g \mid k_2) = \frac{\mathcal{P}_H(X \approx x_g \mid k_1) \cdot \mathcal{P}_R(y_2 \mid k_0, x_g)}{\mathcal{P}_R(y_2 \mid k_1)} \qquad (12\text{-}4)$$

Hence, given further observations y_3 .... y_i .... y_n, we have, for similar reasons:-

$$\mathcal{P}_H(X \approx x_g \mid k_{i+1}) = \frac{\mathcal{P}_H(X \approx x_g \mid k_i) \,.\, \mathcal{P}_R(y_{i+1} \mid k_0, x_g)}{\mathcal{P}_R(y_{i+1} \mid k_i)} \qquad (12\text{-}5)$$

That is, on making a new observation y_{i+1}, we obtain the updated estimate of $\mathcal{P}_H(X \approx x_g \mid k_{i+1})$, by taking the value of the term $\mathcal{P}_H(X \approx x_g \mid k_i)$ from the result of the previous calculation and we compute the values of the terms $\mathcal{P}_R(y_{i+1} \mid k_0, x_g)$ and $\mathcal{P}_R(y_i \mid k_i)$ to substitute in *(12-5)*. Also, it follows from equation *(3-51e)*, that the total probability of observing y_{i+1}, as represented in the denominator of *(12-5)* is found by integrating over all the possible values, the product:-

$$\mathcal{P}_H(X \approx x_g \mid k_i) \,.\, \mathcal{P}_R(y_{i+1} \mid k_0, x_g)$$

expressed in the numerator, *i.e.* the probability of $X \approx x_g$, given the information contained in k_0 and the previously observed y_1.....y_i, multiplied by the probability of observing y_{i+1} if $X \approx x_g$. That is, although the errors affecting the observations are independent, the fact that we have previously observed a value such as y_i exerts a definite influence on the probability, relative to our knowledge, that the next observation will be reasonably close to y_i.

For example, we can imagine a group of children who are engaged in a game of guessing the height of a telegraph pole and the variation in their guesses is known rarely to exceed *1 metre*. Then, if the first child guesses a height of say, *7 metres*, it is very probable that the next child will guess a height between say, *6 metres* and *8 metres*. Thus, when we are given y_i and the form of the error distribution, the prior probabilities attaching to the various possible values of y_{i+1} are thereby affected. Therefore, replacing the denominator in *(12-5)* by the integral of the numerator, as in *(3-51f)*, the resulting distribution of probabilities over the possible values of X is given by:-

$$\mathcal{P}_H(X \approx x_g \mid k_{i+1}) = \frac{\mathcal{P}_H(X \approx x_g \mid k_i) \,.\, \mathcal{P}_R(y_{i+1} \mid k_0, x_g)}{\int_{-\infty}^{\infty} \mathcal{P}_H(X \approx x_g \mid k_i) \,.\, \mathcal{P}_R(y_{i+1} \mid k_0, x_g) dx} \qquad (12\text{-}6)$$

Yet, although there are many instances in practical life where we are required, as above, to estimate the value of a fixed parameter, there are also

important cases where the parameter in which we are interested is not fixed but changes from observation to observation according to some law or distribution of state-transition probabilities. Thus, using $\mathcal{P}_F(.)$ to denote the forecasting function, we may symbolise such transitions by $\mathcal{P}_F(X_{i+1} \approx x_g \mid X_i = x_f)$.

To model this more complex example of a Markov process, we modify *(12-6)* to reflect the fact, following a series of observations $y_1 \ldots\ldots y_i$ the resulting probability that X_i has a value x_g is given by: -

$$\mathcal{P}_H(X_i \approx x_g \mid k_i)$$

$$= \frac{\mathcal{P}_H(X_i \approx x_g \mid k_{i-1}) \cdot \mathcal{P}_R(y_i \mid k_0, x_g)}{\mathcal{P}_R(y_i \mid k_{i-1})} \qquad (12\text{-}7)$$

However, to evaluate *(12-7)* we have to go back to the posterior distribution over the previous state X_{i-1} at the point where we had received and evaluated the previous observation to determine $\mathcal{P}_H(X_{i-1} \approx x_f \mid k_{i-1})$. Given this information, we use the forecasting function $\mathcal{P}_F(.)$ to project the prior probability that the next object X_i will have a value x_g. This requires us to consider all combinations of all possible values of both X_{i-1} and of X_i, which we designate by the intermediary variables x_f and x_g respectively. Notionally, we start by fixing an extreme, but possible, value for x_f and, having done so, we sweep through all the possible values of x_g and, at each point[1], we compute the probability of the hypothesis $(X_{i-1} \approx x_f \mid k_{i-1})$ and the probability of a transition to the state $X_i \approx x_g$ if $X_{i-1} \approx x_f$. Multiplying together these probabilities gives us the contribution to the prior probability that $X_i \approx x_g$ due to the possibility that X_{i-1} might have had the value x_f. Hence, by integrating these contributions over all possible values of x_f, we obtain the total prior probability appropriate to the postulated value of x_g. Therefore we have to perform, or be able to approximate these computations for all possible values and combinations of x_f and x_g. Hence, the prior distribution over all the possible values of X_i is given by:-

$$\mathcal{P}_R(X_i \approx x_g \mid k_{i-1})$$

$$= \int_{-\infty}^{\infty} \mathcal{P}_F(X_i \approx x_g \mid X_{i-1} = x_f) \cdot \mathcal{P}_H(X_{i-1} \approx x_f \mid k_{i-1}) dx_f \qquad (12\text{-}8)$$

[1] See the comment in the 'Notation' section, regarding our treatment of the probability at a point on a continuous distribution.

and the full expression describing the posterior distribution, after the observation y_i is:-

$$\mathcal{P}_H(X_i \approx x_g \mid k_i)$$
$$= \int_{-\infty}^{\infty} \mathcal{P}_F(X_i \approx x_g \mid X_{i-1} = x_f) \cdot \mathcal{P}_H(X_{i-1} \approx x_f \mid k_{i-1}) dx_f$$
$$\times \frac{\mathcal{P}_R(y_i \mid k_0, x_g)}{\mathcal{P}_R(y_i \mid k_{i-1})} \qquad (12\text{-}9)$$

Although the approach outlined above is conceptually pure, such generalised approaches have hitherto received little attention for two remarkably different, but connected reasons. The first and probably major reason, stems from the computational demands in terms of both memory and processing effort which made such approaches almost impossible to conceive and totally impossible to implement prior to the advent of electronic digital computers. The second reason concerns the analytic intractability of the procedures, *i.e.* the difficulty of showing by rigorous deduction that they possess properties which are often considered desirable, a typical example being the yielding of an 'unbiased' estimate.

In contrast, however, to the generalised procedure, it emerged early in the nineteenth century, following the efforts of Simpson, de Moivre, Legendre, Bayes[1], Laplace, Gauss[2], and, no doubt, many others, that the assumption of a 'normal' distribution of errors directly justified both the taking of the arithmetic mean as the most probable[3] value of a fixed quantity and also the use of the method of 'least squares' to estimate the line of best fit to a linear trajectory. The consequent procedures are both minimally demanding of computational effort and are amenable to clear algebraic analysis.

One hundred years after Gauss, Fisher generalised and formalised such approaches in his notable *1922* paper. In that paper, Fisher defined and

---

[1] Not in 'The Essay', which was little read at that time, but in conversation and correspondence. See Stigler (1986a)

[2] See Stigler, (1986a)

[3] The reasoning which lead to the 'most probable' value seems to have been based on the belief that it was necessary to assume a uniform prior over the possible values of the true magnitude. Keynes seems to find this acceptable for the analysis of errors, *(Keynes (1921) p 196)*, but unacceptable for the inversion of Bernoulli's theorem, or Bayes' experiment, *(ibid. p 387)*. Fisher found it totally unacceptable.

explained numerous concepts which are of great utility in the algebraic analysis of statistical procedures, in particular the concept of a 'sufficient statistic'. The essence of this concept is that, in cases where a population is known to be, or, as is perhaps more often the case, is assumed to be, distributed according to a simple algebraic expression, a small set of parameters, computed from the observed values, is notionally sufficient to summarise all the information provided by the observations. This is, in modern terms, a form of 'data compression' and it is enormously useful in simplifying statistical calculations. In particular, where we have to make repeated observations of an unknown object, and each observation is corrupted by error, it leads to procedures which, being synoptic and recursive, can be executed in simple electronic circuits, or algorithms, and thus produce, with minimum demands on the computational memory, progressively refined distributions of resultant probabilities over the values in which we are interested.

The synoptic property of these procedures stems from the fact that the existence of 'sufficient statistics' allows us, at least nominally, to summarise all the information in a small set of parameters. The recursive property stems from the application of Bayes-Markov techniques where, on receipt of a new observed value, the previously estimated distribution of probabilities is treated as a prior distribution and is then combined, according to Bayes' theorem, with the new observation to produce a further and more refined estimate of the probability distribution over the possible values of the attribute in which we are interested.

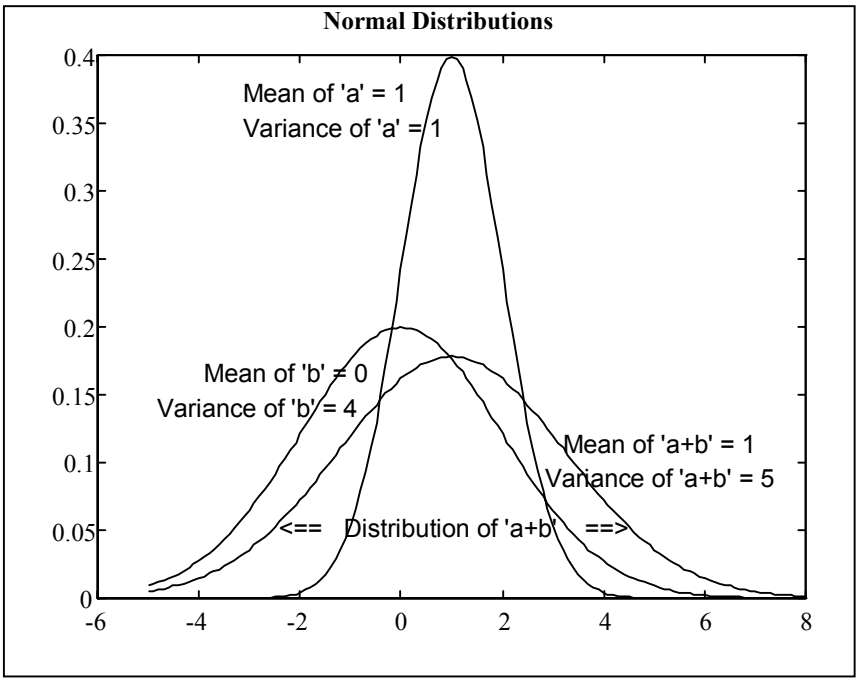


Figure 12.1

The key features are therefore the combination of Bayes' method, with a Markov process and the concept of synoptic statistics, such that the prior probabilities over a new state are fully determined by our knowing the one or two parameters which summarise the probability distribution over the previous state and the assumed transition function.

Some notable cases to which these procedures can be applied, and which are often encountered in contemporary practice, arise when we know, or, more often, we assume, again for computational and analytic convenience, that the errors on the observations have independent 'normal' distributions, similar to those illustrated in figure *12.1*. This is a technique with which some readers will be so familiar that it may seem to them strange that anyone who does not share that familiarity should read this book. Our concern however, being with matters of logic and philosophy, makes it possible that the readers will include others who may have no such familiarity. Therefore, rather than relegate the matter to footnotes and appendices, we take here a few lines to set the scene for the examples which follow.

Historically, the name[1] 'normal' seems to have become attached to these curves, by accident; an alternative name is 'Gaussian': but neither name

[1] The name is discussed in Stigler (1980)

is easily justified. In this chapter, we reluctantly use the name 'normal', but caution that in this usage, 'normal' is a peculiar proper name and should not be taken to mean that the curve is normal in any ordinary sense. We also find that the name is sometimes loosely applied to any bell-shaped curve, of the general shape illustrated in figure *12.1*, which can be defined, or approximated, by an equation of the form:-

$$\mathcal{P}(x_r) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp \frac{-(x-\hat{x})^2}{2\sigma^2} \qquad (12\text{-}10)$$

This equation has many convenient properties; but it is important to note that, while many other equations generate bell-shaped curves, they do not, in general, have the convenient properties of *(12-10)*, where the curve is defined by the two parameters $(\hat{x},\sigma)$. However, one can often find values for the parameters $(\hat{x},\sigma)$ which will produce an adequate fit, for practical purposes, to a set of points which lie roughly on such a curve. In the central portion, this is usually fairly easy; the problem becomes more difficult as one involves the tails. These symbols differ slightly from common usage; the choice of $\hat{x}$ to denote the central (mean) value is unusual: $'\mu'$ being more common. The symbol σ denotes the half-width of the curve at the points of inflexion which occur symmetrically about the mean value; it is often called the 'standard deviation' of the distribution. However, to keep the wording as simple as possible, we later write of the 'width' of a normal distribution being 'defined by σ'. The phrase should therefore be understood in the sense that the distance between the points of inflexion is equal to 2σ. The value of σ^2, known as the 'variance', is equal to the mean-square deviation from the mean value of the distribution. In statistical writing, it is common to define a normal curve by the mean and the variance, *e.g.* $(\hat{x},\sigma^2)$: here, however, in order to reduce the symbols to the simplest possible form, we write $(\hat{x},\sigma)$ and we symbolise the fact that a quantity has a normal distribution by an expression of the form $\sim\mathcal{N}(\hat{x},\sigma)$. We also use expressions of the form $\mathcal{G}(\hat{x},\sigma,x)$ as abbreviations for the full form of *(12-10)*; the symbol '$\mathcal{G}$' being to acknowledge the 'Gaussian' name.

Looking briefly at some properties of the normal distribution, we find that an expression of the simple form:

$$y = \exp \frac{-(x-\hat{x})^2}{2\sigma^2} \qquad (12\text{-}11)$$

has many of the properties sought by early investigators for the 'law of error' with which they were largely concerned. A theoretical defect, however, is that its value does not fall (quite) to zero beyond a certain point, but it does

fall to a value which is often small enough to be negligible for practical purposes.

Another problem is that the area under such a curve, representing the total probability of all possible values, does not integrate to unity but rather to the value $\sqrt{(2\pi\sigma^2)}$. However, while it is merely a convention that the total probability should sum to unity, it is an extremely useful convention, and since the problem can be regarded merely as a matter of adjusting a scale, the factor $1/\sqrt{(2\pi\sigma^2)}$ is introduced to produce the required result. If this strikes the reader as a somewhat cavalier act, it should be acknowledged that this whole field is populated by arbitrary and *'ad hoc'* acts of a similar kind, their purpose having been, historically, to produce algorithms which would yield answers of practical value from the computing devices which were available at the time.

Further, a remarkably useful property of the normal distribution arises because probabilities relating to the addition of variables or the joint occurrence of events, are easily computed by simple operations on the parameters. Thus, we have the simple theorems:-

Theorem 12.a *: A variable which is the sum of two normally-distributed variables is also normally distributed*

That is, if variables *a* and *e* are normally distributed such that:-

$$\mathcal{P}_R(a \approx x \mid k_0) = G(\hat{a},\sigma_a,x) \quad \text{and} \quad \mathcal{P}_R(e \approx x \mid k_0) = G(\hat{e},\sigma_a,x)$$

then their sum is also normally distributed such that:-

$$\mathcal{P}_R(a+b \approx x \mid k_0) = G((\hat{a}+\hat{e}), \sigma_{a+e}, x)$$

where $\sigma_{a+e} = \sqrt{(\sigma_a^2 + \sigma_b^2)}$. This is illustrated in Figure 12.1 which shows the broadening of the resultant distribution when variables are added arithmetically[1].

Theorem 12.b *: A variable which is a direct multiple of a normally-distributed variable is also normally distributed*

That is, if variable *a* is normally distributed with probability:-

$$\mathcal{P}_R(a \approx x \mid k_0) = G(\hat{a},\sigma_a, x)$$

and there is a second variable *e* such that $e = ca$, then *e* will be normally distributed such that:-

$$\mathcal{P}_R(e \approx x \mid k_0) = G(c\hat{a}, c\sigma_a, x) \qquad (12\text{-}12)$$

[1] This has to be distinguished from the narrowing which occurs when *information* is added, as shown by the resultant variance in equation *(12-19)*

Resuming now the main theme, if we assume that the errors on a series of observations $y_1 \ldots y_i \ldots y_n$ have independent normal distributions, and are symmetric about the true value of X, we can write:-

$$y_i = X + \varepsilon_i ; \quad \varepsilon_i \sim \mathcal{N}(0,\sigma_i),$$

and this assumption or knowledge we include in k_0. Hence the probability that an error ε_i has a value $'e'$ is denoted by:-

$$\mathcal{P}_R (\varepsilon_i \approx e \,|\, k_0) \;=\; \mathcal{G}(0,\sigma_i,e) \qquad (12\text{-}13)$$

In such cases, when we are given the first observation y_1 and the width of its error distribution $2\sigma_1$, the probable correctness of a guess that $X \approx x_g$ is equivalent to guessing that an error $\varepsilon_1 \approx y_1 - x_g$ has occurred. Hence:-

$$\mathcal{P}_H(X \approx x_g \,|\, k_0,y_1,\sigma_1) \;=\; \mathcal{P}_R(\varepsilon_1 \approx (y_1 - x_g) \;|\; k_0)$$
$$= \; \mathcal{G}(0,\sigma_1,(y_1 - x_g)) \qquad (12\text{-}14)$$

Following the second observation y_2, we need to compute, as in *(12-6)* the probability of $X \approx x_g$ given that we have now observed, with unknown errors, both y_1 and y_2. Thus:-

$$\mathcal{P}_H(X \approx x_g \,|\, k_2) = \frac{\mathcal{P}_H(X \approx x_g \,|\, k_1) \;.\; \mathcal{P}_R(y_2 \,|\, k_0 , x_g)}{\mathcal{P}_R(y_2 \,|\, k_1)} \qquad (12\text{-}15)$$

Taking the terms of *(12-15)* in order, we have:-

(1) from *(12-12)*:-

$$\mathcal{P}_H (X \approx x_g \,|\, k_1) \;=\; \mathcal{G}(0,\sigma_1,(y_1 - x_g)) \qquad (12\text{-}16)$$

(2) $\mathcal{P}_R(y_2 \,|\, k_0, x)$ is the probability of occurrence of an error $e_2 = y_2 - x_g$, whence:-

$$\mathcal{P}_R(y_2 \,|\, k_0, x_g) \;=\; \mathcal{G}(0,\sigma_2,(y_2 - x_g)) \qquad (12\text{-}17)$$

(3) following *(12-6)*, we have for the denominator:-

$$\mathcal{P}_R(y_2 \,|\; k_1) \;=\; \int_{-\infty}^{\infty} \mathcal{G}(0,\sigma_1,(y_1 - x_g)) \; \mathcal{G}(0,\sigma_2,(y_2 - x_g)) \, dx \qquad (12\text{-}18)$$

and therefore, combining *(1),(2),(3)* the resultant probability that $X \approx x_g$ simplifies as shown in Appendix E, to a normal distribution of the form:-

$$\mathcal{P}_H (X \approx x_g \;|\; k_0,y_1,y_2) \;=\; \mathcal{G}(\hat{y}_2, \hat{\sigma}_2 , x_g) \qquad (12\text{-}19)$$

where, the *'hat'* symbols denote synoptic parameters:-

$$\hat{\sigma}_2^2 = \frac{\sigma_1^2.\sigma_2^2}{(\sigma_1^2+\sigma_2^2)} \text{ and } \hat{y}_2 = \frac{\sigma_2^2 y_1 + \sigma_1^2 y_2}{\sigma_2^2 \sigma_1^2} \qquad (12\text{-}20)$$

The reasoning can then be generalised to show that as further observations are made, the information which they contain can be expressed by the synoptic parameters which are produced by each iteration, that is:-

$$\mathcal{P}_H(X \approx x_g \mid k_i) = \mathcal{G}(\hat{y}_i, \hat{\sigma}_i, x_g) \qquad (12\text{-}21a)$$

where $\hat{\sigma}_i^2 = \frac{\sigma_{i-1}^2.\sigma_i^2}{(\sigma_{i-1}^2+\sigma_i^2)}$ and $\hat{y}_i = \frac{\sigma_i^2 y_{i-1} + \sigma_{i-1}^2 y_i}{\sigma_i^2 \sigma_{i-1}^2}$ *(12-21b)*

Having thus dealt with normally distributed errors in the case of a fixed attribute, we now illustrate a case where we know or assume a transition function such that the true value of the attribute X_i comprises a known multiple of the true value of its predecessor, cX_{i-1} , plus a random variation ν_i so that $X_i = cX_{i-1} + \nu_i$. We also assume that the random variation has a normal distribution, $\mathcal{N}(0,\sigma_{vi}^2)$, centred in each case on the systematic component cX_{i-1} but having a width which can change from case to case. Each observation y_i also remains subject, in addition, to a random, normally distributed error ε_i which has a zero mean value and a width $2\sigma_{\varepsilon i}$. In this scenario, therefore, each observed value, after the first observation, can be regarded as the sum of three components:-

$$y_i = cX_{i-1} + \nu_i + \varepsilon_i \qquad (12\text{-}22)$$

However, because it is also true to make the equation $y_i = X_i + \varepsilon_i$ it follows by simple arithmetic that $X_i = y_i - \varepsilon_i$ and therefore, taking into account only a single observed value y_i, the probability that the true value X_i at that point was equal to some guessed value x_g is the probability that the error ε_i should have been equal to $y_i - x_g$. That is:-

$$\mathcal{P}_H(X_i \approx x_g \mid k_0, y_i) = \mathcal{P}_R(\varepsilon_i \approx (y_i - x_g) \mid k_0, y_i)$$
$$= \mathcal{G}(y_i, \sigma_{\varepsilon i}, x_g) \qquad (12\text{-}23)$$

At the first observation therefore, because there is no predecessor, the probability distribution over the possible values of X_1 is given by *(12-16)* . However, with a view to further generalisation, we can also write:-

$$\mathcal{P}_H(X_1 \approx x_g \mid k_0, y_1) = \mathcal{G}(\hat{y}_1, \hat{\sigma}_1, x_g) \qquad (12\text{-}24)$$

where $\hat{y}_1 = y_1$, the mean of the posterior distribution, and the spread is given by $\hat{\sigma}_1 = \sigma_{\varepsilon 1}$. At the next step, we observe y_2 and require to compute the consequent distribution over the possible values of X_2, taking also into

account the information provided by y_1 and the prior knowledge k_0. That is[1]:-

$$\mathcal{P}_H(X_2 \approx x_g \mid k_2) = \frac{\mathcal{P}_R(X_2 \approx x_g \mid k_1) \cdot \mathcal{P}_R(y_2 \mid k_0, x_g)}{\mathcal{P}_R(y_2 \mid k_1)} \qquad (12\text{-}25)$$

Taking first the term $\mathcal{P}_R(X_2 \approx x_g \mid k_1)$ we use the transition function $\mathcal{P}_F(.)$ and, at least notionally, a guessed value x_g to compute the prior distribution over the possible values of X_2, given that our information k_1 now includes both k_0 and y_1, *viz*:-

$$\mathcal{P}_R(X_2 \approx x_g \mid k_1) = \int_{-\infty}^{\infty} \mathcal{P}_F(X_2 \approx x_g \mid X_1 = x_f) \cdot \mathcal{P}_H(X_1 \approx x_f \mid k_1) dx_f \qquad (12\text{-}26)$$

However, the assumptions of independent normal distributions over ν_i and ε_i allow us to escape from the need to deal explicitly with the integration because the forecasting function, combined with some standard theorems, allows us to proceed directly from *(12-22)* to the result we require in *(12-23)*. This simple resolution of what may otherwise be somewhat complicated, stems from the reasoning that, because we have an estimate $\hat{y}_1$, from the first observation, which is normally and symmetrically distributed $\sim\mathcal{N}(0, \hat{\sigma}_1)$ with respect to X_1, we can project ahead to the second observation and denote a variable δ_2 which, notionally, comprises:- *(i)* the projection of the difference between the first observed value and the first true value, plus, *(ii)* the random variation ν_2 on the true value of the second object. That is:-

$$\delta_2 = c(y_1 - X_1) + \nu_2 \qquad (12\text{-}27)$$

Therefore, as we know that the value of the term $(X_1 - y_1)$ is distributed $\sim\mathcal{N}(0, \hat{\sigma}_1)$ it follows from *Theorem 12b* above that the value of the term $c(X_1 - y_1)$ will be normally distributed $\sim\mathcal{N}(0, c\hat{\sigma}_1)$. Also, as ν_2 is, by definition, normally distributed $\sim\mathcal{N}(0, \sigma_{v2})$ it follows that δ_2 is the sum of two normally-distributed variables, each of which has a zero mean value, and therefore δ_2 will also be normally distributed about a zero mean with a spread $\sigma_{\delta 2} = \sqrt{(c^2\hat{\sigma}_1^2 + \sigma_{v2}^2)}$. Further, since the prior distribution which we are here seeking, $\mathcal{P}_R(X_2 \approx x_g \mid k_1)$, concerns a random event, and also because:-

[1] See equation *(12-9)* above.

$$X_2 = cX_1 + \nu_2 = c\hat{y}_1 - c(X_1 - \hat{y}_1) + \nu_2 = c\hat{y}_1 - \delta_2 \qquad (12\text{-}28)$$

we can proceed directly to show that the prior distribution of X_2, given k_0 and y_1 will be a normal distribution of width $2\sigma_{\delta 2}$ around the projected value cy_1, that is:-

$$\begin{aligned} &\mathcal{P}_R(X_2 \approx x_g \mid k_1) = \mathcal{P}_R((c\hat{y}_1 - \delta_2) \approx x_g \mid k_1) \\ &= \mathcal{P}_R(\delta_2 \approx (x_g - c\hat{y}_1) \mid k_1) \\ &= \mathcal{G}(c\hat{y}_1, \sigma_{\delta 2}, x_g) \end{aligned} \qquad (12\text{-}29)$$

Returning to *(12-22)*, we now consider the term, $\mathcal{P}_R(y_2 \mid k_0, x_g)$, where we know, as with the first observation, that the parameter $y_2 - X_2 = \varepsilon_2$ is normally distributed $\sim\mathcal{N}(0, \sigma_{\varepsilon 2})$, whence it follows that the probability $\mathcal{P}_R(X_2 \approx x_g \mid k_0, y_2)$ is given by:-

$$\mathcal{P}_R(X_2 \approx x_g \mid k_0, y_2) = \mathcal{G}(y_2, \sigma_{\varepsilon 2}, x_g) \qquad (12\text{-}30)$$

Further, since we know from *(3-51f)* that the denominator in *(12-25)* is equivalent to the integral of the numerator over all possible values of x_g it follows that we can express *(12-25)* as:-

$$\begin{aligned} &\mathcal{P}_H(X_2 \approx x_g \mid k_2) \\ &= \frac{\mathcal{G}(c\hat{y}_1, \sigma_{\delta 2}, x_g) \,.\, \mathcal{G}(y_2, \sigma_{\varepsilon 2}, x_g)}{\int_{-\infty}^{\infty} \mathcal{G}(c\hat{y}_1, \sigma_{\delta 2}, x_g) \,.\, \mathcal{G}(y_2, \sigma_{\varepsilon 2}, x_g)\, dx_s} \end{aligned} \qquad (12\text{-}31)$$

Therefore, as it is shown in Appendix E that an expression of the form of *(12-31)* results in a further normal distribution, we have:-

$$\mathcal{P}_H(X_2 \approx x_g \mid k_2) = \mathcal{G}(\hat{y}_2, \hat{\sigma}_2, x_g) \qquad (12\text{-}32)$$

where $\hat{\sigma}_2^2 = \dfrac{\sigma_{\delta 2}^2 \,.\, \sigma_{\varepsilon 2}^2}{(\sigma_{\delta 2}^2 + \sigma_{\varepsilon 2}^2)}$ and $\hat{y}_2 = \dfrac{\sigma_{\varepsilon 2}^2 \hat{y}_1 + \sigma_{\delta 2}^2 y_2}{\sigma_{\varepsilon 2}^2 \sigma_{\delta 2}^2}$ *(12-33)*

and therefore, by extension of the same reasoning:-

$$\mathcal{P}_H(X_n \approx x_g \mid k_n) = \mathcal{G}((\hat{y}_n, \hat{\sigma}_n, x_g) \qquad (12\text{-}34)$$

where, for $n = 1$:-

$$\hat{\sigma}_1 = \sigma_{\varepsilon 1} \text{ and } \hat{y}_1 = y_1; \qquad (12\text{-}35)$$

and, for $n > 1$:-

$$\sigma_{\delta n}^2 = c^2 \hat{\sigma}_{n-1}^2 + \sigma_{\nu n}^2 \qquad (12\text{-}36)$$

$$\hat{\sigma}_n^{\,2} = \frac{\sigma_{\delta n}^2 . \sigma_{\varepsilon n}^2}{(\sigma_{\delta n}^2 + \sigma_{\varepsilon n}^2)} \qquad (12\text{-}37)$$

$$\hat{y}_n = \frac{\sigma_{\varepsilon n}^2 \hat{y}_{n-1} + \sigma_{\delta n}^2 y_n}{\sigma_{\varepsilon n}^2 \sigma_{\delta n}^2} \qquad (12\text{-}38)$$

Although the above expressions may seem not exactly simple to a lay eye, Fig 12.2 shows the ease with which the expressions can be implemented in a recursive device, such as a Kalman filter[1].

In such applications, however, the use of a calibrated measuring device is fundamental and, where the errors are assumed to have a normal distribution, calculation of mean and variance from the calibration data is a matter of simple arithmetic - a further case of the apparently magical power of a normal distribution to dispose of awkward problems. Yet, despite the large following enjoyed by devices based on assumptions of normal distributions, there are people who have long been unhappy at the temptation, presented by the amenity of these devices, to the skimping of the analysis and to the over-simplifying of the physics and mathematics.

It is therefore a welcome fact that the speed and capacity of electronic computers now makes it practical to handle distributions and their integrals, such as *(12-9)* above, which were previously beyond practical possibility. However, instead of computing the values of the integrals by traditional techniques of numerical approximation, it is also possible to use 'Monte Carlo' simulation in order to obtain results which, while still approximate, appear often to be much faster and of an accuracy comparable to those obtained by traditional methods. We must however point out that there may be in a Monte Carlo procedure, a hidden postulate of the uniform prior. This assumption may often be entirely rational and defensible, provided it is made clear both to the perpetrator and to the reader.

[1] Kalman (1960). See also Norton (1986).

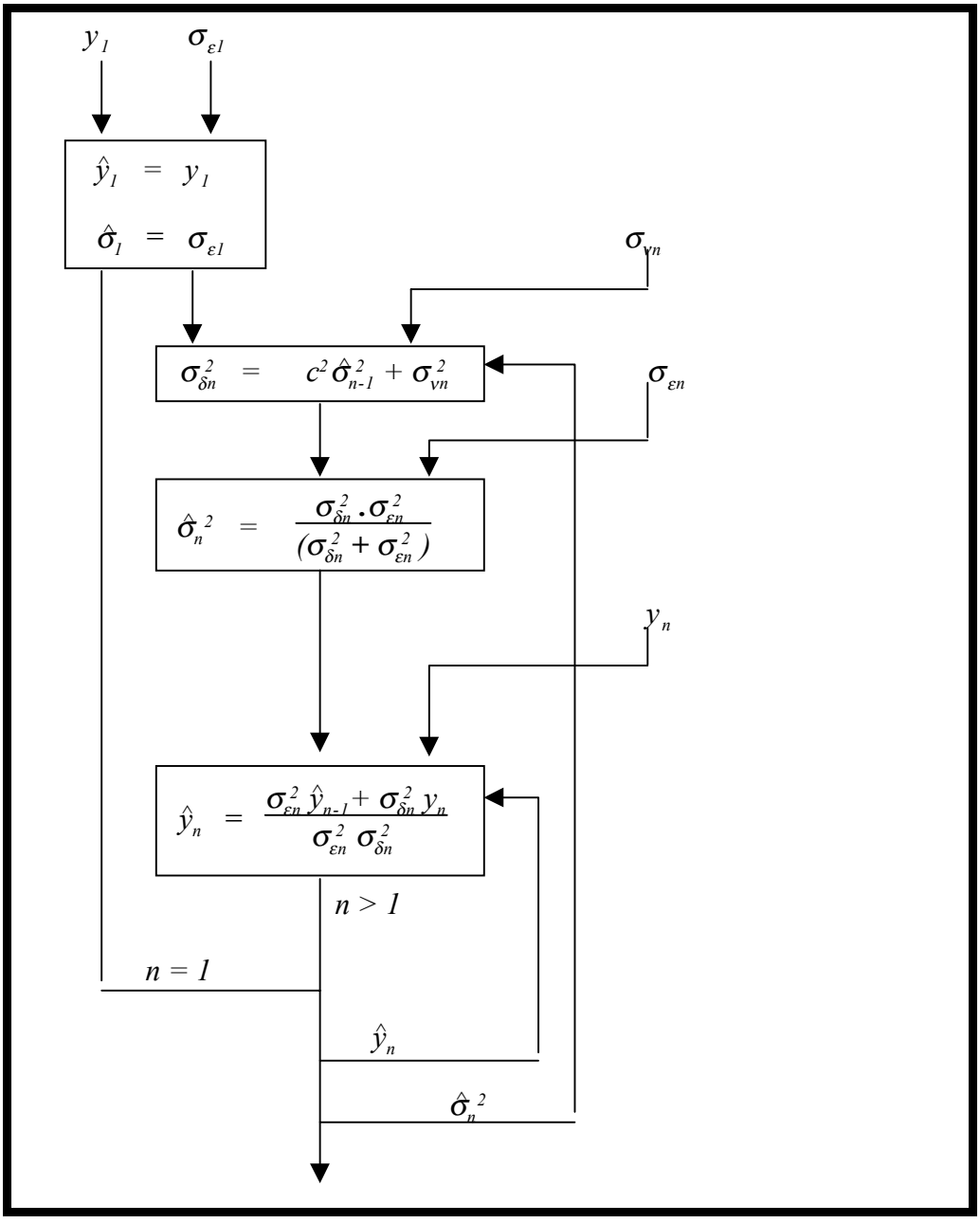


Fig 12.2

Chapter 13

Base Rate

Introduction

Although it may seem clear from the previous discussion that population statistics will, in general, provide no information about an individual, there is, it seems, at least one school of thought where the opposite view is held. The view that the population is indeed informative, seems to be especially common among cognitive psychologists, who are concerned to detect, understand and correct failings which are widespread and potentially serious, in the modern world, in our intuitive responses to situations which involve issues of probability and uncertainty. This concern has led to the publication, since the *1955* paper by Meehl and Rosen[1], of a considerable volume of books and papers, notable among which are: *"Judgement under Uncertainty"* by Kahnemann, Slovic and Tversky[2], Poulton's *"Behavioral Decision Theory[3],* and, of special interest in relation to this chapter, the papers by Bar-Hillel[4] Cohen[5] and Koehler[6]. Another author who gives excellent insights into the matter is Piattelli-Palmarini[7] whose very readable book shows many entertaining and enlightening examples of cases where we are all prone to cognitive slips. There is, however, a school of thought among the cognitive psychologists where, in our view, perceptions seem undesirably limited and over simplified. In this school of thought - the 'base rate school' - it is believed that, when computing a post-trial probability concerning an individual, it is axiomatic that the prior probability must be taken from the rate of occurrence among the population, group, or reference class[8] to which the individual belongs, *i.e.* the 'base rate.' Members of the school seem to believe that the use of the base rate is an integral part of what they refer to as

---

[1] Meehl and Rosen (1955)
[2] Kahnemann (1982)
[3] Poulton, (1994)
[4] Bar-Hillel, (1980), (1983), (1990)
[5] Cohen, L. J. (1979) *et al.*
[6] Koehler (1996)
[7] Piattelli-Palmarini, (1994).
[8] For definitions of the reference class, see Kyburg (1974) and (1983).

Bayes' Law[1]. It does however seem that few of the school may have actually read Bayes' essay, where the base rate plays no part in the derivation of the theorem[2], and Bayes hints at the use of the base rate only very hesitantly in the Scholium.

However, in the belief that the base rate is fundamental to Bayes' theorem, the school have produced a number of paradigm cases to illustrate what they believe to be common 'cognitive mistakes' where people ignore the base rate in the evaluation of test results and testimonies. The school have also, in conjunction with at least one reputable statistical society, set up a task group to indoctrinate young lawyers with their view of 'correct thinking' on this matter[3].

The Taxi Cab Case

One well-known paradigm used by the base rate school concerns a taxi cab which is involved in a hit-and-run accident[4]. There is just one witness who reports the colour of the cab as blue. However, there is in the city a total of *1000* cabs operated by two companies. One company operates *850* cabs, and they are green. The other company operates *150* blue cabs. Tests show that the witness has imperfect colour vision and gets the colour right on only *80%* of occasions. Given these facts, we are asked to compute the probability that the cab involved in the accident really was blue. As the base rate school make no distinction between probability in the frequentist $\mathcal{P}_R$ sense for the expected rate of occurrence of a random event and the $\mathcal{P}_H$ sense concerning hypotheses about a value that is fixed for the duration of a trial, they hold that the correct calculation has the form :-

$$\mathcal{P}(T_B \mid k,D_B) = \mathcal{P}(T_B \mid k) \cdot \frac{\mathcal{P}(D_B \mid k,T_B)}{\mathcal{P}(D_B \mid k)} \qquad (13\text{-}1)$$

where:-

T_B $\rightarrow$ *the true colour is blue*
k $\rightarrow$ *background data and assumptions*
D_B $\rightarrow$ *a blue cab is declared*

We now expand the denominator in terms of cases when the true colour is green, (T_G) and cases where it is blue, giving :-

$$\mathcal{P}(D_B \mid k) = \mathcal{P}(D_B \mid k,T_B).\mathcal{P}(T_B \mid k) + \mathcal{P}(D_B \mid k,T_G).\mathcal{P}(T_G \mid k)$$

[1] Piattelli-Palmarini, (1994).
[2] *cf.* Ch. 2, Prop 5; Ch.3 Prop 5.
[3] See : http://www.maths.nott.ac.uk/personal/rsscse/lawyersp1.html
[4] Piattelli-Palmarini, (1994), pp.83 and 207.

Assigning values :-

$$\mathcal{P}(T_B \mid k) = 0.15 \qquad \mathcal{P}(D_B \mid k,T_B) = 0.8$$

$$\mathcal{P}(T_G \mid k) = 0.85 \qquad \mathcal{P}(D_B \mid k,T_G) = 0.2$$

thus gives $\mathcal{P}(D_B \mid k) = 0.8x\ 0.15 + 0.2 \times 0.85 = 0.29$

and therefore

$$\mathcal{P}(T_B \mid k) \cdot \frac{\mathcal{P}(D_B \mid k,T_B)}{\mathcal{P}(D_B \mid k)} = 0.15 \times \frac{0.8}{0.29}$$

i.e. $\mathcal{P}(T_B \mid k,D_B) = 0.4138$ *(13-2)*

whence it is argued that there is a probability of only just over *40%* that the cab truly was blue.

If however we differentiate between $\mathcal{P}_R$ as the probability of a random event and and $\mathcal{P}_H$ as the probability of an hypothesis, and define:-

K_P	*the population statistics*
C(W)	*the calibrated performance of the witness*
k	*background data and assumptions*
K_1	= *(C(W), k), i.e. C(W) and k are both given*
K_2	= *(K_P, C(W), k) i.e. the population statistics and C(W) and k are given*

we find by analogy with *(13-2)* that $\mathcal{P}_R(T_G \mid K_2,D_B) = 0.4138$. If however we formulate precisely the question to which this result is the answer, we find that it is as follows:- *If all the cabs in the city were to be driven past the witness a large number of times, and the witness were asked in each case to identify the colour, then, in what proportion of the cases in which the colour was identified as blue, would we expect the identification be correct ?*

In contrast, the question we are really trying to answer concerns the probability that a specific cab, declared to have been blue, really was blue. Thus, given that only one cab was involved, it seems strange that *999* other cabs, which were not involved, are somehow deemed to provide evidence about the cab which actually was involved. However, as with other types of observation, a single observation in such a case is merely a sample from a potentially large set. That is, there could have been, in principle, any number of witnesses to the event in which the unique cab was involved. Such witnesses may then provide a set of randomly varying observations about the specific cab, from which our concern is to determine the probability of an hypothesis concerning the colour, which is a fixed value in the context of the accident. Therefore, as the single witness has been calibrated and the prob-

ability of error is asserted to be independent of the true colour of the taxi, we can, by analogy with the calibrated ruler, write:-

$$\mathcal{P}_H(E|D_B,K_1) = \mathcal{P}_R(E|K_1) \cdot \frac{\mathcal{P}_R(D|E,K_1)}{\mathcal{P}_R(D|K_1)} \qquad (13\text{-}3)$$

where *E* signifies that the witness was in error and *D* signifies simply that the witness declares a colour. Assuming then, as described in Chapter 9, that there has been no change in the error-probabilities since the calibration of the witness, we can set the ratio

$$\frac{\mathcal{P}_R(D|E,K_1)}{\mathcal{P}_R(D|K_1)} = 1$$

whence

$$\mathcal{P}_H(E|D_B,K_1) = \mathcal{P}_R(E|K_1) = C(W) = 0.2 \qquad (13\text{-}4)$$

and therefore, as the probability that the taxi truly was blue is the compliment of the probability that the witness was in error, we have:-

$$\mathcal{P}_H(T_B|D_B,K_1) = 1 - \mathcal{P}_H(E|D_B,K_1) = 0.8 \qquad (13\text{-}5)$$

that is, there is a probability of *80%* that the hypothesis that the taxi was blue is correct.

Clearly, therefore, in cases which are logically equivalent to that of the taxi cab, there can be marked differences between the value of $\mathcal{P}_R$ which is given by the base rate approach, taken over the whole population, and the value of $\mathcal{P}_H$ which is given by evaluating the probability of an hypothesis concerning a specific object, in relation to specific evidence.

A further point of special interest emerges if we consider the human aspect of the term $\mathcal{P}_R(D|E,K_1)$: that is, the probability that the witness will declare a colour, given that the declaration will be in error. For, a human witness, being conscious of a tendency to err in certain cases, may refuse to make a declaration if it is felt that the case in point is doubtful. Similarly, it is easy to envisage an automatic measuring device which will suppress the declaration of a measured value if the 'signal-to-noise' ratio is below some threshold of acceptability. Clearly, if we assume that the performance of the witness or measuring device is stable, the suppression of a declaration will not invalidate the calibration: the point is, rather, in clarifying the real-world meaning of an algebraic term that may be otherwise obscure.

Medical screening

Another case used as a paradigm by the base rate school[1] concerns a medical screening test which is loosely said to have a one-percent false positive rate. This appears to mean that if a large number of people who are clear of the condition are tested, then, on average, one percent of those people will produce a result which, wrongly, indicates that the condition is present. It is also assumed that the test never fails to indicate the condition if it is actually present. In addition, we are told that, taken over the whole population, only one person in *1,000* has the condition. It is then shown, by using these figures in conjunction with Bayes' theorem, that, if a person selected at random from the population is tested 'positive,' the probability that the person actually has the condition, is only about *9%*. In outline, the reasoning behind this conclusion is of the form that, if one million people, selected randomly, are tested, then *1,000* of those people will have the condition and will show positive. In addition the test will show positive for *1%* of the remaining *999,000* people, *i.e.* in nearly *10,000* cases. Thus, out of a total of *11,000* positive declarations, only *1,000*, or about *9%* will be correct.

To apply Bayes' theorem, we first have to ask whether the probability that is to be determined concerns an attribute which varies randomly throughout a trial, or whether it concerns the value of an attribute in an individual case that is fixed for the trial ? Clearly, in terms of the reasoning just shown, the concern is with an attribute which varies randomly as different members of the population are tested. We therefore define the symbols to be used:-

D_+ *a positive detection is declared*
D_{F+} *a false positive detection is declared*
D_{T+} *a true positive detection is declared*
T_+ *the true condition in the case is positive*
T_- *the true condition in the case is negative*
K_2 *the supporting data and assumptions:-*
- K_P *the population statistics*
- C_1 *the calibrated performance of the test in this context*
- k *background data and assumptions*

This gives for the average expected rate of occurrence across the population:-

[1] A closely similar case is given by Piattelli-Palmarini, (1994), pp.80 and 204-207.

$$\mathcal{P}_R(T_+ \mid D_+, K_2) = \mathcal{P}_R(T_+ \mid K_2) \cdot \frac{\mathcal{P}_R(D_+ \mid T_+, K_2)}{\mathcal{P}_R(D_+ \mid K_2)} \qquad (13\text{-}6)$$

We now expand the denominator in terms of cases when the true condition is positive and when it is negative:-

$$\mathcal{P}_R(D_+ \mid K_2) = \mathcal{P}_R(D_+ \mid T_+, K_2) \cdot \mathcal{P}_R(T_+ \mid K_2) \ldots\ldots + \mathcal{P}_R(D_+ \mid T_-, K_2) \cdot \mathcal{P}_R(T_- \mid K_2) \qquad (13\text{-}7)$$

whence

$$\mathcal{P}_R(T_+ \mid D_+, K_2) \qquad (13\text{-}8)$$

$$= \frac{\mathcal{P}_R(T_+ \mid K_2) \cdot \mathcal{P}_R(D_+ \mid T_+, K_2)}{\mathcal{P}_R(D_+ \mid T_+, K_2) \cdot \mathcal{P}_R(T_+ \mid K_2) + \mathcal{P}_R(D_+ \mid T_-, K_2) \cdot \mathcal{P}_R(T_- \mid K_2)}$$

Substituting:- $\mathcal{P}_R(D_+ \mid T_+, K_2) = 1.0$; $\mathcal{P}_R(T_+ \mid K_2) = 0.001$;
$\mathcal{P}_R(D_+ \mid T_-, K_2) = 0.01$; $\mathcal{P}_R(T_- \mid K_2) = 0.999$;

gives:- $\mathcal{P}_R(T_+ \mid D_+, K_2) = 0.090992 \qquad (13\text{-}9)$

That is, relative to the information and assumptions K_2 there is only a 9.1% probability that the condition will actually be present in a member of this population, selected at random, and for whom the test gives a positive result. Superficially, therefore, this result seems to contradict the premiss of the one percent false positive rate on which it is based and gives the appearance of a paradox.

The truth is, however, that the presentation of the problem is ambiguous and involves deceptive shifts of meaning. The appearance of self-contradiction, for example, stems from the fact that the false positive rate is initially defined as the ratio of positive declarations to total declarations when a large number of condition-negative people are tested. The final result of 9% is, in contrast, the ratio of correct positive results to total positive results when a large population of a mixed, but known, composition is tested.

Serious points also concern the definition of the false positive rate and its relevance to the individual. Introducing the problem, above, we took care to explain that a *"one-percent false positive rate"* appeared to mean that, if a large number of people who are clear of the condition are tested,

then, on average, *1%* will produce a false positive result. This is to be contrasted with an alternative definition that, if a particular condition-negative individual is tested a large number of times, then, on average, *1%* of those tests will show a false positive result. A significant feature of this alternative definition is that it relates to a fixed value, namely the true state of the individual under examination, and therefore supports a $\mathcal{P}_H$ statement of probability of the form:-

$$\mathcal{P}_H(T_+ \mid D_+,K_1) = \mathcal{P}_H(T_+ \mid K_1) . \frac{\mathcal{P}_R(D_+ \mid T_+,K_1)}{\mathcal{P}_R(D_+ \mid K_1)} \qquad (13\text{-}10)$$

where K_1 denotes:-

C_2 *the calibrated performance of the test in this context*
k *background data and assumptions*

Unfortunately, although the error-probability has been derived by calibration, the errors in this case are neither randomly added to a true value, nor is the probability of error independent of the true state of the individual. (For it is given in the definition of the problem that the test never fails to indicate the condition if it is actually present). Hence, the techniques which are available in the case of the calibrated ruler, are not available here: but the facts do allow us to set the value $\mathcal{P}_R(D_+ \mid T_+,K_1) = 1$.

To evaluate the term $\mathcal{P}_R(D_+ \mid K_1)$ in the denominator of *(13-10)*, requires expansion in terms of the mutually exclusive positive or negative true state of the individual. This invokes the prior probabilities $\mathcal{P}_H(T_+ \mid K_1)$, and $\mathcal{P}_H(T_- \mid K_1)$ but excludes any reference to the population statistics. However, because the positive and negative states T_+ and T_- are mutually exclusive and exhaustive, $\mathcal{P}_H(T_- \mid K_1) = 1 - \mathcal{P}_H(T_+ \mid K_1)$. The (assumed) absence of information concerning the individual prior to the test therefore requires that we set the prior probability that the true condition is negative equal to the prior probability that the true condition is positive, *i.e.* $\mathcal{P}_H(T_- \mid K_1) = \mathcal{P}_H(T_+ \mid K_1) = 0.5$. Expansion of the denominator into components corresponding to these two possibilities and the calibrated performance of the test then gives:-

$$\begin{aligned} &\mathcal{P}_R(D_+ \mid K_1) \\ &= \mathcal{P}_R(D_+ \mid T_+, K_1).\mathcal{P}_H(T_+ \mid K_1) + \mathcal{P}_R(D_+ \mid T_-, K_1).\mathcal{P}_H(T_- \mid K_1) \\ &= 1 \times 0.5 + 0.01 \times 0.5 = 0.505 \qquad (13\text{-}11) \end{aligned}$$

which, substituted into *(13-10)* gives:-

$$\mathcal{P}_H(T_+ \mid D_+, K_1) = \mathcal{P}_H(T_+ \mid K_1) \cdot \frac{\mathcal{P}_R(D_+ \mid T_+, K_1)}{\mathcal{P}_R(D_+ \mid K_1)} \qquad (13\text{-}12)$$

$$= 0.5 \times \frac{1.0}{0.505} = 0.99 \qquad (13\text{-}13)$$

i.e. on this basis it appears there is a *99%* probability that a positive result is correct.

Three factors make important differences between these two results. The first factor is the distinction between the $\mathcal{P}_H$ probability of an hypothesis concerning a fixed value in testing an individual and the $\mathcal{P}_R$ probability of a random event in a group. The distinction explains why we are able to derive probabilities, in individual cases, from calibrated procedures without knowing or referring to a population from which the individual is deemed to have been drawn. This is a crucial re-assurance. Without this distinction, the estimation of probabilities after, for example, weighing a bag of unknown contents, would appear to be arbitrary or illogical, and would, in effect, deny the possibility of objective measurement.

The second factor is the difference between the calibrated variability of a test when applied repeatedly to a given individual and the variability which is encountered in testing different members of a group. In the former case, it is important to understand the reasons for the variability of any given test. In some cases, the variability may stem entirely from 'noise' in the measuring process. In such cases, the technique of the calibrated ruler may apply. Variability may also arise because a substance may be not quite homogeneously distributed *e.g.* within the blood stream.

In a different category, a wrong result from a test of an individual may stem from an imperfect correlation between the test criteria and the condition which is of interest. This is important for indirect medical screening procedures which are based on examination of samples for characteristics which are associated only statistically with a given condition but are simple, quick and inexpensive compared with direct examination. Indirect tests may also involve much less risk to a patient. However, in an indirect test, there may be no significant error in the measured value of the test parameter, but a wrong interpretation may stem from a lack of correlation between that value and the existence of the condition in the specific individual.

The third factor concerns differences between the supporting information and assumptions. In the group case, the statistics are essential but, in the individual case, they are excluded. This is vital for the first detection of a disease in a population where there have previously been no confirmed

cases. That is, if we were to assume a zero probability that a person selected at random would have the condition, then, regardless of the accuracy of the test, substitution of $\mathcal{P}_R(T_+ | K_2) = 0.0$ in *Eqn.(13-6)* will give $\mathcal{P}_R(T_+ | D_+, K_2) = 0$. This answer is clearly spurious and brings out a further, implicit assumption in the base rate theory *i.e.* that the base rate is fixed.

These factors are crucial for the understanding and use of the resultant probabilities. Bayes' theorem supports valid inference in both cases. Yet, whereas a direct observation of an individual provides evidence about that individual, a random probability derived by base rate thinking provides no evidence about an individual; its evidence is to the rarity or otherwise of an event within an arbitrary population.

The length of a pencil

We now turn to a problem concerning the length of a pencil. The pencil is chosen at random from a box containing *1000* pencils which have been cut precisely to a length of *200mm ±0.01mm* and one further pencil which has been cut precisely to a length of *201mm ±0.01mm*. The chosen pencil is placed in a calibrated measuring device for which there is a probability of *0.99* that the measurement error will be less than *±0.01mm* and a probability of *0.01* of an error of *+1mm ±0.01mm*. The measured length of the pencil is *201mm* and we need to know the probabilities, on the different methods of reckoning, that the true length of the pencil is, to within *±0.01mm*, either *200mm* or *201mm*.

Taking the base rate view, we will argue that, if all the pencils are measured, then we will expect *10* of the *200mm* pencils to be measured at *201mm* and the single *201mm* pencil also to be measured at *201mm*, with a slight possibility that it could be measured at *202mm*. On this basis, there is a *90%* probability that the pencil in the test-case is actually *200mm* in length and a *10%* probability that it is *201mm* in length. Taking the independent view, however, the corresponding probabilities are *1%* and *99%* - a fairly dramatic difference.

To this point, our reasoning is similar to that in the previous examples. Now, however, we make the point that the base rate view seems to be impacting on our ability to measure a pencil, using a calibrated device, and state the resultant distribution of probabilities over the possible values. That this is not truly the case can be seen by recalling that the base rate gives us

the $\mathcal{P}_R$ value on the basis that all the pencils are tested, whereas the independent view gives us the $\mathcal{P}_H$ value for a single, selected pencil.

For further illustration, the example can be extended to a case where the lengths of the pencils have a normal distribution $\mathcal{N}(\ell_o, \sigma_\ell)$ and the error probabilities also have a normal distribution $\mathcal{N}(0, \sigma_\varepsilon)$. If we then measure the length of a given pencil as *m*, we may reasonably ask what is the most probable true value ? On the base rate view, we have the probability of the random event that the true length of a pencil is *x*, given that its measured length is *m* :-

$$\mathcal{P}_R(\ell \approx x \mid k,m) = \mathcal{P}_R(\ell \approx x \mid k) \cdot \frac{\mathcal{P}_R(m \mid k,x)}{\mathcal{P}_R(m \mid k)} \qquad (13\text{-}14)$$

Using the notation described in Chapter 12 for normally distributed variables, the terms of the numerator can be immediately written:-

$$\mathcal{P}_R(\ell \approx x \mid k) = \mathcal{G}(\ell_o, \sigma_\ell, x) \qquad (13\text{-}15a)$$

$$\mathcal{P}_R(m \mid k,x) = \mathcal{G}(0, \sigma_\varepsilon, (m-x)) \qquad (13\text{-}15b)$$

Evaluation of the denominator $\mathcal{P}_R(m \mid k)$ requires, however, that we compute the prior probability, taken over all the pencils and all the possible errors, that the sum of the true length of a randomly selected pencil and the error on the measurement will equal the measured length, that is:-

$$\mathcal{P}_R(m \mid k) = \int \mathcal{P}_H(\ell = x \mid k) \cdot \mathcal{P}_R(\varepsilon = m - x)\, dx$$

$$= \int \mathcal{G}(\ell_o, \sigma_\ell, x) \cdot \mathcal{G}(0, \sigma_\varepsilon, (m-x))\, dx \qquad (13\text{-}16)$$

Hence, substituting into *(13-14)*:-

$$\mathcal{P}_R(\ell \approx x \mid k,m) = \frac{\mathcal{G}(\ell_o, \sigma_\ell, x) \cdot \mathcal{G}(0, \sigma_\varepsilon, (m-x))}{\int_{-\infty}^{\infty} \mathcal{G}(\ell_o, \sigma_\ell, x) \cdot \mathcal{G}(0, \sigma_\varepsilon, (m-x))\, dx} \qquad (13\text{-}17)$$

$$= \frac{\mathcal{G}(\ell_o, \sigma_\ell, x) \cdot \mathcal{G}(m, \sigma_\varepsilon, x)}{\int_{-\infty}^{\infty} \mathcal{G}(\ell_o, \sigma_\ell, x) \cdot \mathcal{G}(m, \sigma_\varepsilon, x)\, dx} \qquad (13\text{-}18)$$

Therefore, as shown in Appendix E, putting

$$\sigma_c^2 = \sigma_\ell^2\sigma_\varepsilon^2/(\sigma_\ell^2 + \sigma_\varepsilon^2)$$

and $c = (\ell_o\sigma_\varepsilon^2 + m\sigma_\ell^2) / (\sigma_\ell^2 + \sigma_\varepsilon^2)$

gives

$$\mathcal{P}_R(\ell \approx x \mid k,m) = \mathcal{G}(c, \sigma_c, x) \qquad (13\text{-}19)$$

which has a maximum when $x = c$. Hence, the 'most probable' value, on the base rate view, is a weighted average between the mean of the pencil population and the value actually measured. It follows that, the further the measured value is from the mean of the population, the greater will be the difference between the measured value and the apparently 'most probable' value. That is, the 'accuracy' of the measuring process is apparently affected by a relationship between the set from which the pencil is drawn and the actual length of the pencil. This has the remarkable implication that we could move the pencil to a box where its length is very close to the most common value, then take it back to the same measuring device, using the base rates of the new box, and obtain a 'most probable' value extremely close to the measured value. The silliness in such reasoning is a simple consequence of the failure to distinguish between probability as the $\mathcal{P}_R$ frequency of a random event in a population, and the probability of an hypothesis in the $\mathcal{P}_H$ sense about a specific object and specific evidence concerning that object. The value of c derived by base rate thinking actually tells us that, if we were to measure all the pencils many times, and consider all the different combinations of true length and error which could give rise to a measured length m, then the most common combination would be that in which the true length was c and the error $m - c$. That was not however the question : we were asked to measure a specific pencil and state its most probable length. We also have to bear in mind that a single measurement of a given object is but a representative of a set of measurements of that same object. To enlarge the set, we might well ask a number of different people to measure the pencil and we might well use a number of different rulers: but no sane person would set about measuring other pencils in order to produce a better answer for the pencil in hand.

Thus, it is clear that, not only does the base rate view fail to distinguish the expected frequency of a random event from the probability of an hypothesis concerning a specific case, it also fails to acknowledge the fundamental rôle of direct and repeatable observations in the specific case. For, without the independence and objectivity of specific observation, there can

be no knowledge of a base rate. The base rate view requires, and tacitly assumes, the pre-existence of independent observation.

Yet this does lead to a further question. Suppose that we measure each pencil in a consignment of *10,000* pencils, using a calibrated device which has a normal error distribution, and then construct histograms showing the measured rates in different bands: the fact is that the measured rates will generally not be the true rates. Fortunately, in many cases, we can use repetitive independent measurements in order to improve the accuracy of the independent process. For example, if the errors on a single measurement of a pencil have a Gaussian distribution with a zero mean and $\sigma = 0.2mm$, then averages of two independent measurements will also have a Gaussian, but sharper, distribution with a zero mean and a width of $\sigma = 0.144mm$. Nevertheless, the accuracy with which a base rate can be known remains subject to the uncertainties on individual measurements and the consequent distribution of probabilities over hypotheses concerning the possible true values[1].

Aviation and medicine

The practical implications of the above considerations are deeply important in many areas. If base rate thinking were applied directly to aircraft navigation and control, the chances of safely completing any flight would be slim indeed, for we would be relying not only on what the instruments tell us about this particular flight, but also on flight plans and instrument readings relating to other aircraft. In fact, it is only with the assured objectivity of calibrated instruments[2], that we can relate what is known to be safe practice in aviation with the mathematical fact of Bayes' theorem. In aviation, the logical status of a calibrated instrument, with the vital term reflecting the probability that the instrument is still working as calibrated, is reflected in the duplication of instruments and in the training of aviators to constantly cross-check the coherence of the readings.

Base rates are, however, by no means irrelevant to the safety of aviation. If an aircraft, on taking off, appears, according to an altimeter, not to be gaining height as would be normally expected, given the power settings, attitude, airspeed *etc.,* the discrepancy should raise the question whether the altimeter might have failed. Even though the base rate for altimeter failures is extremely low, instances are known where this has occurred and led to terrible accidents. Thus, comparison of the base rate for take-off perform-

[1] This matter will be addressed in more detail in a future section on the number of measurements which are required to provide any defined degree of probability that a true rate is greater than or less than a defined value.
[2] Ch. 9, *"The Ruler"*

ance against the reading of the altimeter can provides a valuable warning of an unusual and possibly dangerous condition. In medicine, a doctor will, understandably, be cautious about declaring the existence of a rare and serious disease if general experience in the population is that the test often produces a false positive result and the disease is extremely rare. But a low true-positive rate in the population at large gives no a reason to equate that rate with the probability for a person who gives a positive result under test. There are many serious diseases where prior indoctrination with base rate thinking may tragically distract the doctor's attention from the surprising and unwelcome conclusion to which the individual evidence actually points.

False doctrine

There are therefore serious dangers in teaching students to believe that the base rate version of Bayes' theorem is a mathematical law, when it is merely a special case to which the theorem can be applied. It is also seriously wrong to teach the base rate doctrine while failing to point out the distinctions between probability in the $\mathcal{P}_H$ sense and in the $\mathcal{P}_R$ sense - probabilities concerning individuals and probabilities in populations. We also need to understand both the objectivity of calibrated instruments and the possibilities of drift and failure in such instruments.

Arbitrary populations

It should also be emphasised that the assignment of an individual to a population is often a totally arbitrary act and that, when we are dealing with an individual of any complexity, there are numerous populations, groups, or 'reference sets' which we can construct around that individual. It is a consequence of this little-acknowledged freedom to define arbitrary populations that Bayes' theorem can be too-easily used to provide an appearance of support for fallacious inference concerning an individual when the inference is in fact a projection of a property which is merely a characteristic of an arbitrary set.

Another important aspect, which seems little-acknowledged in base rate thinking, is that, if the probability of correct diagnosis really were dependent upon knowledge of the relevant base rate, it would be impossible to form any view of the probability in cases where the relevant base rate is not known. Further, whatever set is chosen, it will be possible to indicate other sets of which the individual is undoubtedly a member, and for which the base rate may be, in practice, utterly unknowable. These things present,

however, no obstacle to our knowing the probabilities over the possible hypotheses concerning the specific individual stemming from a specific observation. They are merely an obstacle to our knowing the frequency with which certain phenomena will be observed in repeated tests of a given set.

The objectivity of a calibrated process gives us coherent, disciplined scientific, engineering, medical and navigational practice. It also directs our attention to the independence and integrity of the individual.

The ace detector

We now consider what is, perhaps, the most difficult problem raised by our criticism of base rate thinking. The problem concerns the rational assignment of resources, following Bayes' definition of probability, which we discussed in Chapter 6. The cases of the taxi cab and the medical screening test, however, raise ethical issues where it is not easy to assign monetary values, *e.g.* to a human life or the wrongful conviction of an innocent person. To explore the question, we therefore postulate a set of trials involving a stack of playing cards and an hypothetical device known as an 'ace detector'. The total number of cards in the stack and the number of aces in the stack can be of any size we choose. The ace detector is a device which samples a number of points on the back of a card and, on that basis, reports 'ace' or 'not ace'. The detector also has the following properties:-

- it is designed to change, on each test, the points which are sampled and thus ensure that the probability of a false positive on any given test is not affected by the occurrence of a false positive on any other test.
- it has a zero false negative rate
- it can be tuned to give any desired probability of false positive detection.

We now set up a situation in which the stack comprises *1001* cards, of which just one card is an ace. We tune the detector to have a false positive rate of *1:4*; that is, when tested against a large number of non-ace cards, one quarter of those cards are, on average, wrongly reported as aces. A card is then taken at random from the stack, is tested and is given a positive assessment. We now require to consider how much it would be worth paying for a lottery ticket which will yield a prize of *$1000* if the card really is an ace.

On the base rate view, we reason that, if all *1001* cards were passed through the detector, there would be *251* positive declarations, of which just one would be correct. That is, by analogy with the base rate view of the medical screening test, $\mathcal{P}_R(T_+ \mid D_+, K_2) = 1/251 = 0.00398$ and it would

Version zh25a.doc at 18:36 on 13 July, 2001

therefore be worth paying a maximum of *$3.98* for a ticket[1]. Taking the independent view, however, that the base rate provides no evidence about any given card, we get, by similar analogy $\mathcal{P}_H(T_+ \mid D_+,K_1) = 0.75$ and it would be worth paying up to *$750* for a ticket. Clearly there is here a problem, and if this game were played a large number of times under the same conditions, anyone following the independent line, as just described, would probably lose a lot of money.

There are however many variations that we can make to this scenario. We may, for instance, tune the detector to a false positive rate of 10^{-6}, which makes a base rate of *1:1000* effectively irrelevant. If however the stack is increased in size to, say, *50,000,000* cards, of which just *50* are aces, the base rate again becomes significant in relation to the false positive rate and the appearance of conflict between the two views returns.

However, the appearance of conflict may actually stem from unstated assumptions and implications in the scenarios we have drawn. There are, for example, unstated assumptions of simplicity and financial symmetry in the lottery scenario. This becomes clear if, instead of considering just simple gains and losses, we consider penalties, rewards and costs for backing different views in a given situation. If the detector declares an ace and we back the detector and it is correct, there may be a reward of *$500*. If we refuse to back the detector and it is correct, there may be a penalty of *$1000*. If we ask for a further test, we may have to pay a fee of *$100*. If we do not make a decision within a certain time, there may be a penalty of *$2000*. And so on.

Fundamental point

Yet there is, in all this reflection, an implicit 'bottom line' which tells us that while, in many situations, both the base rate and the independent view can be relevant to decisions in which costs and penalties are involved, the independent view in the individual case is fundamental. There are two reasons for this. First, there is the dependence of a knowable base rate upon the assessment of individual cases. Second, there is the fact that, in principle, there will be a set of tests which can be integrated to provide any required degree of refinement and which will, in any individual case, overwhelm the base rate with precise evidence concerning the individual. Thus the threat of ruin can often be avoided by refining the independent tests to the point where, taking account of all the costs, benefits and risks, the

[1] Not being experienced gamblers, we are assuming that the price paid for the ticket is not returned in the event of winning.

ratio of false-positives to true positives is either favourable or is, at least, in a state of balance.

Complex decisions

Nevertheless, in the real world, refinement will often incur costs and cause delays. In medicine, particularly in public-funded, resource-limited situations, doctors are implicitly forced into an intuitive balancing of extremely complex issues of probability, cost and risk, taking into account all their patients. In privately-funded situations, where a patient's entitlement to treatment is more directly determined by individual insurance cover and the ability to pay, the responsibility for balancing the probabilities costs and risks is, in the last resort, a matter for the individual patient. We must however leave to others further discussion of the enormously complex moral, mathematical and political issues to which these reflections point.

Dangers of indoctrination

Finally, we return to the dangers in the indoctrination of people such as doctors and aircrew, who are often called upon to make rapid decisions under conditions of great stress. It is entirely desirable to encourage such people to understand the probabilities, risks and costs which have to be balanced. To propagate the idea that there is a single, simple and correct way of viewing all such situations is reprehensible and must inevitably lead to unnecessary accidents, suffering and litigation. Even more dangerous, is the strong possibility that simplistic rules of base rate thinking will be built into 'expert systems' and that people will be taught that, without question or reflection, such systems should be believed and obeyed.

Chapter 14

The Probable Cause

In previous chapters the discussion has related largely, if at times only implicitly, to issues of measurement. While such matters are important for our understanding of the theoretical basis on which much of modern life stands, there are many questions of probability which we encounter in the control of communications devices, in the operation of radars, in using clinical instruments and in a wide range of situations where we are concerned in one way or another with diagnosing a probable cause and deciding on some form of corrective, compensatory or retaliatory action. Some questions are debated prominently in the media when questions of causation, criminal guilt, and so forth seem likely to catch the attention of the public. While the concept of a cause has proved very difficult to analyse and define precisely to the satisfaction of the philosophical community, people at large, (certainly in 'Western' societies), seems to suffer no inhibitions on that account and simply take causation as a primitive, self-evident matter[1]. Bayes' analysis, however, is not directly concerned with causation, but with the more general case of events E_1 and E_2 which may be correlated in the sense that event E_1 may be generally held to cause event E_2, but a causal connection is by no means necessary. In this chapter, however, we are concerned with the common, if loose, view of causation which requires the causative event E_1 always to precede the caused event E_2.

There can be, however, serious difficulties in computing the probability of cause. Frequently, an honest and competent analysis will be stalled by the possible existence of causes, or agents, of which we may be completely ignorant. Such difficulties, will be obvious to any doctor or engineer who has grappled with the diagnosis of intermittent faults in a complex process; but the difficulties may be not so obvious to others. It is also saddening to see, in many cases which are raised for popular debate, the tendency for people to assign, implicitly and unconsciously, zero dogmatic prior probabilities to possibilities of which they may be totally unaware. In other cases, an 'expert' witness may assert the probability of an hypothesis concerning an empirical fact to be zero - 'impossible' - and give the impression that this is assertion is based on scientific knowledge, when it is, in truth, an

[1] See *e.g.* the article on *Causation* in the Cambridge Dictionary of Philosophy.

Version zh25a.doc at 18:36 on 13 July, 2001

assertion of a dogmatic prior based on the prejudice of an academic or professional clique. It was precisely under the duress of such dogmatic priors that Galileo was kept under house arrest for the last two years of his life[1]; and that Semmelweis[2] was put in a strait jacket and locked in a dark room, where he died. In too many cases, 'It is the theory which decides what we can observe'[3] .

Statistical frequency is often, however, a subject of great popular interest and debate, especially when related to the incidence of misfortunes which are given a high profile in the media. This readily leads us to wonder whether the probability governing a given event in one population is probably the same as the probability governing a similar event in a different population, or whether it is more probable that the governing probabilities in the two populations[4] are different. However, a difficulty with a question of this kind is that the probabilities governing the two sets may be only minutely different. In such cases, it may be extremely difficult to detect the difference, and, even if detected, a small difference may be of little significance. Hence, our real concern is to know the probability that the probabilities governing the events in the sets differ by at least a given amount.

To this end, we define:-

m_1 the number of events in a sample from the first population.

m_2 the number of events in a sample from the second population.

n_1 the size of the sample from the first population.

n_2 the size of the sample from the second population.

D the matrix of observational data, $[m_1, m_2; n_1,n_2]$

$\mathcal{P}_{R1}(e)$ the probability which determines the rate of occurrence of the event *e* in the first population.

$\mathcal{P}_{R2}(e)$ the probability which determines the rate in the second population.

To answer the above question by Bayes' reasoning, we first compute the raw measured rates in the populations, *i.e.* m_1/n_1 and m_2/n_2.

---

[1] See *Galileo* in the Cambridge Dictionary of Philosophy p 291

[2] Semmelweis had found that puerperal fever - lethal for newly-delivered mothers and their babies - could be prevented by requiring obstetricians and midwives to wash their hands in a chlorinated solution before touching their patients. See Broad, W. (1985)

[3] Einstein, in a letter to Heisenberg, cited by Broad, W. (1985) p138

[4] Such populations are generally arbitrary sets but that is usually, in such cases, valid and reasonable.

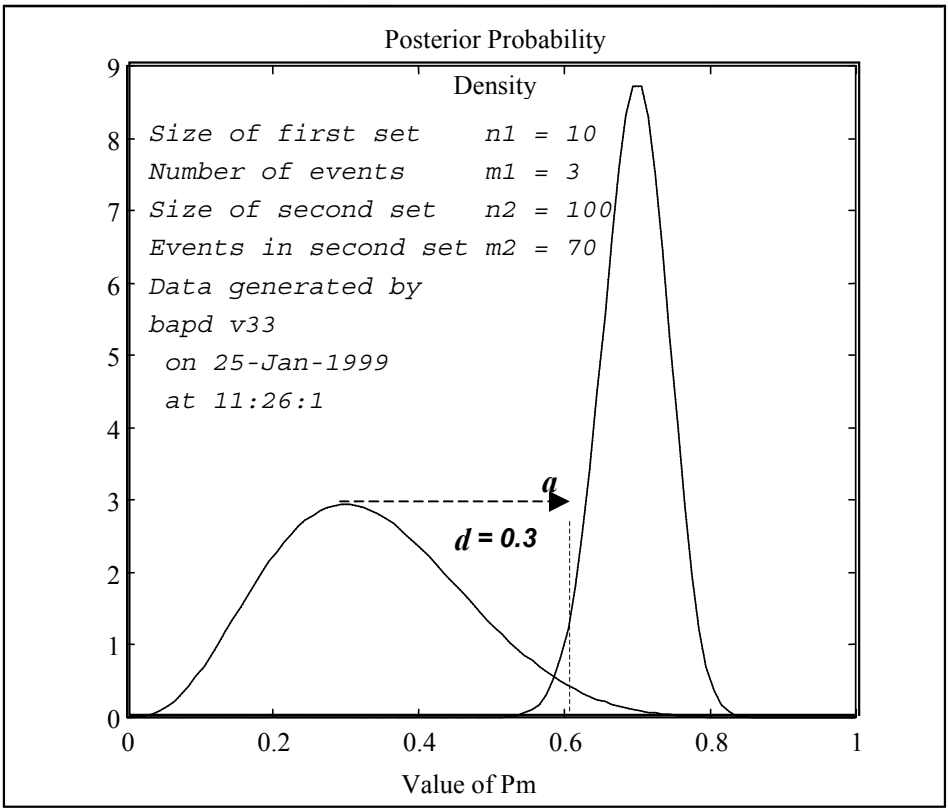


Figure 14.1

Then, taking the lower of these, we add an allowance d in order to form a dividing line between the two measured rates[1] at a point $a = m_1/n_1 + d$. Hence, the probability, on the stated assumptions, that the lower rate is less than a is given by:-

$$\mathcal{P}_H\{0 < P_{m1} < a \mid m_1, n_1, k, I_Z\} = \frac{\int_0^a x^{m1}\,(1-x)^{(n1-m1)}\,dx}{\int_0^1 x^{m1}\,(1-x)^{(n1-m1)}\,dx} \qquad (14\text{-}1a)$$

and the corresponding probability that the higher rate is greater than a is given by the converse formula, whence the joint probability, on the given assumptions, that $\mathcal{P}_{R_1}(e)$ is in the interval $0 \rightarrow a$ and that $\mathcal{P}_{R_2}(e)$ is in the interval $a \rightarrow 1$ is given by the product:-

[1] We assume that the lower rate is found in the (m_1, n_1) sample.

$$\mathcal{P}_H\{(0<P_{m1}<a)\wedge(a<P_{m2}<1)\mid D,k,I_Z\}$$

$$= \mathcal{P}_H\{0<P_{m1}<a\mid m_1,n_1,k,I_Z\}\ldots\ldots.$$

$$\times\ \mathcal{P}_H\{a<P_{m2}<1\mid m_2,n_2,k,I_Z\} \qquad (14\text{-}2)$$

An example in which this formula is used, is shown in Figure 14.1, above with details in Table 13a below:-

Size of first set	*n1 = 10*
Number of events in first set	*m1 = 3*
Size of second set	*n1 = 100*
Number of events in second set	*m1 = 70*
d = .3	
Probability in first sector	*= .97072*
Probability in second sector	*= .97922*
Probability that separation > 0.3	*= .95055*

Table 14a

However, having deduced the formula *(14-2)*, we can only imagine with trepidation, what Boole, Keynes or Fisher might have said to such an assertion, bearing in mind, for example, the fierce condemnation which, for example, Keynes heaped upon Pearson[1]. Yet the truth seems to be that such condemnations stem from the seriously false belief that the assertion of a probability is in some way absolute, and from a corresponding failure to observe and understand the relative nature of the assertion. Even so, although Keynes' assertion that such results invest any conclusion with 'far too high a degree of probability', is technically fallacious[2], he is probably better understood in terms of the superficial strength of such results being, arguably, much greater than is warranted by the evidence. It is, however, extremely difficult to see how one could ever prove or justify in depth Keynes' aspersion, for that would require a new definition of probability, the general acceptance of that definition, and an algorithm appropriate to the situation, such that a value significantly lower than that given above, should be obtained.

[1] Keynes (1921) p382. Strictly speaking, we are quoting Keynes slightly out of context. Even stronger views are however expressed by Keynes on p388 of that same volume.
[2] See Ch 7 above.

Also, there is an opposing view, for if one examines the results shown in Figure 14.1, the computed probability of separation in the example shown, *i.e. 0.95*, may seem, from an intuitive point of view, to be rather low. It is therefore interesting, and indeed it is important, in such cases, to consider the sensitivity of the results to fluctuations in the data. Thus, for example, Table 13b shows a marked sensitivity to any increase in the number of events in the first set above the level of *3* assumed above, but a remarkable insensitivity to any decrease in that number.

Size of first set	Number of events	Probability of separation > 0.3 from second set
10	*0*	*0.98*
10	*1*	*0.97*
10	*2*	*0.97*
10	*3*	*0.95*
10	*4*	*0.47*

Table 14b

If however the size of the sample can be increased from *10* to, say, *20*, then the results are distinctly clarified, as shown in Table 13c:-

Size of first set	Number of events	Probability of separation > 0.3 from second set
20	*0*	*0.999*
20	*2*	*0.997*
20	*4*	*0.996*
20	*6*	*0.976*
20	*7*	*0.841*
20	*8*	*0.475*

Table 14c

In practical situations, therefore, one should proceed with great caution in such matters and be particularly careful about drawing conclusions about causation, or its absence, from the presence or absence of statistical correlations. For, as is well known, one can often observe statistical correlations between factors where there is no possibility of a causal link. It is however less widely recognised that an absence of any observable correlation does not imply the absence of causal links - this being a matter in which the distinction between the individual and the population can be crucially important. For example, a food or activity which produces beneficial effects in a small number of people but produces adverse effects in a roughly equal number,

Version zh25a.doc at 18:36 on 13 July, 2001

may appear, statistically, to have 'no effect'. None, that is, on the statistics while on certain individuals the effects may be dramatic.

Turning now to the more general matter of diagnosing causes, an example which shows some of the difficulties, is provided by a flash of light in the night sky. We then look at some possible causes, *e.g.* lightning, gunfire, an electric train, and so on. We let E_0 denote the observed event, including all its observable parameters, and we let E_a, ... E_n denote the possible causative events[1] of which we are aware. E_ℓ denotes lightning. The corresponding hypotheses concerning the occurrence of causative events are denoted H_a ... H_ℓ, ... H_n. Applying Bayes' theorem to the situation we have, purely formally as the probability that lightning was the source of the flash:-

$$\mathcal{P}_H(H_\ell \mid E_0, k) = \frac{\mathcal{P}_R(E_\ell \mid k)}{\mathcal{P}_R(E_0 \mid k)} \times \mathcal{P}_R(E_0 \mid E_\ell, k) \qquad (14\text{-}3)$$

However, if we are pursuing the probability of lightning as a statistical frequency over all flashes seen in the sky, the probability can only be calculated if we make all sorts of arbitrary assumptions about the parameters of space, time, optical spectra and other characteristics of the flashes which constitute the population. Further, we must assume either that that we know all possible causes and can assign to them appropriate frequencies, or we must assign frequencies to the known causes and an arbitrary frequency to the totality of all unknown causes. It is all unsavoury. Further, because such causative events are not, in general, mutually exclusive, we ought, strictly speaking, to consider the possibility that lightning may not occur alone, but may also occur in conjunction with other causes of the observed flash; for example, an electric train may have crossed the points at precisely the same instant as a flash of lightning. This is tricky ground indeed.

If however we are concerned with the degree of rational belief about the cause of a particular flash, then, as a first step, we take the term $\mathcal{P}_R(E_\ell \mid k)$ which denotes the probability that we should observe lightning, given the information which is implicit in the symbol k. This information comprises data which are separate from and independent of the event E_ℓ in which we observe the flash and its associated parameters, but is relevant to the probability of cause. Such separate data could include, for instance, the weather conditions, a crackle on a radio set at the time of the flash, or a crash of thunderous noise just after the flash. The denominator in *(14-3)* is, therefore, as in *(3-51d)*, given by the summation of the numerators in *(14-3)* taken over all possible causes, or, at least, over all the causes which are assumed to be possible:-

---

[1] We are here making an implicit distinction between 'causes' and 'causative events'.

$$\mathcal{P}_R(E_0|k) = \sum_i \mathcal{P}_R(E_0|E_i,k).\mathcal{P}_R(E_i|k) \qquad (14\text{-}4)$$

There are, however, some situations in which we can avoid the difficulties indicated above, by confining the problem to pairs of well-defined, mutually exclusive and exhaustive hypotheses, such that the probability of one hypothesis is the compliment of the probability of the other. If necessary, this situation can be created by imposing a dogmatic prior to this effect. In a radar situation, for example, we may define a classification such that every 'blip' is deemed to be produced either by an aircraft, in conjunction with noise[1], which is always present, or by noise with no aircraft. Assuming that we have no independent information relating to the blip, we define the symbols:-

E_0	*a blip has been observed with parameters* $\{x_0\}$
h_1	*there is an aircraft where the blip appeared*
h_2	*there is no aircraft where the blip appeared*
k	*independent data, assumptions etc.*

whence, taking a simple and direct approach we would evaluate and compare:-

$$\mathcal{P}_H(h_1|k,E_0) = \frac{\mathcal{P}_R(h_1|k)}{\mathcal{P}_R(E_0|k)} \times \mathcal{P}_R(E_0|k,h_1) \qquad (14\text{-}5a)$$

$$\mathcal{P}_H(h_2|k,E_0) = \frac{\mathcal{P}_R(h_2|k)}{\mathcal{P}_R(E_0|k)} \times \mathcal{P}_R(E_0|k,h_2) \qquad (14\text{-}5b)$$

There are however, in this simple and direct approach, some substantial difficulties; for, having no independent information about the blip, k tells us nothing about E_0 , nor about h_1 , nor about h_2 . To avoid these difficulties, we therefore invoke the fact that k is of zero information value, as defined in Ch.11, whence:-

$$I(h_1,h_2|k) = \ln \frac{\mathcal{P}_R(h_1|k)}{\mathcal{P}_R(h_2|k)} = 0 \qquad (14\text{-}6)$$

Thus, taking the ratio of *(14-5a)* and *(14-5b)*:-

$$\frac{\mathcal{P}_H(h_1|k,E_0)}{\mathcal{P}_H(h_2|k,E_0)} = \exp\{I(h_1,h_2|k)\} . \frac{\mathcal{P}_R(E_0|k,h_1)}{\mathcal{P}_R(E_0|k,h_2)} \qquad (14\text{-}7)$$

[1] 'noise' is the name given to the unwanted 'rubbish' signals which are always present in a radar receiver. The noise arises from many different sources.

whence, because the hypotheses h_1 and h_2 are exclusive and exhaustive, it follows that the unit interval is shared by their probabilities in the same proportion as $\mathcal{P}_R(E_0 | k,h_1)$ and $\mathcal{P}_R(E_0 | k,h_2)$ which gives us:-

$$\mathcal{P}_H(h_1 | k,E_0) = \frac{\mathcal{P}_R(E_0 | k,h_1)}{\mathcal{P}_R(E_0 | k,h_1) + \mathcal{P}_R(E_0 | k,h_2)} \qquad (14\text{-}8a)$$

and

$$\mathcal{P}_H(h_2 | k,E_0) = 1 - \mathcal{P}_H(h_1 | k,E_0) \qquad (14\text{-}8b)$$

Returning to the question of radar blip classification, these results tell us that the probability that there was an aircraft at the blip, given k and E_0, is equal to *the probability that an aircraft, together with noise, would produce that blip*, divided by *the sum of that same probability and the probability that noise alone would produce that blip.* This feels like firmer ground, but we are by no means out of difficulty, for we still have to consider exactly what we mean by the italicised phrases and to avoid the ever-present pitfall of thinking in terms of frequencies and populations. On this basis, we have to interpret the phrase *the probability that an aircraft, together with noise, would produce that blip* in terms of the probability that any aircraft would produce such a blip: in principle, that is, an aircraft of a type known to us, or even of a type which is not known to us. This is slippery ground indeed, for, if we include 'stealth' aircraft which are designed to produce no identifiable reflection of the radar signal[1], there may be no detectable difference between a blip which happens to coincide with such an aircraft and a blip which is produced entirely by noise.

In such cases, our only escape is to narrow the class of aircraft to those types which are visible to radar and to correspondingly widen the complementary field to include aircraft which are not visible. Even so, many difficulties remain. For example, the phenomenon of 'scintillation' can cause any aircraft to be transiently invisible to a radar, forcing us to grasp yet another escape ladder, this time by substituting for the aircraft an idealised reflecting object of a known 'size'. On this basis, the first task becomes that of assessing *the probability that an idealised reflecting object of the given size would produce such a blip*, which is, in principle, easily computed. Conversely, the *probability that noise alone would produce that blip* is also easily computed, whence we can compute, subject to all the assumptions stated, the probability that the blip was produced by an idealised reflecting object of the given size. As we saw in Chapter 10, this information may, in fact, be useful but the essential point,

[1] We are deliberately avoiding many technical issues here which are not germane to the discussion.

however, remains that the algebra is almost trivially simple compared with the enormous difficulties of relating the algebra to the real world.

This leads to the further and important point that we should be wary of allowing such difficulties to drive us too quickly into the use of statistical expectations. The danger can be illustrated by the anecdote, dating from the closing days of the 'Cold War', of the Qantas airliner which, flying from Australia to Singapore, had been told there were no other aircraft on that route, yet found itself flying in the condensing vapour trail of another aircraft. Slowly catching up on the other aircraft, various characteristics progressively became clear until it was eventually identified as a Soviet long-range reconnaissance aircraft. On a conventional statistical basis, there would have been no expectation of finding a Soviet aircraft in that place, while, on the other hand, had the Qantas aircrew, prior to visual identification, picked up a strong radio transmission in a language which was recognisably Russian, a prior probability, directly relevant to the specific situation, would have been created. When we need to beware of rare but dangerous events, statistics and expectations based on past experience, can be totally misleading. Pearl Harbour showed this, and it might be wise for us to consider, in that light, for example, the desirability of radar surveillance against meteors and similar objects which might hit and seriously damage our planet Earth.

Of interest also in relation to probability of causation, is the technique known as a 'Bayesian Network' which has been developed in the area of artificial intelligence. This technique allows us to depict graphically the application of Bayes' theorem to problems of probable causation and to summarise and present, with remarkable directness, brevity and clarity, quite large sets of relationships which, presented in any other way, could be extremely difficult to grasp[1].

In such a network, the basic units are groups of three 'events', between which there are probabilistically causal, or similar relationships, the direction of causality being indicated by directional lines, which are also known as 'arcs' or 'edges'. This gives a 'directed graph' which, to be 'Bayesian', also has to be acyclic, in that it must not contain any recursive paths forming an un-broken sequence of arcs, *a,b,c ..... n* such that the *n'th* arc is also a previous member of the sequence. Hence a Bayesian network is a 'directed acyclic graph[2]'. When these conditions are fulfilled, a rich field of visualisation is opened for exploration.

As an example, Figure 14.6 shows a network of relationships between factors and events in the field of radar-assisted air traffic control, where,

---

[1] Ripley (1981); Heckerman (1995); Buntine (1996)

[2] Often abbreviated to *'DAG'*

initially a provisional flight plan is required. Prior to its being 'filed', such a flight plan can be envisaged as a probablistic function of circumstantial facts, such as the identity of the airport at which the flight is to originate, the destination airport, the type of aircraft, the advertised departure time and the weather forecast.

Within the network, the standard forms of Bayes' rule are used and allow us, in principle, to infer the probability of hypotheses concerning ancestral events when it is given that various descendant events have occurred. For example, in Figure 14.6, the detection of a track by a surveillance radar and the association of that track with an active flight plan can tell us something, in probability, about the weather forecast at the time the flight was planned. Conceptually, therefore, the output of a Bayesian network is a distribution of probabilities over a set of hypotheses. However, all the provisos concerning the use of prior probabilities apply as fully in the case of Bayesian networks as in any other case and it is just as important, here as elsewhere, to note the distinctions between populations and individuals. Often, however, in demanding real-life situations, there is tremendous pressure and a great temptation, to use population statistics as a stop-gap when we do not really understand or do not have enough detail to analyse an individual case. Nevertheless, we must express reservations about the assertion by Jensen[1] that, in the absence of specifically relevant prior information, a Bayesian network should be initialised by population probabilities when, in our view, it would be sounder, for the reasons given earlier, to use 'Information-zero' distributions.

The situation in which a precise diagnosis of a single cause is possible is a very special case. In the general case, there are many possible, independent factors and, in real life, there may be many interactive factors, with which a Bayesian network may be totally unable to cope. However, it is not clear that *any* kind of formal analysis is capable of unravelling situations in which factors inter-act. In practice, therefore, the problem is, often, not to identify 'the cause', nor even any distribution of probabilities over the possible causes, but to decide rationally, upon a course of action**.** This is a hugely important matter, but totally beyond the scope of this present work.

---

[1] Jensen (1995) p26 *'If nothing is known of the phenogroups of the parents, they are given a prior probability equal to the frequencies of the various phenogroups'*

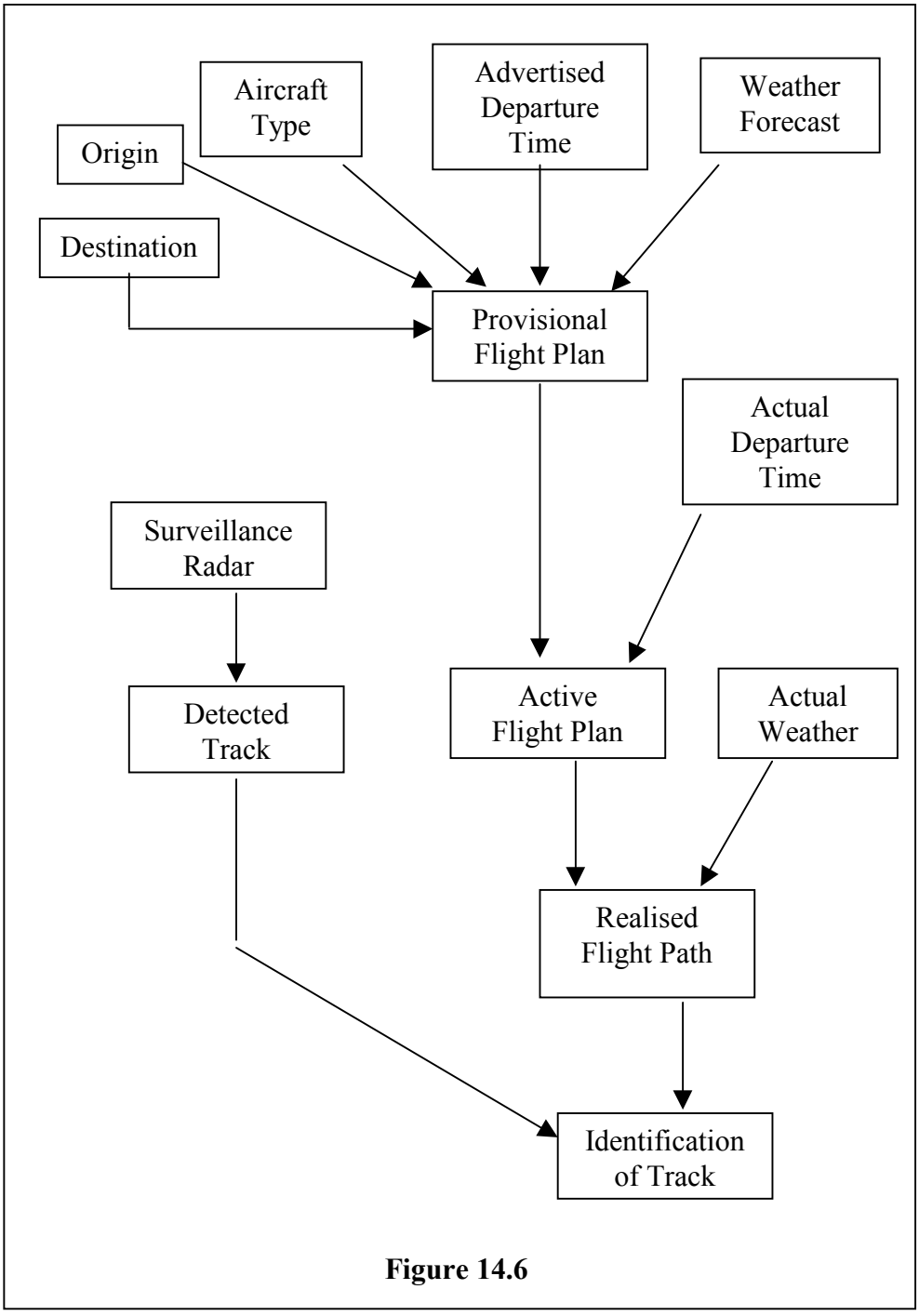


Figure 14.6

Version zh25a.doc at 18:36 on 13 July, 2001

A combination of population-oriented procedures with procedures which are oriented towards the individual, can however be of great practical importance in the design of cost-efficient diagnostic procedures, where the interpretation of evidence ought always to be directed at the individual case, but the order in which costly tests are performed, may be determined by considerations which will minimise the expected costs of such procedures when taken over a population.

In criminal investigations, for example, it may be statistically and economically sensible, taken over the populations of all crimes and all criminals, to take account of previous criminal records in deciding the order and allocation of resources to alternative lines of enquiry. This does not apply, however, in considering the probability of guilt in any specific case any more than it is rational to bias our weighing of a particular potato towards the statistical norm, or, as noted above to fly an aeroplane, not according to the readings of the instruments on this flight, but according to what they showed at a similar point on previous flights. It is however disturbing that the judicial process, at least in the English-speaking world, seems too often to be directed, not at the degree to which evidence objectively indicates the probable truth of an hypothesis concerning the guilt of an accused person, but at whether a jury can be persuaded, often by appealing to dogmatic and population priors, to return a 'guilty' verdict.

In this process, prosecutors can cite as evidence of motive, characteristics which may indeed be true of many people, but, it seems, without being required to show any evidence that an accused person was so-motivated in the commission of an alleged crime. Nor is it unknown for a prosecutor to appeal to the lack of any other known explanation as 'evidence' that the accused person actually did commit the crime. This is very disturbing to anyone who is concerned about the objective assessment of probability and its relationship to the administration of justice. It may lead us indeed to suggest that if engineers and doctors were to adopt the standards of reasoning on probability which are apparently acceptable in legal argument, the lawyers would rightly judge the engineers and doctors to be guilty of serious incompetence if not indeed gross negligence.

The problem presented, however, by the possibility of unknown causes is so fundamental that some people may feel that it invalidates the whole process of inferring the probability of causation. Yet, from the viewpoint of practical reason, it is vitally important for the making of rational decisions in uncertain situations, for example in the diagnosis of severe illness, or in emergency engineering situations, where there is no possibility of acquiring further evidence, that we are able to balance the probabilities, and, in effect,

use a form of 'identification', as we discussed in Chapter 10. Moreover, provided we make clear the assumptions we are making, the consequent assertions are, as a matter of pure logic, valid. The doubt concerns the veracity with which they reflect reality. But such doubts assail almost every rational deed we do. Rarely, if ever, can we be certain beyond any possibility of being wrong, that our knowledge or perceptions of the world around us are, in any sense, faultless. Nor indeed that they are of an accuracy sufficient to support beyond doubt our decisions. Yet, probability, as expounded by Thomas Bayes is remarkable for the extent to which it allows us to use mathematical reason in the resolution of practical issues, where we have to balance moral or ethical values against uncertainties. In such situations it is not that we should apologise for the difficulties or lack of perfection in matching the rules to the realities, but rather that we should be thankful that Bayes left us an instrument of such flexibility, simplicity, power and breadth.

Version zh25a.doc at 18:36 on 13 July, 2001

Chapter 15

In Conclusion

We have now reached a point where we must pause our exploration and offer our journal to others. We do not, in this conclusion, review all the ground we have covered, but rather we concentrate on a few salient aspects. The issue of notation is clearly not fundamental from a philosophical point of view, but is simply basic to clear thinking and self discipline. Laziness in such matters is a midwife to disaster. The calibrated ruler may, again, be not fundamental but it is crucial in releasing us from the stranglehold of arbitrary and contentious priors when we are in a measuring situation. On the basis of a calibrated measurement, we can make a plain statement of probability about, for example, the true length of a pencil without the inhibitions concerning priors which have for so long precluded such statements. This probability can be multiplied by a monetary value in order to achieve the magnitudes of a probable cost - often essential for the making of rational decisions in uncertain situations.

In the remainder of this chapter, we look first at the issue of prior probability in the individual case, where we feel that Bayes, and many after him, went astray and thus inadvertently lent support to numerous fallacies which seem to permeate both legal and political thinking. After that we consider, briefly, the relative nature of probabilities and the fundamental place in natural science which is occupied by Bayes' experiment and, further, the general matter of estimation under the conditions of observational uncertainty which permeate all science. Finally, we look again at Bayes' definition of probability and its profound implications for the integration of values over many, many aspects of human life.

Looking back at Bayes' treatment of prior probability, it is remarkable that, despite the introduction of a prior process in the Experiment, nowhere does Bayes explicitly equate knowledge of a prior probability with knowledge of the process by which the value of the probability in question is predetermined. Obviously, his mind had, at some point, moved in that direction, but nowhere does he assert, as Fisher indeed did assert, that:- *Such a*

problem is indeterminate without knowing the statistical mechanism under which different values ... come into existence[1]. Bayes is, in fact, thoroughly circumspect, for, having described the experiment, he merely suggests:- *therefore I shall take for granted that the rule given concerning the event M in Prop.9 is also the rule to be used in relation to any event concerning the probability of which nothing at all is known antecedently*[2]. Sadly, however, by connecting 'the rule to be used' in general with the process governing the position of the first ball in the experiment, Bayes unwittingly committed, if not himself, then many generations of later analysts to a concept of prior probability which engendered confusion and confrontation. Our solution to the problem of The Ruler, in Chapter 9, may therefore cause some unease to those who are deeply accustomed to thinking of prior probabilities in terms of populations, or other forms of *a priori* expectation concerning the object which is to be measured. To some, our use of the calibration may seem like an illusionist's trick to avoid a much deeper problem. Yet, this is no mere trick, for the approach based on calibration simply, but powerfully, switches our view from a mesmerised focus on illusory populations of objects to be measured - such as the population of sticks to be found on the beach, or those brought home by the dog - to the process of measurement itself[3].

Some very serious issues, stem however, from the widespread and primitive attitude to prior probability, which can, all too easily, be inferred from the *Scholium*. These issues seem to arise to a particularly acute degree in criminal proceedings, where, it seems, alleged motives, (for the existence of which there may be no direct evidence in an individual case), are often taken to imply an *a priori* probability of guilt where there may be, in fact, no direct evidence to link an accused person with a crime[4]. While it seems to us intuitively probable that, in allegations of motive and thence guilt, lawyers and most other people, were unconsciously using populations priors centuries before Bayes[5], it was only with Bayes' essay that the scientific

---

[1] Fisher (1921) p24

[2] See Chapter 5 above.

[3] Interestingly, the definition of the experimental set-up in Bayes' experiment automatically defines the calibration. Bayes' table is also, like the median, remarkably robust in this respect.

[4] The use of such reasoning to persuade juries to convict an accused person has led, in recent years, to at least two very seriously wrongful convictions in the English courts, which have been later overturned on appeal. See *e.g. The Times*, Friday 26 April 1996, p.1, *'... a victory for love and truth'*.

[5] This seems a useful topic for further research.

community became actively, if unwittingly, involved with the issue. Thus, it is unfortunate that, the concept of prior probability which Bayes introduced in the *Scholium*, seems to have been accepted without question by a large majority of the scientific and philosophical communities for some *250* years, and it is perhaps neither surprising nor reprehensible that lawyers, unchallenged by those communities, should have persisted with the primitive view. Thus, lawyers can reason that a tendency which is arguably present in people at large can be assumed to have been present and effective in the case of any person who is suspected of committing a crime.

Fortunately, it cannot be said that practitioners in the application of science, such as doctors and engineers, generally follow such reasoning in their work. On the contrary, as we have seen earlier, for them to allow *a priori* expectations to influence a process of measurement would be judged by lawyers to be incompetent and fraudulent. Unfortunately, very few practising doctors and engineers have time to reflect upon the epistemological foundations of the measurements they make when they are trying to cure a sick person or repair a collapsed bridge. In consequence, there has been little, if any challenge to the primitive fallacy that a distribution of probability within a population can be applied to an individual taken from that population.

Even more troubling is that, in everyday life if not in courts of law, the alleged existence of a motive can be taken to constitute evidence as to the commission of a crime, again, for which there may be no substantive evidence. School life abounds with petty instances of this kind, in which children are constantly at risk of false accusation. A $10 dollar bill disappears from a school desk and it is immediately assumed that 'someone has taken it' when, in fact, what really happened was 'someone' actually opened a window, the wind blew into the room and the note flew into hiding under the desk. Yet because it is assumed that children are susceptible to temptation and that a $10 bill constitutes a big temptation to some abstract child, therefore, it is reasoned, the note has been stolen. Had the note been a useless piece of scrap paper, there would be no presumption whatsoever of any wrongdoing. By similar perverted reasoning, it can also be argued, and indeed is argued, that if Jack is known to have bullied little Jimmy and, at a later date, 'someone' does something beastly to Jack, then Jack's bullying of Jimmy constitutes a motive, and therefore a prior probability of Jimmy's guilt. Hence, by extension of this corrosive reasoning, the more monstrous

Jack's bullying of Jimmy, the more it can be held to raise the probability of Jimmy's guilt. To the scientific or philosophical mind, such arguments are indeed monstrous, yet the grim truth is that, in its logical structure, Bayes' concept of the prior probability over the experiment on the flat table, embodies precisely this same perversion, albeit it would be considered better manners in scientific circles to call it merely a fallacy.

Yet, in the context of a criminal trial, under the concepts of justice which rule in western society, the true implications of the conventional, primitive view of prior probability point unwaveringly at the relevance of factors such as the genetic make up and the social background of an accused person in the determination of the probability of guilt. And, while such reasoning may be nowadays politically unacceptable in many countries, that does not show it to be fallacious and its practitioners are beating a long and hard-fought retreat. The fallacy is visible, as we noted earlier, in the fact that, logically, such reasoning would cause us to bias our weighing of a particular potato towards the statistical norm, or bias a pilot's flying of an aeroplane according to the readings of instruments on other flights. The fallacy is to assume that a probability distribution over a population is applicable to an individual selected from that population. Individuals and populations are distinct objects and, as we showed in Chapter 10, it is only in the very special case of a 'classification' dependent upon a fully determined attribute, that precise knowledge can be transferred syllogistically from a population to a member of that population. Always assuming, that is, a zero probability of mistakes in the physical process of selecting the population. Equally serious as the use of population statistics as *a priori* probabilities in criminal proceedings, is the practice, mentioned above, whereby certain British 'health authorities' set up dogmatic priors which arbitrarily exclude pre-defined groups of people from certain kinds of medical treatment, without any reference to specific evidence concerning the individual patient or doctor.

We also have to face the common illusion which believes that adding witnesses adds to the probability of obtaining the truth. A naked person 'streaks' across the Melbourne cricket ground in the course of a test match, is seen by 50,000 spectators, and then disappears into the crowd. Later, a person is arrested and all 50,000 spectators are asked whether they can identify the arrested person as the same who streaked across the pitch. All agree that this is the person: but does this really constitute massive evidence

Version zh25a.doc at 18:36 on 13 July, 2001

of guilt ? In simplistic terms, that might seem to be the case, but, if we consider people as 'classifying filters' it is arguable that we have here 50,000 very similar filters to which we apply an identical signal and get 50,000 very similar outputs. Hence, if the truth is merely that the arrested person 'looks like' the person who streaked across the cricket pitch, then the opinions of all 50,000 spectators actually tells us little more than the opinion of any single one: the testimonies are merely, in fact, highly correlated. Sadly, however, it would appear that reasoning accepted in some courts does not encompass even the most elementary insight into the concept of correlation and blindly accepts that the probabilities of joint events are given by their naïve product. The resulting, utterly unjustifiable, numbers can then be put to juries as evidence[1].

Rather similar considerations can also be applied to the findings of juries and cast grave doubt indeed upon the logical basis of majority verdicts. It may also be instructive to consider why we find it acceptable to pick twelve citizens at random and require them to weigh matters of great technical complexity in order to determine the 'guilt' or otherwise of an accused. Why then, one may ask, do we not pick 12 passengers at random when our aeroplane is ready to take off, and ask them to decide the point at which to ease the plane off the ground and into the air[2] ? At an even deeper level of disquiet, we have the grim fact that a prosecution may be brought against a person, not because the substantive evidence actually indicates to a very high degree of probability, the guilt of that individual, but because the evidence indicates a high degree of probability that a jury - for whom the dogmatic prior prejudice may be easily computed - will, by a majority, vote that person guilty.

This ghastly error, of assuming that a characteristic which is statistically typical of a population, constitutes evidence or likelihood concerning a

---

[1] The BBC news at 6pm on Friday 26 November 1999 quoted a case in which odds of 73 million to one were stated in a murder trial as the probability against cot death striking two children in one family. This evidence was apparently submitted by an expert medical witness and has since been challenged - see Wansell, (2000).
See also http://news.bbc.co.uk/hi/english/health/newsid_249000/249946.stm *"BBC News | Health | Abuse blamed for some cot deaths"*

[2] The English Court of Appeal is reported as having ruled that a juror who has relevant expert knowledge may not make that knowledge known to other jurors. ('The Times' Law supplement, 19 October 1999 p 11, 'When the juror is an expert'). It is therefore intriguing to consider just what is the epistemological basis - beyond total ignorance - upon which an English jury may lawfully base its deliberations.

specific individual is the root of much evil, of many mistakes in our society, and of little, if any good. If each person is treated truly as an individual, then, in the aggregate, we shall, to a very high degree of probability, 'get it right', for it is thus that we minimise the probability of a statistical catastrophe. While Bayes' mistake in identifying, not a previous, or independent observation, as the valid prior over the position of the first ball in his experiment, but rather the distribution within the population from which that position was drawn, may have had little practical impact on the findings of scientists, engineers and doctors, who have simply ignored it as a philosophical aberration, the impact of that same mistake, whether consciously derived from Bayes or elsewhere, on the practice of law, of politics, of education and of commerce may have been, and may still be, widespread and grim.

A different fallacy, but one which all too often occurs, is the assertion of probability as an absolute value, rather than as a value relative to and dependent upon the supporting evidence and assumptions. Assertions of probability are empirical statements and are fundamentally a matter of physics. Even in matters of quantum events, the probability that a given event will occur in a defined situation may have an absolute value, but our knowledge of that value will inevitably be to some degree uncertain, and the probability that it lies in a given range will depend, in the last resort, upon experimental information. Bayes deals with probability not just as a concept, but also as an object on which we can experiment, and thence estimate numerical values.

Deeper still, the issues raised by Bayes' essay are fundamental for quantitative science. Bayes' experiment requires no rulers, no scales, no clocks, indeed no dimensional metrics of any kind. But every empirical process involves uncertainty - a problem to which we respond by invoking the concept of probability. To handle uncertainty and probability when measuring dimensional parameters, we have to, first, calibrate the rulers, clocks, and other meters by means of a Bayes trial, and, second, use Bayes' theorem in order to combine the calibrations and measurements. Hence, the fact that the measurement of probability, in a Bayes trial, requires only the counting of binary events, suggests that probability, though a late arrival on the scene of scientific consciousness, is a concept rather more primitive than dimensional concepts such as length, time and mass. Hence, we might reasonably expect the principles employed in the estimation of parametric

values, especially the values of natural constants, to be regarded as a fairly important branch of natural science. Yet, Bayes has been largely ignored by the natural science community and has been both abused and misrepresented by others for nearly 250 years, while yet other scientists regard the issues raised by Bayes' essay as rather trivial matters of technology to be left in the hands of statisticians.

This is a sad state of affairs, for, in the course of the twentieth century, many statisticians have moved from a rather harmless *ad hoc* pragmatism in their approach to estimation, to beliefs in 'optimality' and other dogmatic virtues. As a result, estimation has become a mechanical process in which data are fed into pre-defined computational engines, which can be shown, by analysis, to possess dogmatically correct properties[1]. Hence, it is not perhaps surprising that natural scientists should have turned their backs on the theory of estimation. Yet there is no law, neither of man nor of nature, which requires us to confine the design of estimators to those for which dogmatically correct properties can be analytically proven - usually in relation to yet further dogmatic assumptions. Such estimators are an infinitesimally small proportion of those which are conceivable and there is here, a tract of unexplored territory, in which, using the power of modern computing techniques, new approaches to estimation can be evaluated by the kinds of empirical criteria - including trials in the style of Bayes' experiment - which are the true norm in natural science. Only those who have hung their hats inexorably on the dogmatic pegs of 'correct thinking' need fear the result.

We turn, finally, to the wider implications of Bayes' definition of probability. Mathematics, money, love and morality are rare bedfellows, yet, if we follow to its logical conclusion the philosophy of probability expounded in Bayes' essay, we find that, despite the gap which, in our culture, divides morality and affective values from mathematics, the chasm can perhaps be bridged. Yet, although such a bridge could bring relief and rationality where today we experience only pain and confusion when we confront situations which combine both deep feelings and great uncertainty, there is, sadly, a real danger that intellectual vandals will be the first across the bridge in the pursuit of destructive agenda. At a time when values of scholarship, compassion, and the pursuit of excellence are globally under

[1] Always provided that the data also correspond to the dogmatic models: the difficulty of proving which, was admirably demonstrated by Jeffreys in his 1934 paper to the Royal Society.

attack, we can scarcely contemplate with equanimity the prospect of placing in the hands of vandals a conceptual instrument which might further their destructive intent. So why should we proceed with this exploration if we believe that it may yield an instrument of evil ? The answer is, simply, that our own values, our dogmatic faith that truth and goodness are ultimately one, requires us to do so.

For, there are grounds to believe that issues of value, even of affective value, may be more prevalent in matters of scientific explanation than is commonly perceived. There is, for example, the issue of 'simplicity', where, following Occam, and before him Empedocles[1], explanations which are simple and economic in the assumptions they invoke, are much preferred. Yet the evaluation of simplicity may be a highly subjective matter. Although one can conceive of a machine which might automatically yield a measure of simplicity, it is arguable that the measure must itself always be relative to some standard which has been arbitrarily decreed.

Finally, there are, in human life, all too many agonisingly serious problems where it is our moral or professional duty to balance probabilities and costs which are not obviously commensurable. Yet the fact that we often are able to make decisions in appallingly difficult cases may be taken to imply that, from a mathematical or logical point of view, we have found it within ourselves to assign the necessary costs and make the corresponding decisions. Yet such is the gap that separates the cultural fields of mathematics and natural science, not to mention accountancy, from the fields of morality and emotive values, that few will not experience qualms, possibly revulsion, at the implications of ordering and valuation that are implicit in Bayes' philosophy. For, there can be no doubt that the Bayes' essay contains the seeds of a vibrantly fertile philosophy which, in principle and in practice, offers us conceptual instruments by which we can, if we so desire, rationally resolve some of the most difficult problems in life. But, for this possibility to be realised, we have first to build, in our culture and in our hearts, a bridge across this great divide. We have to deal with the taboo which makes us feel sick at the thought of relationships between morality, aesthetics and mathematics on anything deeper than trivial utilitarian issues. Of putting a dollar sign beside a human life. Where we are directly and personally involved, this may be totally, and perhaps rightly, impossible; yet there are many people in positions of authority and responsibility from whom, every

---

[1] Cited by Aristotle, 'Physics' section 188a, lines 17-18.
βελτιον τε ελαττω και πεπερασμενα λαβειν, οπερ ποιει Εμπεδοκλης .
See Ross (1936) p 487

day, such judgements are demanded. However it may be hidden from conscious thought, it is hard to believe that the 'value expected' is other than the criterion which they must, implicitly, employ.

Yet, despite the fact that probability is a dangerous solvent, too easily abused into producing clear solutions from turbid thinking, it is fortunate that probability belongs to all: we are free to discuss it, and carry out our own experiments, at the races, on a poker machine in a neighbourhood club, or in the hallowed precincts of an ancient university. In a world of science that is not entirely free from class distinctions, cliques, snobbery and hypocrisy, the widespread human feel for probability is an amazingly common and unifying factor. It is one of the great blessings of Thomas Bayes' essay that he defines probability in terms which are fully within the experience and grasp of most people.

It is however surprising perhaps, that such a perspective springs from the approach to probability adopted by an eighteenth century Nonconformist cleric and mathematician, who posed a problem in 'The Doctrine of Chances' which stems from a simple counting of events in repeated trials. But there are other probabilities, which apply not to repeatable cases, but to unique and un-repeatable events, where we cannot measure by counting instances. Bayes faced this problem very squarely with his definition, not of 'probability in itself', but of a rule by which we can measure a probability by means of a ratio of values, and which can be applied quite generally to the truth of hypotheses, to unique events, and to repeatable events. Whether Bayes conceived this measure for himself, or borrowed it from another, perhaps Huygens, we do not know; but we do know, to a fair degree of certainty, that it was Bayes who showed, by applying this measure to the fundamental problem of quantitative inference, that a rational solution could be found.

Implicit in that solution is a rule for the making of decisions in questions which involve the balancing of values and probabilities. Thus, having shown how we can make a perfectly rational connection between money as a measure of value and quantitative science, Bayes had implicitly provided a means for bridging the gap that has long separated the measures of mathematics and natural science from the measures of human values. Therefore, because Bayes' approach to probability has suffered much censure and misrepresentation, it is a matter of quite wide importance to have expounded his argument as clearly as we are able and to have decided for ourselves, not perhaps whether Bayes was simply right or wrong, but rather the extent to which we can accept his argument as valid, useful, enlightening.

On a wider view, the language of human discourse concerning matters of probability is rich. It is, or it ought to be, salutary for the analyst to consider that the wealth of concepts betokened by the language, is generally much greater than any monograph can encompass. The language of ordinary folk, moreover, continues to evolve, and to show probability in ever new aspects. We have touched upon just a few of those aspects, but there are many other aspects which we have not even mentioned. We have looked into certain situations in which we can test the chances that an hypothesis concerning the cause of certain events is true, but there are many other situations where we may form hypotheses concerning causation, and at which we have cast scarcely a glance. There are, for instance, whole areas of the law where questions concerning the probability of guilt and responsibility arise, areas which were of great concern to philosophers of earlier times, such as Leibniz, but of which one hears little today. Arguably, however, Bayes gave us, a wonderful new instrument by showing us the way towards maximising the expected utility of decisions across an enormous range of political, technical and commercial activities. The penalty is that, when those in authority fail to understand the distinction between the population and its members, their decisions, in terms of individual human lives, are often disastrous, providing endless motivation for the media to undermine authority of any form, fanning the flames of irrational blame and Luddite retribution.

This book is simply, therefore, the report of our expedition along the paths we chose to take. Others would no doubt have taken different paths. As in many human activities, we had to balance prospects and penalties against the probabilities, as they appeared to us at the time. In exploring these paths, stretching nearly 250 years from Bayes' essay to the present day, we have inevitably had to move with haste in many places where there was much over which we could well have lingered and pondered in depth. Of the wide field over which probability ranges, we have addressed just a small portion: great areas still present wonderful, and vitally important opportunities for exploration. The fields of discourse and enquiry which are permeated by the concept of probability are wide: too wide, fortunately, to be captured by the sort of linguistic piracy which removes words from everyday use and constrains them to meanings which can sometimes be almost diametrically opposed to their meanings in ordinary life. We use these metaphors deliberately: for despite occasional efforts by some, to confine probability to a narrow and rather sterile axiomatic strip, probability remains defiantly at large, a wonderful creature of the human imagination, yet pos-

sessed of properties which seem amazingly founded in external realities over which our imagination has no control[1].

In choosing his definition of probability, Thomas Bayes tied his treatment of probability to the human experience of life and made monetary value fundamental. While such an equation might seem remarkably foreign to the world of science, there can be little doubt that it is abundantly meaningful to people at large. And if probability is indeed as fundamental in our natural philosophy as time, length, mass - perhaps even more fundamental - and if ratios of values - indeed, monetary values - are fundamental to the measurement of probability, then perhaps we have here a point of view which can, in time, integrate our sense of economic, scientific, aesthetic and ethical values in a world where uncertainty, probability and a compensating creativity are openly accepted as fundamental elements.

[1] Paraphrasing slightly: *Science is a magnificent adventure of the human spirit. The aim .... is not to achieve certainty but to invent better and better theories ..... more and more powerful searchlights.* (Popper (1973) p.361).

Appendix A

Fisher's Mice

In Chapter 2 of *'Statistical Methods and Scientific Inference*[1]*'*, Fisher develops a criticism of Bayes, based on an imaginary experiment in which mice, of different genetic types, mate and produce offspring. Unfortunately, Fisher's specification of the experiment is not clear, and we found it necessary to expand it at some length. Fisher is also, (like Bayes before him), unhelpful in his choice of colours, for he specifies the mice as being black or brown, leading to the confusing abbreviations *BB* and *Bb*. We therefore postulate an equivalent experiment with imaginary red (*R*) and grey (*g*) mice. The experiment is assumed to be governed by the rules of Mendelian genetics[2] in which each mouse carries a pair of genes which together determine the colour of its fur. The red gene is dominant and is denoted by the upper-case *'R'*; the grey gene is recessive, and is denoted by the lower-case *'g'*. Hence we have the rules given in Appendix B which determine, according to the genes of the parents, the probable, or expected, frequencies governing the distribution of colour and genes in their offspring. It is however extremely important to phrase this relationship correctly, for, while the genes of the parents determine the probability distribution over their offspring, considered as a population, the joint genetic constitution of the parents does not, in general, constitute evidence as to the actual genetic constitution of any specific offspring[3] and, given the possibilities of mutation and genetic modification *etc.*, this is more than a philosophic nicety. In a normal case, however, the genes of a descendant do provide evidence concerning the genes of its ancestors.

The experiment takes the form of a test mating between a red female and a grey male. The outcome is a litter of seven red mice[4], and Fisher's aim

---

[1] Fisher (1956) pp 18 *et seq.*

[2] The rules are given in Appendix B, below

[3] Special cases occur, *e.g.* where both parents are *RR* or *gg*, but even here it is debatable whether the genes of the parents actually constitute *evidence* about the genes of the offspring.

[4] Additional conditions, necessary for the validity of the following argument, are (i) that each offspring shall be the product of independent fertilisation of a separate ovum, (ii) that

is to compare the probabilities of two competing hypotheses concerning the genetic constitution of the red female *viz:-*

h_{RR} *is the hypothesis that the red female is of genetic type RR.*
h_{Rg} *is the hypothesis that the red female is of genetic type Rg.*

Fisher also asks that we consider, in addition, two possibilities concerning the parents of the red female *viz*:-

- in the first case, it is known that the red female is herself the offspring of a mating between two *Rg* mice, thus giving computable prior probabilities[1] for the genetic types of red females from such parents; *i.e.* they can be *RR* or *Rg*. The expected frequencies are given below.

- in the second case, the parentage of the red female is unknown and the expected frequencies of the various genetic types in their red female offspring are therefore also unknown.

In the first case, Fisher's argument is that we can evaluate the competing hypotheses by Bayes' theorem, *(3-51b)*, *viz:-*

$$\mathcal{P}_H(h_{RR} \mid {}^7R_7) = \frac{\mathcal{P}_R(RR \mid Rg \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})} \quad (A\text{-}8)$$

$$\mathcal{P}_H(h_{Rg} \mid {}^7R_7) = \frac{\mathcal{P}_R(Rg \mid Rg \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})} \quad (A\text{-}9)$$

where the bold type $\boldsymbol{Rg} \wedge \boldsymbol{Rg}$ indicates the genes of the maternal grandparents and the further symbols signify:-

${}^7R_7$ *the event that a litter of seven red mice* (*and no grey mice*) *are born from the mating.*

$\mathcal{P}_R(RR \mid Rg \wedge Rg)$
The probability that an offspring selected at random from a mating of two Rg parents will be of type RR.
By Rule 7a, $\mathcal{P}_R(RR \mid Rg \wedge Rg) = 0.25$.

---

the probability of successful fertilisation and survival to birth is independent of the genetic type in question. There may well be numerous further conditions of this kind, which could be identified by experts in this field. None, however, are mentioned by Fisher.

[1] These 'probabilities' are, of course, the expected frequencies in the population of descendants and do not apply to any specific individual.

$\mathcal{P}_R\,(\,Rg \mid Rg \wedge Rg)$

The probability that an offspring selected at random from a mating of two Rg parents will be of type Rg.

By Rule 7b, $\mathcal{P}_R(\,Rg \mid Rg \wedge Rg) = 0.5$

$\mathcal{P}_R\,(\,{}^7R_7 \mid RR \wedge gg)$

the probability that, in a litter of seven offspring, produced by the mating of an RR parent, with a grey mouse, all will be red. In this case, by Rule 6a above, $\mathcal{P}_R(\,{}^7R_7 \mid RR \wedge gg) = 1.0$

$\mathcal{P}_R\,(\,{}^7R_7 \mid Rg \wedge gg\,)$

the probability that, in a litter of seven offspring, produced by the mating of an Rg parent, with a grey mouse, all will be red. As the probability of any such single offspring, selected at random, being red is, by Rule 8a above, 0.5, the probability that all seven will be red is given by Bayes' Proposition 6 as $\mathcal{P}_R(\,{}^7R_7 \mid Rg \wedge gg) = (0.5)^7$

$\mathcal{P}_R(\,{}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})$

the probability that, following a mating of a red female which is known only to be the offspring of two Rg grandparents, (*indicated by bold type*)*, with a grey mouse, every offspring in a resultant litter of seven, will be red.*

$\mathcal{P}_H(\,h_{RR} \mid {}^7R_7)$

the probability that the red female was of type RR in a trial in which, in a litter of seven offspring, all were red.

$\mathcal{P}_H(\,h_{Rg} \mid {}^7R_7)$

the probability that the red female was of type Rg in a trial in which, in a litter of seven offspring, all were red

Thus, as the other values are as given above, we require only the further value of $\mathcal{P}_R(\,{}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})$ in order to evaluate the required expressions. In cases of this type, however, the hypotheses h_{RR} and h_{Rg} are mutually exclusive and exhaustive alternatives, (*i.e.* a red mouse must be either *RR* or *Rg*):-

$$\mathcal{P}_H(\,h_{RR} \mid {}^7R_7) + \mathcal{P}(\,h_{Rg} \mid {}^7R_7) = 1 \qquad (A\text{-}10)$$

thus

$$\frac{\mathcal{P}_R(RR \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg}) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})}$$

$$+ \quad \frac{\mathcal{P}_R(Rg \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg}) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})} \quad = \quad 1 \qquad (A\text{-}11)$$

and thus

$$\begin{aligned} &\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg}) \quad = \\ &\mathcal{P}_R(RR \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg}) \times \quad \mathcal{P}_R({}^7R_7 \mid RR \wedge gg) \\ &+ \ \mathcal{P}_R(Rg \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg}) \times \quad \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg) \qquad (A\text{-}12) \end{aligned}$$

Substituting then the numerical values for $\mathcal{P}_R(RR \mid Rg \wedge Rg)$ *etc.* as given above, we have:-

$$\begin{aligned} \mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg}) \quad &= \quad (0.25)^*(1.0) + (0.5)^*(0.5)^7 \\ &= \quad 0.254 \qquad (A\text{-}12a) \end{aligned}$$

Hence, by Fisher's argument

$$\mathcal{P}_H(h_{RR} \mid {}^7R_7) \quad = \quad \frac{\mathcal{P}_R(RR \mid Rg \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})}$$

$$= \quad \frac{(0.25).(1.0)}{(0.254)} \quad = \quad \boldsymbol{0.9846} \qquad (A\text{-}13)$$

Also

$$\mathcal{P}_H(h_{Rg} \mid {}^7R_7) \quad = \quad \frac{\mathcal{P}_R(Rg \mid Rg \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg} \wedge \boldsymbol{Rg})}$$

$$= \quad \frac{(0.5) \times (0.5)^7}{(0.254)} \quad = \quad 0.0154 \qquad (A\text{-}14)$$

apparently showing the much greater probability that a red female in such a situation was of type *RR*.

In our view, however, the above results are fallacious, for, while they purport to show the probabilities of hypotheses, they are in fact expectations taken over populations and should be correctly expressed as:-

$$\mathcal{P}_R(RR \mid {}^7R_7, \boldsymbol{Rg} \wedge \boldsymbol{Rg}) \quad = \quad \boldsymbol{0.9846}$$

and

$$\mathcal{P}_R(Rg \mid {}^7R_7, \mathbf{Rg} \wedge \mathbf{Rg}) \quad = \quad 0.0154$$

which is to say that, if a large number of red females, all of whom are the offspring of $Rg \wedge Rg$ parents, produce litters of seven red mice, and have no other offspring, then we would expect *98.46 %* of those females to be *RR* and *1.54%* to be *Rg*. That is, where a population prior is invoked, as by Fisher, the resultant value, in relation to an hypothesis, is more akin to a likelihood than to a probability, *i.e.* it is a very general assertion of a probability relating to the class of all red females in a defined situation. The invoked knowledge of the grandparents provides no evidence whatsoever as to the genetic make-up of any specific descendant, albeit a descendant does provide evidence as to the genetic make-up of its ancestors.

Fisher then turns his attention to the case when we do not know the genes of the maternal grandparents. He argues:- *If, therefore, the experimenter knows that the animal under test is the offspring of two heterozygotes*[1]*, ... cogent knowledge 'a priori' would have been available, and the method of Bayes could properly be applied. But if* (*that*) *knowledge were lacking, no experimenter would feel he had a warrant for arguing as if he knew that of which in fact he was ignorant, and for lack of adequate data Bayes' method of reasoning would be inapplicable to his problem*[2].

Superficially, these assertions are plausible and seductive, yet, when we look at them in detail, and compare them with what Bayes actually wrote, we find them misleading and aspersive. First we have the implication that *the method of Bayes* is only applicable when prior knowledge is available: yet anyone who has read Bayes' essay, (and Fisher clearly had read the essay, for he quotes it *verbatim*), must know that Bayes' objective is firmly defined as being to address the case in which *a priori* knowledge is not available. Fisher then concludes his remark with an intensified assertion, not merely against 'Bayes' method' but against Bayes' 'method of reasoning', which, we are told, is inapplicable to this problem. In fact, however, we can determine, exactly as above, the so-called 'probabilities' for each possible gene-set in the parents of the red female.

The analysis is tedious[3]. The first step is to construct a table to show the expected frequency, according to each possible combination of gene-sets in the parents of the red female, with which a red descendant of her mating with a grey male will be *RR* or *Rg*. Five combinations are possible, as shown in Table A1, where the top line shows the genes of the female's

---

[1]In our example, mice of type *Rg* are *heterozygotes*.

[2]Fisher (1956) pp 19-20.

[3] Details of the calculations are given in Appendices B and C below.

parents, the left hand column designates the genes of the offspring and the cells show the expected frequencies:-

	$RR \wedge RR$	$RR \wedge Rg$	$RR \wedge gg$	$Rg \wedge Rg$	$Rg \wedge gg$
$\mathcal{P}_R(RR \mid \ldots)$	*1.0*	*0.5*	*0.0*	*0.25*	*0.0*
$\mathcal{P}_R(Rg \mid \ldots)$	*0.0*	*0.5*	*1.0*	*0.5*	*0.5*

Table A1

We then use again *(A-8)* and *(A-9)* in order to compute, for each possible parentage of the red female, the expected frequency with which, following a mating with a grey male, in a litter of seven, all will be red. To this end, we substitute the values from Table A1, together with the values of $\mathcal{P}_R({}^7R_7 \mid RR \wedge gg) = 1.0$ and $\mathcal{P}_R({}^7R_7 \mid Rg \wedge gg) = (0.5)^7$ as derived previously, to give Table A2:-

Parents of red female ⇓	$\mathcal{P}_R(RR \mid \ldots)$	$\mathcal{P}_R(Rg \mid \ldots)$	$\mathcal{P}_R({}^7R_7 \mid \ldots)$
$RR \wedge RR$	*1.0*	*0.0*	***1.000***
$RR \wedge Rg$	*0.5*	*0.5*	***0.5039***
$RR \wedge gg$	*0.0*	*1.0*	***0.0078***
$Rg \wedge Rg$	*0.25*	*0.5*	***0.2539***
$Rg \wedge gg$	*0.0*	*0.5*	***0.0039***

Expected Frequency of ${}^7R_7$ Litter
According to Parentage of Red Female
Table A2
(Details of the calculations are given in Appendix C)

If we now use Fisher's reasoning to compute values of the competing hypotheses, h_{RR} and h_{Rg}, in each of the above cases, we get the results shown in Table A3. This is achieved by substituting the above values of $\mathcal{P}_R({}^7R_7 \mid \ldots)$ in expressions corresponding to *(A-8)* and *(A-9)* above, *i.e.*

$$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7)$$

$$= \frac{\mathcal{P}_R(RR \mid \ldots) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \ldots)} \qquad \textit{(A-15)}$$

$$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7)$$

$$= \frac{\mathcal{P}_R(Rg \mid \ldots) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \ldots)} \qquad \textit{(A-16)}$$

Mother's parents ⇓ ⇒		***RR***	***Rg***	***gg***
RR	$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7)$	***1.0***	***0.9921***	*0.0*
	$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7)$	*0.0*	*0.0078*	***1.0***
Rg	$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7)$	***0.9921***	***0.9844***	*0.0*
	$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7)$	*0.0078*	*0.0155*	***1.0***
gg	$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7)$	*0.0*	*0.0*	
	$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7)$	***1.0***	***1.0***	

Pseudo-Probabilities $\mathcal{P}_{H^*}(\,.\,)$ of hypotheses h_{RR} and h_{Rg} according to genetic types of mother's parents

Details of calculations are shown in Appendix D.

Table A3

In Table A3, the upper line in each cell shows the pseudo $\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7)$ for the case where the genes of the maternal grandparents are as shown at the heads of the corresponding rows and columns. The lower line in each cell shows the pseudo $\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7)$ in each corresponding case.

Table A3 however shows just how seriously misleading it can be to allow a prior process to influence our assessment of the probability in an individual case. For, if no independent evidence is available, the values of *1.0, 0.9921,* and *0.9844* are so close that, quite apart from our fundamental objections, the evidence of this test could provide no grounds for deciding between the competing hypotheses. Furthermore, and contrary to popular myth[1], there is here no attenuation of the importance of the prior with increasing size of the litter, so long as all the offspring are red.

A single grey offspring in the litter is, however, sufficient to eliminate the hypothesis h_{RR} ; for a red parent whose genes are *RR*, cannot produce a grey offspring. Thus, under the Mendelian rules, a grey offspring in the litter would also eliminate the possibility that the genes of the maternal grandparents could be $\boldsymbol{RR \wedge RR}$**.** These are, in our view, highly significant results. We conclude, therefore, that Fisher is simply wrong in his claim that this example demonstrates fatal weakness in Bayes' conclusion and method of reasoning. Quite to the contrary, the example leaves Bayes unscathed.

[1] *cf* the remarks of Laplace *et al* quoted in Molina (1931).

Appendix B

Mendelian Rules

The experiment is assumed to be governed by the rules of Mendelian genetics in which each mouse carries a pair of genes which determine the colour of its fur. The red gene is dominant and is denoted by the upper-case *'R'*; the grey factor is recessive, and is denoted by the lower-case *'g'*. Hence we have the following rules:-

(*1*) *If* *a given mouse is red*
And *its parentage is unknown*
Then *its genes can be RR or Rg*

(*2*) *If* *a given mouse is grey*
Then *its genes are gg.*

(*3*) *If* *RR mates with RR,*
Then *a. All offspring are red*
b. All carry RR genes

(*4*) *If* *RR mates with Rg*
Then *a. All offspring are red*
b. There is a probability of 0.5 that an offspring selected at random will have RR genes.
c There is a probability of 0.5 that an offspring selected at random will have Rg genes

(*5*) *If* *RR mates with gg*
Then *a. All offspring are red*
b. All carry Rg genes

(*6*) *By* (*3*), (*4*), (*5*)
If *RR mates with RR or Rg or gg*
Then
a. All offspring are red.

Version zh25a.doc at 18:36 on 13 July, 2001

(*7*) *If* *Rg mates with Rg*
Then
a. There is a probability of 0.25 that an offspring selected at random will be red and will have RR genes.
b. There is a probability of 0.5 that an offspring selected at random will be red and will have Rg genes
c. There is a probability of 0.25 that an offspring selected at random will be grey and will therefore have gg genes.

(*8*) *If* *Rg mates with gg*
Then
a. There is a probability of 0.5 that an offspring selected at random will be red and will have Rg genes.
b. There is a probability of 0.5 that an offspring selected at random will be grey and will have gg genes.

(*9*) *If* *gg mates with gg*
Then
a. There is a probability of 1.0 that an offspring selected at random will be grey and will have gg genes.

Appendix C

Expressions and Calculations Used to Produce the Results Shown in Table A2

The probability - *i.e.* in the sense of 'expected frequency' - that, in litters of seven mice resulting from matings between Red females and grey males, all will be Red, is computed according to the genes of the maternal grandparents. The genes of the grandparents assumed in each case are shown in bold type.

$$
\begin{aligned}
&\mathcal{P}_R({}^7R_7 \mid \boldsymbol{RR \wedge RR}) \\
&= \mathcal{P}_R(RR \mid RR \wedge RR) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg) \\
&+ \mathcal{P}_R(Rg \mid RR \wedge RR) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg) \\
&= 1.0 \times 1.0 + 0.0 \times (0.5)^7 = 1.0
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{P}_R({}^7R_7 \mid \boldsymbol{RR \wedge Rg}) \\
&= \mathcal{P}_R(RR \mid RR \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg) \\
&+ \mathcal{P}_R(Rg \mid RR \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg) \\
&= 0.5 \times 1.0 + 0.5 \times (0.5)^7 = 0.5039
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{P}_R({}^7R_7 \mid \boldsymbol{RR \wedge gg}) \\
&= \mathcal{P}_R(RR \mid RR \wedge gg) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg) \\
&+ \mathcal{P}_R(Rg \mid RR \wedge gg) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg) \\
&= 0.0 \times 1.0 + 1.0 \times (0.5)^7 = (0.5)^7
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg \wedge Rg}) \\
&= \mathcal{P}_R(RR \mid Rg \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg) \\
&+ \mathcal{P}_R(Rg \mid Rg \wedge Rg) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg) \\
&= 0.25 \times 1.0 + 0.5 \times (0.5)^7 = 0.2539
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{P}_R({}^7R_7 \mid \boldsymbol{Rg \wedge gg}) \\
&= \mathcal{P}_R(RR \mid Rg \wedge gg) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg) \\
&+ \mathcal{P}_R(Rg \mid Rg \wedge gg) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg) \\
&= 0.0 \times 1.0 + 0.5 \times (0.5)^7 = (0.5)^8
\end{aligned}
$$

Appendix D

Calculations of pseudo-probabilities $\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7)$ and $\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7)$ according to genetic make up of maternal grandparents

If maternal grandparents are RR ∧ RR

$$\mathcal{P}_R(RR \mid \mathbf{RR} \wedge \mathbf{RR}) = 1.0$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{RR}) = 1.0$$

$$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7) = \frac{\mathcal{P}_R(RR \mid \mathbf{RR} \wedge \mathbf{RR}) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{RR})}$$

$$= 1.0 \times 1.0 / 1.0 = \mathbf{1.0}$$

If maternal grandparents are RR ∧ Rg

$$\mathcal{P}_R(RR \mid \mathbf{RR} \wedge \mathbf{Rg}) = 0.5$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{Rg}) = 0.504$$

$$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7) = \frac{\mathcal{P}_R(RR \mid \mathbf{RR} \wedge \mathbf{Rg}) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{Rg})}$$

$$= 0.5 \times 1.0 / 0.504 = \mathbf{0.9921}$$

If maternal grandparents are RR ∧ gg

$$\mathcal{P}_R(RR \mid \mathbf{RR} \wedge \mathbf{gg}) = 0.0$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{gg}) = 0.0078$$

$$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7) = \frac{\mathcal{P}_R(RR \mid \mathbf{RR} \wedge \mathbf{gg}) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{gg})}$$

$$= 0.0 \times 1.0 / (0.5)^7 = \mathbf{0.0}$$

If maternal grandparents are Rg ∧ Rg

$$\mathcal{P}_R(RR \mid \mathbf{Rg} \wedge \mathbf{Rg}) = 0.25$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{Rg}) = 0.254$$

$$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7) = \frac{\mathcal{P}_R(RR \mid \mathbf{Rg} \wedge \mathbf{Rg}) \times \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{Rg})}$$

$$= 0.25 \times 1.0 \;/\; 0.254 = \mathbf{0.9843}$$

If maternal grandparents are Rg ∧ gg

$$\mathcal{P}_R(RR \mid \mathbf{Rg} \wedge \mathbf{gg}) = 0.000$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{gg}) = 0.004$$

$$\mathcal{P}_{H^*}(h_{RR} \mid {}^7R_7) = \frac{\mathcal{P}_R(RR \mid \mathbf{Rg} \wedge \mathbf{gg}) \quad \mathcal{P}_R({}^7R_7 \mid RR \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{gg})}$$

$$= 0.0 \times 1.0 \;/\; 0.004 = \mathbf{0.0}$$

We next evaluate the probability of the hypothesis that the red mother is of type *Rg* under each possible assumption concerning the genes of her parents:-

If maternal grandparents are RR ∧ RR

$$\mathcal{P}_R(Rg \mid \mathbf{RR} \wedge \mathbf{RR}) = 0.0$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{RR}) = 1.0$$

$$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7) = \frac{\mathcal{P}_R(Rg \mid \mathbf{RR} \wedge \mathbf{RR}) \times \quad \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{RR})}$$

$$= 0.0 \times (0.5)^7 / 1.0 = \mathbf{0.0}$$

If maternal grandparents are RR ∧ Rg

$$\mathcal{P}_R(Rg \mid \mathbf{RR} \wedge \mathbf{Rg}) = 0.5$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{Rg}) = 0.504$$

$$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7) = \frac{\mathcal{P}_R(Rg \mid \mathbf{RR} \wedge \mathbf{Rg}) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{Rg})}$$

$$= 0.5 \times (0.5)^7 / 0.5039 = \mathbf{0.0078}$$

If maternal grandparents are RR ∧ gg

$$\mathcal{P}_R(Rg \mid \mathbf{RR} \wedge \mathbf{gg}) = 1.0$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{gg}) = 0.008$$

$$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7) = \frac{\mathcal{P}_R(Rg \mid \mathbf{RR} \wedge \mathbf{gg}) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{RR} \wedge \mathbf{gg})}$$

$$= 1.0 \times (0.5)^7 / (0.5)^7 = \mathbf{1.0}$$

If maternal grandparents are Rg ∧ Rg

$$\mathcal{P}_R(Rg \mid \mathbf{Rg} \wedge \mathbf{Rg}) = 0.5$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{Rg}) = 0.254$$

$$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7) = \frac{\mathcal{P}_R(Rg \mid \mathbf{Rg} \wedge \mathbf{Rg}) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{Rg})}$$

$$= 0.5 \times (0.5)^7 / 0.254 = \mathbf{0.0154}$$

If maternal grandparents are Rg ∧ gg

$$\mathcal{P}_R(Rg \mid \mathbf{Rg} \wedge \mathbf{gg}) = 0.5$$

$$\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{gg}) = (0.5)^8$$

$$\mathcal{P}_{H^*}(h_{Rg} \mid {}^7R_7) = \frac{\mathcal{P}_R(Rg \mid \mathbf{Rg} \wedge \mathbf{gg}) \times \mathcal{P}_R({}^7R_7 \mid Rg \wedge gg)}{\mathcal{P}_R({}^7R_7 \mid \mathbf{Rg} \wedge \mathbf{gg})}$$

$$= 0.5 \times (0.5)^7 / (0.5)^8 = \mathbf{1.0}$$

Appendix E

The Gaussian Ratio

Given an expression of the form:-

$$Y = \frac{G(a, \sigma_a, x) . G(b, \sigma_b, x)}{\int_{-\infty}^{\infty} G(a, \sigma_a, x) . G(b, \sigma_b, x)\ dx} \qquad (E\text{-}1)$$

we can define a variable y such that:-

$$y = G(a, \sigma_a, x) . G(b, \sigma_b, x)$$

and therefore

$$Y = \frac{y}{\int_{-\infty}^{\infty} y\ dx} \qquad (E\text{-}2)$$

Expanding y gives:-

$$y = \frac{1}{2\pi[\sigma_a\sigma_b]^2} \exp\frac{-1}{2}\left\{\frac{(a-x)^2}{\sigma_a^2} + \frac{(b-x)^2}{\sigma_b^2}\right\}$$

$$= \frac{1}{2\pi[\sigma_a\sigma_b]^2} \exp\frac{-1}{2}\left\{\frac{\sigma_b^2(a-x)^2 + \sigma_a^2(b-x)^2}{\sigma_a^2\,\sigma_b^2}\right\} \qquad (E\text{-}3)$$

and separating the terms in x gives:-

$$y = \frac{1}{2\pi[\sigma_a\sigma_b]^2} \exp\frac{-1}{2}\left\{\frac{a^2}{\sigma_a^2} + \frac{b^2}{\sigma_b^2}\right\}$$

$$\times \quad \exp\frac{-1}{2}\left\{\frac{x^2(\sigma_a^2+\sigma_b^2)}{\sigma_a^2\sigma_b^2} - \frac{2x(a\sigma_b^2+b\sigma_a^2)}{\sigma_a^2\sigma_b^2}\right\} \qquad (E\text{-}4)$$

If we now define σ_c such that:-

$$\sigma_c^2 = \sigma_a^2\sigma_b^2/(\sigma_a^2 + \sigma_b^2) \qquad (E\text{-}5)$$

and c such that:-

$$c = (a\sigma_b^2 + b\sigma_a^2) / (\sigma_a^2 + \sigma_b^2) \qquad (E\text{-}6)$$

we can take the final terms of *(E-4)* and express them as:-

$$\frac{x^2(\sigma_a^2+\sigma_b^2)}{\sigma_a^2\sigma_b^2} = \frac{x^2}{\sigma_c^2} \qquad (E\text{-}7a)$$

and $$\frac{2x(a\sigma_b^2+b\sigma_a^2)}{\sigma_a^2\sigma_b^2} = \frac{2xc}{\sigma_c^2} \qquad (E\text{-}7b)$$

whence

$$\frac{x^2(\sigma_a^2+\sigma_b^2)}{\sigma_a^2\sigma_b^2} - \frac{2x(a\sigma_b^2+b\sigma_a^2)}{\sigma_a^2\sigma_b^2}$$

$$= \frac{(x-c)^2}{\sigma_c^2} - \frac{c^2}{\sigma_c^2} \qquad (E\text{-}8)$$

and therefore

$$\exp\frac{-1}{2}\left\{\frac{x^2(\sigma_a^2+\sigma_b^2)}{\sigma_a^2\sigma_b^2} - \frac{2x(a\sigma_b^2+b\sigma_a^2)}{\sigma_a^2\sigma_b^2}\right\}$$

$$= \exp\frac{1}{2}\left\{\frac{c^2}{\sigma_c^2}\right\} \exp\frac{-1}{2}\left\{\frac{(x-c)^2}{\sigma_c^2}\right\} \qquad (E\text{-}9)$$

Hence, substituting in *(E-4)*:-

$$y = \frac{1}{2\pi[\sigma_a\sigma_b]^2} \exp\frac{-1}{2}\left\{\frac{a^2}{\sigma_a^2}+\frac{b^2}{\sigma_b^2}-\frac{c^2}{\sigma_c^2}\right\} \exp\frac{-1}{2}\left\{\frac{(x-c)^2}{\sigma_c^2}\right\} \qquad (E\text{-}10)$$

which is equivalent to:-

$$y = \frac{\sqrt{(2\pi\sigma_c^2)}}{2\pi[\sigma_a\sigma_b]^2} \exp\frac{-1}{2}\left\{\frac{a^2}{\sigma_a^2}+\frac{b^2}{\sigma_b^2}-\frac{c^2}{\sigma_c^2}\right\}$$

$$x \quad \frac{1}{\sqrt{(2\pi\sigma_c^2)}} \exp\left\{\frac{(x-c)^2}{\sigma_c^2}\right\} \qquad (E\text{-}11)$$

Thus, putting

$$k_c = \frac{\sqrt{(2\pi\sigma_c^2)}}{2\pi[\sigma_a\sigma_b]^2} \exp\frac{-1}{2}\left\{\frac{a^2}{\sigma_a^2}+\frac{b^2}{\sigma_b^2}-\frac{c^2}{\sigma_c^2}\right\} \qquad (E\text{-}12)$$

gives

$$y = k_c G(c, \sigma_c, x) \qquad (E\text{-}13)$$

But,

$$\int_{-\infty}^{\infty} k_c . G(c, \sigma_c, x)\ dx = k_c \qquad (E\text{-}14)$$

therefore by substitution in *(E-2)*

$$Y = \frac{k_c \mathcal{G}(c, \sigma_c, x)}{k_c} \qquad \textit{(E-15a)}$$

$$= \mathcal{G}(c, \sigma_c, x) \qquad \textit{(E-16)}$$

Bibliography

Abbott, T.K. (1959), *Kant's Critique of Practical Reason and Other Works on the Theory of Ethics*, London, Longmans.

Andrewes, William J. H. (1996) *The Quest for Longitude : The Proceedings of the Longitude Symposium Harvard University, November 4-6, 1993*; Cambridge, Massachusetts, Harvard University Press

Apostol T.M. (1974) *Mathematical Analysis,* Reading Massachusetts, Addison-Wesley

Aristotle - See Ross, W.D. (1936)

Audi, R. (Ed) (1995) *The Cambridge Dictionary of Philosophy,* Cambridge University Press.

Bar-Hillel, M. (1980) *The base-rate fallacy in probability judgments*. Acta Psychologica 44:211-233

Bar-Hillel, M. (1983) *The base rate fallacy controversy*. In: *Decision making under uncertainty* (pp. 39-61), ed. R. W. Scholz. Elsevier Science

Bar-Hillel, M. (1990) *Back to base rates.* In: *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 200-216), ed. R. M. Hogarth. University of Chicago Press

Bar-Shalom,Y. and Fortmann,T.E. (1988) *Tracking and Estimation*, Boston, Academic Press.

Barnard G.A. (1967) *The Bayesian Controversy in Statistical Inference,* Journal of the Institute of Actuaries, London, Vol 93 PtII No 395, pp229-269

Barnard G.A. (1958) *Thomas Bayes - A Biographical Note,* Biometrika, Vol 45, Parts 3 and 4, (Studies in the History of Probability and Statistics).

Bayes, T. (1763) *An Essay towards solving a Problem in the Doctrine of Chances* Phil. Trans. Roy. Soc. vol 53**.**

Bennett J.H. *ed* (1990) *A Re-issue of 'Statistical methods for Research Workers', 'The Design of Experiments', and 'Statistical methods and Scientific Inference' by R.A.Fisher*, Oxford University Press.

Bennett, J.H. (ed) (1990), *'Statistical Inference and Analysis: Selected Correspondence of R.A.Fisher'*, Clarendon Press, Oxford.

Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian Theory,* Wiley, Chichester.

Bernoulli, Jacques (*or* James, Jakob), (1713) *Ars Conjectandi,* Basle.

Bernstein, Peter L. (1996) *Against The Gods,* Wiley, New York.

Bondi H. (1967) *Assumption and Myth in Physical Theory,* Cambridge, England C.U.P.

Boole G. (1854) *The Laws of Thought,* Reprinted, New York; Dover, 1951.

Boole G. (Various) *Studies in Logic and Probability,* Reprinted with corrections 1952, La Salle, Illinois; Open Court Publishing Company. Page references are to the 1952 edition.

Box, J.F. (1978) *R.A.Fisher, the life of a scientist*, New York, Wiley

Braithwaite R.B. (1968) *Scientific Explanation,* Cambridge, England; C.U.P.

Broad, C.D. (1918) *On The Relation Between Induction And Probability, (Part I)* MIND, New Series No 108, October 1918, pp389-404

Broad, W. and Wade, N. (1985) *Betrayers of the Truth*, Oxford, O.U.P.

Buderi, Robert (1996) *The Invention that Changed the World*, New York: Simon & Schuster.

Buehler, R.J. and Feddersen, A.P. (1963). Note on a conditional property of Student's *t*. *Ann. Math. Statist.* **34** 1098-1100

Buntine, Wray (1996) *A Guide to the Literature on Learning Probablistic Networks from Data*, IEEE Transactions on Knowledge and Data Engineering, Vol 8, No 2, April 1996

Cohen, L. J. (1979) *On the psychology of prediction: Whose is the fallacy?* Cognition 7: 385-407

Dale A.I. (1982) *Bayes or Laplace ? An examination of the origin and early applications of Bayes' theorem.* Archive for History of Exact Sciences Vol 27 pp 23-47

Dale A.I. (1991) *A History of Inverse Probability from Thomas Bayes to Karl Pearson,* New York, Springer-Verlag

Daintith J and Nelson R.D. (1989) *The Penguin Dictionary of Mathematics,* Penguin Books, Harmondsworth, England.

de Moivre A. (1756) *The Doctrine of Chance,* London, (3rd edition)

Doob J.L. (1953) *Stochastic Processes*, New York, Wiley.

Earman J. (1992) *Bayes or Bust ?* MIT Press, Cambridge Mass.

Edwards A.W.F. (1974) *A problem in the doctrine of chances.* In Conference on Foundational Questions in Statistical Inference, memoirs No 1, Dept of Theoretical Statistics, Aarhus University, pp 43-60

Edwards A.W.F. (1978) *'Commentary on the Arguments of Thomas Bayes'* Scandinavian Journal of Statistics

Edwards A.W.F. (1992) *Likelihood, Expanded Edition,* Baltimore; The Johns Hopkins University Press

Ellis B. (1966) *Basic Concepts of Measurement* Cambridge, England, C.U.P.

Ehrenberg, A.S.C., (1982), *How Good is Best*, J.R.Statist.Soc. A, **145,** Part 3, pp364-366

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications; Vol I,* 3rd edition, New York, Wiley

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications; Vol II,* 2nd edition, New York, Wiley

Fine T. (1973) *Theories of Probability* New York; Academic Press

Fisher, R.A. (1912) *On an absolute criterion for fitting frequency curves,* 'The Messenger of Mathematics' Vol XLI, (May1911-April 1912), pp155-160, (Cambridge, Bowes & Bowes; London, MacMillan, 1912)

Fisher, R.A. (1915) *frequency distribution of the values of the correlation co-efficient in samples from an indefinitely large population*, Biometrika, Vol X, pp507-552

Fisher R.A. (1921) *On the 'Probable Error' of a Coefficient of Correlation deduced from a Small Sample'* Metron, Vol 1 N^{o}4, pp3-32.

Fisher, R.A. (1922) *'The mathematical foundations of theoretical statistics'* Phil.Trans. Roy.Soc. A222, pp309-368

Fisher R.A. (1930) *Inverse Probability* Proceedings of the Cambridge Philosophical Society, Vol XXVI pt 4, pp528-535

Fisher R.A. (1937) *Professor Karl Pearson and the Method of Moments*, Annals of Eugenics, Vol VII, Pt. IV, pp303-318.

Fisher R.A. (1945) *The Logical Inversion of the Notion of the Random Variable*, SANKHY Ā, Vol 7, Part 2, pp129 - 132.

Fisher R.A. (1950) *'Contributions to Mathematical Statistics'* Wiley, New York

Fisher R.A. (1956) *Statistical Methods and Scientific Inference,* Edinburgh, Scotland; Oliver and Boyd. (Reprinted with corrections and additions by Oxford University Press 1990).

Fisher R.A. (1959) *Mathematical probability in the natural sciences*, Technometrics, Vol 1, pp21-29

Fisher R.A. (1962) *Some Examples of Bayes' Method of the Experimental Determination of Probabilities A Priori*, J.R.S.S. Series B, Vol 24, pp118 - 124.

Fowler, H.W. and F.G. (1930) *The King's English*, Oxford, Oxford University Press.

Gillies D.A. (1987) *Was Bayes a Bayesian?* Historia Mathematica Vol 14 pp325-346

Good, I.J. (1971) *The Probablistic Explication Of Information*, in *Foundations of Statistical Inference,* (Godambe, V.P. and Sprott, D.A. eds), Toronto: Holt, Rinehart and Winston

Hacking I. (1965) *Logic of Statistical Inference* Cambridge, England; C.U.P.

Hacking I, (1975) *The Emergence of Probability,* Cambridge, England; C.U.P.

Hacking I, (1990) *The Taming of Chance,* Cambridge, England; C.U.P.

Hartley D. (1749) *Observations on Man, his Frame, his Duty, and his Expectations* London: Richardson

Hartley, R.V.L. (1928) 'Transmission of Information', *Bell System Technical Journal,* July 1928 p535.

Heckerman, D. (1995) *'Learning Bayesian Networks: The Combination of Knowledge and Statistical Data';* Technical Report MSR-94-09, Microsoft Research, Advanced Technology Division, Redmond WA98052

Hesse M. (1974) *The Structure of Scientific Inference* , London, Macmillan.

Howson, C. and Urbach, P. (1993) *Scientific Reasoning: The Bayesian Approach*, Chicago: Open Court

Hume D. (1748) *An Enquiry Concerning Human Understanding* Selby-Bigge ed. Oxford: O.U.P. (1963)

Huygens, Christian (1657) *De Ratiociniis in aleae ludo,* published in *Exercitionum Mathematicorum,* ed F. van Schooten, Amsterdam. (See also Hacking (1975) p194).

Jaynes, E.T. (1982) *Papers on Probability, Statistics and Statistical Physics,* Riedel, Dordrecht

Jeffrey, Richard (1992) *Probability and the art of judgement*, Cambridge: C.U.P.

Jeffreys H. (1934) *Probability and scientific method,* in Proc. Roy. Soc. 'A', Vol 146, pp 9-16.

Jeffreys H. (1983) *Theories of Probability* (Third edition), Oxford, England; O.U.P.

Jensen, Finn V. (1996) *An Introduction to Bayesian Networks*, London, UCL Press

Kalman, R.E., *A New Approach to Linear Filtering and Prediction*, Trans A.S.M.E Ser D.J. Basic Eng, **82,** pp35-45

Kant, Immanuel, (1785), *Grundlegung zur Metaphysik der Sitten*

Keynes J.M. (1921) *A Treatise on Probability* London, Macmillan

Koehler, J.J. (1996). *The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges.* Behavioral and Brain Sciences vol.19 (1) pp1-53 (ISSN: 0140525X)

Kahnemann, D., Slovic, P., Tversky, A. (eds), (1982) *Judgement under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kolmogorov A. (1950) *Foundations of the Theory of Probability*, (Translated from the German of 1933 by N.Morrison), New York: Chelsea Publishing Company

Kyburg, H. (1974) *The Logical Foundations of Statistical Inference*, D.Reidel, Dordrecht.

Kyburg, H. (1983) *The reference class*. Philosophy of science, **50**:374-397

Laplace P.S. (1774) *'Mémoire sur la Probabilité des causes par les évènemens'* in Mémoires de l'Academie royale des sciences de Paris, vol 6 pp621-56 (See also Stigler (1986-b))

Laplace P.S. (1995) *A Philosophical Essay on Probabilities*, (English translation of *Essai philosophique sur les probabilités*, Paris 1814) New York, Dover.

Laplace P.S. (1820) *Théorie Analytique des Probabilités,* Paris, France.

Lindley D.V. (1965) *Introduction to Probability and Statistics,* (2 parts), Cambridge, C.U.P.

Luenberger, D.G. (1969) *'Optimisation by Vector Space Methods'*, New York, Wiley.

Markov, A.A. (1906) *'Extension of the law of large numbers to dependent events'* (Russian) Bull. Soc. Phys. Math. Kazan (2) 15, 135-156

Medawar, P.B. (1963) *'Is the Scientific Paper a Fraud?'* , The Listener, 12 Sept 1963, pp377-8. (BBC publications, London).

Meehl, Paul E., and Rosen, Albert (1955) *'Antecedent probability and the efficiency of pyschometric signs, patterns or cutting scores'*, Psychological Bulletin Vol 52, No 3 pp 194-216

Michell J (1767) *An inquiry into the probable parallax and magnitude of the fixed stars,* Phil.Trans.Roy.Soc. Vol 57 pp234-264.

Mossner E.C. (1954) *The Life of David Hume* Oxford, O.U.P.

Murray F.H. (1930) *Note on a Scholium of Bayes* Bulletin of the American Mathematical Society Vol 36 pp129-132

Molina E.C. (1931) *Bayes' Theorem: An Expository Presentation* , Annals of Mathematical Statistics, Vol 2, pp23-37.

Neyman J (1935) *On the problem of confidence intervals* Reprinted in *A selection of early statistical papers of J.Newman* pp142-6 Cambridge, C.U.P. (1967)

Neyman J (1937) *Outline of a theory of statistical estimation based on the classical theory of probability* Reprinted in *A selection of early statistical papers of J.Newman* pp250-290 Cambridge, C.U.P. (1967)

Neyman, J. (1961), *Silver Jubilee of my dispute with Fisher*, Journal of the Operations Research Society of Japan, Vol 3, Number 4 pp145-154

Norton, J.P. (1986), *An Introduction to Identification,* London, Academic Press

Pearson K (1900) *On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it*

can be reasonably supposed to have arisen from Random Sampling. The Philosophical Magazine, Vol 5, pp 157-175.

Pearson K (1907) *'On the Influence of Past Experience on Future Expectation'* in The Philosophical Magazine, 1907, pp 365-378

Pearson, K. (1920) *The Fundamental Problem of Practical Statistics*, Biometrica **XIII,** October 1920, No 1, pp.1-16.

Pearson, K. (1920 ?) *Note on the "Fundamental Problem of Practical Statistics"*, Biometrica **XIII** (?) pp.300-301

Pearson K (1978) *The History of Statistics in the Seventeenth and Eighteenth Centuries against the Changing Background of Intellectual, Scientific and Religious Thought*, London: Charles Griffin

Penrose R (1989) *The Emperor's New Mind,* Oxford University Press.

Piattelli-Palmarini, M. (1994) *Inevitable Illusions,* New York, Wiley

Polak, Elijah (1997) *Optimization: Algorithms and Consistent Approximations,* Springer, New York

Pollock D.S.G. (1999) *A Handbook of Time-Series Analysis, Signal Processing and Dynamics* , New York, Academic Press

Popper, K.R. (1972) *The Logic of Scientific Discovery,* London, England; Hutchinson.

Popper, K.R. (1973) *Objective Knowledge,* Oxford,O.U.P.

Poulton, E.C. (1994) *"Behavioral Decision Theory : A New Approach"*, Cambridge, C.U.P.

Price, R (1767) *Four Dissertations.* London, Millar and Cadell

Ramsey, F.P. (1931) *The Foundations of Mathematics*, edit. Braithwaite, London, Kegan Paul.

Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, (second edition), New York, Wiley

Ridley, M. (1999a) *This is your life - and how to build a man*, The Daily Telegraph, London, Monday, Aug 23rd 1999, p13

Ridley, M. (1999b) *Genome: The Autobiography of a Species in 23 Chapters*, London, Fourth Estate.

Ripley, B.D. (1981) *Spatial Statistics.* New York: Wiley.

Ross, W.D. (1936) *Aristotle's Physics*, Oxford, O.U.P.

Russell, Betrand (1922) Review of Keynes' *'Treatise on Probability'*, printed in *The Mathematical Gazette*, July , 1922

Russell, Bertrand (1948) *Human Knowledge, Its Scope and Limits,* London

Ryan K.C. (1980) *Historical development of R.A.Fisher's fiducial argument* Unpublished Ph.D. thesis, London University.

Savage, L.J. (1954) *The Foundations of Statistics,* New York, Wiley.

Seidenfeld, T. (1992) *R.A.Fisher's Fiducial Argument and Bayes' Theorem* Statistical Science Vol 7 No 3, pp358-368

Shafer G. (1982) *Bayes' two arguments for the rule of conditioning* Annals of Statistics Vol 10, No 4, pp1075-1089

Shannon,C.E. (1948), 'The Mathematical Theory of Communication', *Bell System Technical Journal,* July 1948. Reprinted and published by University of Illinois Press with same title.

SMSI an abbreviation for *Statistical Methods and Scientific Inference*, see Fisher (1956).

Stigler S.M. (1980) *Stigler's law of eponymy*, Transactions of the New York Academy of Sciences, 2nd series 39: 147-157

Stigler S.M. (1982) *Thomas Bayes's Bayesian Inference,* Journal of Royal Statistical Society, Series A, Vol 145, pp250-258.

Stigler S.M. (1983) *Who discovered Bayes's theorem ?* The American Statistician, Vol 37, pp290-296

Stigler S.M. (1986-a) *The History of Statistics* Cambridge, Mass; Belknap Press (Harvard).

Stigler S.M. (1986-b) *Laplace's 1774 memoire on inverse probability* Statistical Science vol 1.

Sobel, Dava. (1996) *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time.* New York, Penguin.

Soper, A.W., Young, A.W., Cave, B.M., Lee, A., Pearson, K., (1916), *Co-operative study On the distribution of the correlation co-efficient in small samples*

Thomas, D.O. (1977) *The Honest Mind: The Thought and Work of Richard Price,* Oxford, O.U.P.

Turberville A.S. (1926) *Men and Manners in the Eighteenth Century,* Oxford, O.U.P.

Turkle, Sherry (1995) *Life on the Screen*, New York, Simon and Schuster.

Van Trees, H.L. (1968) *Detection, Estimation and Modulation Theory*, Volume 1, New York: Wiley.

Von Plato, J. (1994) *Creating Modern Probability*, Cambridge, C.U.P.

Wansell, Geoffrey (2000) *"Four months ago, Sally Clark was jailed for life .....",* the Daily Mail, London, Saturday March 11, 2000 p42.

Waterman, Talbot (1989) *Animal Navigation* Scientific American Press

Whittle P. (1970) *Probability* Penguin Books, Harmondsworth.

Woodward, P.M. (1964) *Probability and Information Theory with Applications to Radar*, Pergamon Press.

Zabell, S.L. (1992) *R.A.Fisher and the Fiducial Argument*, Statistical Science Vol 7, No 3, pp369-387

INDEX