CHAPTER 15

PARADOXES OF PROBABILITY THEORY

"I protest against the use of infinite magnitude as something accomplished, which is never permissible in mathematics. Infinity is merely a figure of speech, the true meaning being a limit."

— C. F. Gauss

The term "paradox" appears to have several different common meanings. Székely (1986) defines a paradox as anything which is true but surprising. By that definition, every scientific fact and every mathematical theorem qualifies as a paradox for someone. We use the term in almost the opposite sense; something which is absurd or logically contradictory, but which appears at first glance to be the result of sound reasoning. Not only in probability theory, but in all mathematics, it is the careless use of infinite sets, and of infinite and infinitesimal quantities, that generates most paradoxes.

In our usage, there is no sharp distinction between a paradox and an error. A paradox is simply an error out of control; *i.e.* one that has trapped so many unwary minds that it has gone public, become institutionalized in our literature, and taught as truth. It might seem incredible that such a thing could happen in an ostensibly mathematical field; yet we can understand the psychological mechanism behind it.

How do Paradoxes Survive and Grow?

As we stress repeatedly, from a false proposition – or from a fallacious argument that leads to a false proposition – all propositions, true and false, may be deduced. But this is just the danger; if fallacious reasoning always led to absurd conclusions, it would be found out at once and corrected. But once an easy, short–cut mode of reasoning has led to a few correct results, almost everybody accepts it; those who try to warn against it are not listened to.

When a fallacy reaches this stage it takes on a life of its own, and develops very effective defenses for self–preservation in the face of all criticisms. Mathematicians of the stature of Henri Poincaré and Hermann Weyl tried repeatedly to warn against the kind of reasoning used in infinite set theory, with zero success. For details, see Appendix B and Kline (1980). The writer was also guilty of this failure to heed warnings for many years, until absurd results that could no longer be ignored finally forced him to see the error in an easy mode of reasoning.

To remove a paradox from probability theory will require, at the very least, detailed analysis of the result and the reasoning that leads to it, showing that:

(1) The result is indeed absurd.

(2) The reasoning leading to it violates the rules of inference developed in Chapter 2.

(3) When one obeys those rules, the paradox disappears and we have a reasonable result.

There are too many paradoxes contaminating the current literature for us to analyze separately. Therefore we seek here to study a few representative examples in some depth, in the hope that the reader will then be on the alert for the kind of reasoning which leads to them.

Summing a Series the Easy Way

As a kind of introduction to fallacious reasoning with infinite sets, we recall an old parlor game by which you can prove that any given infinite series $S = \sum_i a_i$ converges to any number x that your victim chooses. The sum of the first n terms is $s_n = a_1 + a_2 + \ldots + a_n$. Then, defining $s_0 \equiv 0$, we have

$$a_n = (s_n - x) \; - \; (s_{n-1} - x) \,, \qquad 1 \le n < \infty$$

so that the series becomes

$$\begin{aligned} S = (s_1 - x) + (s_2 - x) + (s_3 - x) + \ldots \\ - (s_0 - x) - (s_1 - x) - (s_2 - x) - \ldots \end{aligned} \tag{15–1}$$

The terms $(s_1 - x)$, $(s_2 - x), \ldots$ all cancel out, so the sum of the series is

$$S = -(s_0 - x) = x \qquad\qquad QED. \tag{15–2}$$

The reader for whom this reasoning appears at first glance to be valid has a great deal of company, and is urged to study this example carefully. Such fallacious arguments are avoided if we follow this advice, repeated from Chapter 2:

> *Apply the ordinary processes of arithmetic and analysis only to expressions with a finite number n of terms. Then after the calculation is done, observe how the resulting finite expressions behave as the parameter n increases indefinitely.*

Put more succinctly, passage to a limit should always be the last operation, not the first. In case of doubt, this is the only safe way to proceed. Our present theory of convergence of infinite series could never have been achieved if its founders – Abel, Cauchy, d'Alembert, Dirichlet, Gauss, Weierstrasz, and others – had not followed this advice meticulously. In pre–Bourbakist mathematics (such as Whittaker and Watson, 1927) this policy was considered so obvious that there was no need to stress it. The results thus obtained have never been found defective.

Had we followed this advice above, we would not have tried to cancel out an infinite number of terms in a single stroke; we would have found that at any finite (n'th) stage, instead of the s_i's cancelling out and one x remaining, the x's would have cancelled out and the last s remains, leading to the correct summation of the series.

Yet today, reasoning essentially equivalent to what we did in (15–1) is found repeatedly where infinite sets are used in probability theory. As an example, we examine another of the consequences of ignoring this advice, which has grown into far more than a parlor game.

Nonconglomerability

If $(C_1 \cdots C_n)$ denote a finite set of mutually exclusive, exhaustive propositions on prior information I, then for any proposition A the sum and product rules of probability theory give

$$P(A|I) = \sum_{i=1}^{n} P(AC_i|I) = \sum_{i=1}^{n} P(A|C_iI)\, P(C_i|I) \tag{15–3}$$

in which the prior probability $P(A|I)$ is written as a weighted average of the conditional probabilities $P(A|C_iI)$. Now it is a very elementary theorem that a weighted average of a set of real numbers cannot lie outside the range spanned by those numbers; if

$$L \le P(A|C_i I) \le U \quad , \qquad (1 \le i \le n) \tag{15–4}$$

then necessarily

$$L \le P(A|I) \le U \ , \tag{15–5}$$

a property which de Finetti (1972) called "conglomerability" or more precisely, "conglomerability in the partition $\{C_i\}$", although it may seem too trivial to deserve a name. Obviously, nonconglomerability cannot arise from a correct application of the rules of probability theory on finite sets. It cannot, therefore occur in an infinite set which is approached as a well–defined limit of a sequence of finite sets.

Yet nonconglomerability has become a minor industry, with a large and growing literature. There are writers who believe that it is a real phenomenon, and that they are proving theorems about the circumstances in which it occurs, which are important for the foundations of probability theory. Nonconglomerability has become, quite literally, institutionalized in our literature and taught as truth.

In spite of its mathematical triviality, then, we need to examine some cases where nonconglomerability has been claimed. Rather than trying to cite all of this vast literature, we draw upon a single recent reference (Kadane, Schervish, & Seidenfeld, 1986), hereafter denoted by KSS, where several examples and references to other work may be found.

Example 1: Rectangular Array: First we note the typical way in which nonconglomerability is manufactured, and the illustrative example most often cited. We start from a two-dimensional $(M \times N)$ set of probabilities:

$$p(i,j), \qquad 1 \le i \le M, \qquad 1 \le j \le N \tag{15–6}$$

and think of i plotted horizontally, j vertically so that the sample space is a rectangular array of MN points in the first quadrant. It will suffice to take some prior information I for which these probabilities are uniform: $p(i,j) = (1/MN)$. Then the probability of the event $(A: \ i < j)$ is found by direct counting to be:

$$P(A|I) = \left\{ \begin{array}{rl} (2N - M - 1)/2N, & \quad M \le N \\ (N-1)/2M, & \quad N \le M \end{array} \right\} \tag{15–7}$$

Let us resolve this in the manner of (15–3), into probabilities conditional on the set of propositions $(C_1 \cdots C_M)$, where C_i is the statement that we are on the i'th column of the array: then $P(C_i|I) = (1/M)$ and

$$P(A|C_i I) = \left\{ \begin{array}{rl} (N-i)/N, & \quad 1 \le i \le M \le N \\ (N-i)/N, & \quad 1 \le i \le N \le M \\ 0, & \quad N \le i \le M \end{array} \right\} . \tag{15–8}$$

These conditional probabilities reach the upper and lower bounds

$$U = (N-1)/N, \qquad \textit{all } M, N$$
$$L = \left\{ \begin{array}{rl} 1 - R, & \quad M \le N \\ 0, & \quad N \le M \end{array} \right\} , \tag{15–9}$$

where R denotes the ratio $R = M/N$. Substituting (15–7) and (15–9) into (15–5), it is evident that the condition for conglomerability is always satisfied, as it must be, whatever the values of (M, N). How, then can one possibly create a nonconglomerability out of this?

Just pass to the limit $M \to \infty$, $N \to \infty$, and ask for the probabilities $P(A|C_i I)$ for $i = 1, 2, \cdots$. But instead of examining the limiting form of (15–8), which gives the exact values for all (M, N), we try to evaluate these probabilities directly on the infinite set.

Then it is argued that, for any given i, there are an infinite number of points where A is true and only a finite number where it is false. *Ergo*, the conditional probability $P(A|C_i I) = 1$ for all i; yet $P(A|I) < 1$. We see here the same kind of reasoning that we used in (15–1); we are trying to carry out very simple arithmetic operations (counting), but directly on an infinite set.

Now consider the set of propositions $(D_1, \ldots, D_N)$, where D_j is the statement that we are on the j'th row of the array, counting from the bottom. Now, by the same argument, for any given j, there are an infinite number of points where A is false, and only a finite number where A is true. *Ergo* the conditional probability $P(A|D_j I) = 0$ for all j; yet $P(A|I) > 0$. By this reasoning, we have produced two nonconglomerabilities, in opposite directions, from the same model (*i.e.*, the same infinite set).

But it is even more marvelous than that. In (15–7) it is true that if we pass to the limit holding i fixed, the conditional probability $P(A|C_i B)$ tends to 1 for all i; but if instead we hold $(N - i)$ fixed, it tends to 0 for all i. Therefore, if we consider the cases $(i = 1, i = 2, ...)$ in increasing order, the probabilities $P(A|C_i B)$ appear to be 1 for all i. But it is equally valid to consider them in decreasing order $(i = N, i = N - 1, \ldots)$; and then by the same reasoning they would appear to be 0 for all i. [Note that we could redefine the labels by subtracting $N + 1$ from each one, thus numbering them $(i = -N, \ldots, i = -1)$ so that as $N \to \infty$ the upper indices stay fixed; this would have no effect on the validity of the reasoning.]

Thus to produce two opposite nonconglomerabilities we need not introduce two different partitions $\{C_i\}$, $\{D_j\}$; they can be produced by two equally valid arguments from a single partition. What produces them is that one supposes the infinite limit already accomplished *before* doing the arithmetic, reversing the policy of Gauss which we recommended above. But if we follow that policy and do the arithmetic first, then an arbitrary redefinition of the labels $\{i\}$ has no effect; the counting for any N is the same.

Once one has understood the fallacy in (15–1), then whenever someone claims to have proved some result by carrying out arithmetic or analytical operations directly on an infinite set, it is hard to shake off a feeling that he could have proved the opposite just as easily and by an equally sound argument, had he wished to. Thus there is no reason to be surprised by what we have just found.

Suppose that instead we had done the calculation by obeying our rules strictly, doing first the arithmetic operations on finite sets to obtain the exact solution (15–7); then passing to the limit. However the infinite limit is approached, the conditional probabilities take on values in a wide interval whose lower bound is 0 or $1 - R$, and whose upper bound tends to 1. The condition (15–4) is always satisfied, and a nonconglomerability could never have been found.

The reasoning leading to this nonconglomerability contains another fallacy. Clearly, one cannot claim to have produced a nonconglomerability on the infinite set until the 'unconditional' probability $P(A|I)$ has also been calculated on that set, not merely bounded by a verbal argument. But as M and N increase, from (15–7) the limiting $P(A|I)$ depends only on the ratio $R = M/N$:

$$P(A|I) \to \left\{ \begin{array}{cc} 1 - R/2, & R \le 1 \\ 1/(2R), & R \ge 1 \end{array} \right\} . \qquad (15\text{–}10)$$

If we pass to the infinite limit without specifying the limiting ratio, the unconditional probability $P(A|I)$ becomes indeterminate; we can get any value in $[0, 1]$ depending on how the limit is approached. Put differently, the ratio R contains all the information relevant to the probability of A; yet it was thrown away in passing to the limit too soon. The unconditional probability $P(A|I)$ could not have been evaluated directly on the infinite set, any more than could the conditional probabilities.

Thus nonconglomerability on a rectangular array, far from being a phenomenon of probability theory, is only an artifact of failure to obey the rules of probability theory as developed in Chapter 2. But from studying a single example we cannot see the common feature underlying all claims of nonconglomerability.

Strong Inconsistency

We now examine a claim that nonconglomerability can occur even in a one–dimensional infinite set $n \to \infty$ where there does not appear to be any limiting ratio like the above M/N to be ignored. Also we now consider a problem of inference, instead of the above sampling distribution example. The scenario has been called the "Strong Inconsistency Problem" (Stone, 1970). We follow the KSS notation for the time being – until we see why we must not.

A regular tetrahedron with faces labelled e^+ (positron), e^- (electron), μ^+ (muon), μ^- (antimuon), is tossed repeatedly. A record is kept of the result of each toss, except that whenever a record contains e^+ followed immediately by e^- (or e^- by e^+, or μ^+ by μ^-, or μ^- by μ^+), the particles annihilate each other, erasing that pair from the record. At some arbitrary point in the sequence the player (who is ignorant of what has happened to date) calls for one more toss, and then is shown the final record $x \in X$, after which he must place bets on the truth of the proposition $A \equiv$ "Annihilation occurred at the final toss". What probability $P(A|x)$ should he assign?

When we try to answer this by application of probability theory, we come up immediately against the difficulty that in the problem as stated, the solution depends on a nuisance parameter, the unspecified length n of the original sequence of tosses. This was pointed out by B. Hill (1980), but KSS take no note of it. In fact. they do not mention n at all except by implication, in a passing remark that the die is "rolled a very large number of times." We infer that they meant the limit $n \to \infty$, from later phrases such as 'the countable set S' and 'every finite subset of S'.

In other words, once again an infinite set is supposed to be something already accomplished, and one is trying to find relations between probabilities by reasoning directly on the infinite set. Nonconglomerability enters through asking whether the prior probability $P(A)$ is conglomerable in the partition x, corresponding to the equation

$$P(A) = \sum_{x \epsilon X} P(A|x)\, P(x). \qquad (15\text{–}11)$$

KSS denote by $\theta \,\epsilon\, S$ the record just before the final toss (thought of as a 'parameter' not known by the player), where S is the set of all possible such records, and conclude by verbal arguments that:

(a) $0 \le p(A|\theta) \le 1/4\,, \qquad all \quad \theta \epsilon S$

(b) $3/4 \le p(A|x) \le 1\,, \qquad all \quad x \epsilon X.$

It appears that another violent nonconglomerability has been produced; for if $P(A)$ is conglomerable in the partition $\{x\}$ of final records, it must be true that $3/4 \le P(A) \le 1$, while if it is conglomerable in the partition $\{\theta\}$ of previous records, we require $0 \le P(A) \le 1/4$; it cannot be conglomerable in both. So where is the error this time?

We accept statement (a); indeed, given the independence of different tosses, knowing anything whatsoever about the earlier tosses gives us no information about the final one, so the uniform prior assignment 1/4 for the four possible results of the final toss still holds. Therefore, $p(A|\theta) = 1/4$ except when the record θ is blank, in which case there is nothing to annihilate, and so $p(A|\theta) = 0$. But this argument does not hold for statement (b); since the result of the final toss affects the final record x, it follows that knowing x must give some information about the final toss, invalidating the uniform 1/4 assignment.

Also, the argument that KSS gave for statement (b) supposed prior information different from that used for statement (a). This was concealed from view by the notation $p(A|\theta)$, $p(A|x)$ which fails to indicate prior information I. Let us repeat (15–11) with adequate notation:

$$P(A|I) = \sum_{x \epsilon X} P(A|xI)\, P(x|I). \tag{15–12}$$

Now as I varies, all these quantities will in general vary. By 'conglomerability' we mean, of course, 'conglomerability *with some particular fixed prior information* I. Recognizing this, we repeat statements (a) and (b) in a notation adequate to show this difference:

(a) $0 \le p(A|\theta, I_a) \le 1/4\,, \qquad \theta \in S$

(b) $3/4 \le p(A|x, I_b) \le 1\,, \qquad x \in X$

From reading KSS we find that prior information I_a, in effect, assigned uniform probabilities on the set T of 4^n possible tosses, as is appropriate for the case of 'independent repetitions of a random experiment' assumed in the statement of the problem. But I_b assigned uniform probabilities on the set S of different previous records θ. This is very different; an element of S (or X) may correspond to one element of T; or to many millions of elements of T, so a probability assignment uniform on the set of tosses is very nonuniform on the set of records. Therefore it is not evident whether there is any contradiction here; they are statements about two quite different problems.

Exercise 15.1 In $n = 40$ tosses there are $4^n = 1.21 \times 10^{24}$ possible sequences of results in the set T. Show that, if those tosses give the expected number $m = 10$ of annihilations leading to a record $x \in X$ of length 20, the specific record x corresponds to about 10^{14} elements of T. On the other hand, if there are no annihilations, the resulting record x of length 40 corresponds to only one element of T.

Perhaps this makes clearer the reason for our seemingly fanatical insistence on indicating the prior information I explicitly in every formal probability symbol $P(A|BI)$. Those who fail to do this may be able to get along without disaster for a while, judging the meaning of an equation from the surrounding context rather than from the equation as written. But eventually they are sure to find themselves writing nonsense, when they start inadvertently using probabilities conditional on different prior information in the same equation or the same argument; and their notation conceals that fact. We shall see presently a more famous and more serious error (the Marginalization Paradox) caused by failure to indicate the fact that two probabilities are conditional on different prior information.

To show the crucial role that n plays in the problem, let I agree with I_a in assigning equal prior probabilities to each of the 4^n outcomes of n tosses. Then if n is known, calculations of $p(A|nI)$, $p(x|nI)$, $p(A|nxI)$ are determinate combinatorial problems on finite sets (*i.e.* in each case there is one and only one correct answer), and the solutions obviously depend on n. So let us try to calculate $P(A|xI)$; denoting summation over all n in $(0 \le n < \infty)$ by Σ, we have for the prior probabilities

$$
\begin{aligned}
p(A|I) &= \Sigma\, p(An|I) = \Sigma\, p(A|nI)\, p(n|I) \\
p(x|I) &= \Sigma\, p(xn|I) = \Sigma\, p(x|nI)\, p(n|I)
\end{aligned}
\tag{15–13}
$$

and for the conditional one

$$
p(A|xI) = \Sigma\, p(A|nxI)\, p(n|xI) = \frac{\Sigma\, p(A|nxI)\, p(x|nI)\, p(n|I)}{\Sigma\, p(x|nI)\, p(n|I)} \tag{15–14}
$$

where we expanded $p(n|xI)$ by Bayes' theorem. It is evident that the problem is indeterminate until the prior probabilities $p(n|I)$ are assigned. Quite generally, failure to specify the prior information makes a problem of inference just as ill–posed as does failure to specify the data.

Passage to infinite n then corresponds to taking the limit of prior probabilities $p(n|I)$ that are nonzero only for larger and larger n. Evidently, this can be done in many different ways, and the final results will depend on which limiting process we use unless $p(A|nI)$, $p(x|nI)$, $p(A|nxI)$ all approach limits independent of n.

The number of different possible records x is less than 4^n (asymptotically, about 3^n) because many different outcomes with annihilation may produce the same final record, as the above exercise shows. Therefore for any $n < \infty$ there is a finite set X of different possible final records x, and *a fortiori* a finite set S of previous records θ, so the prior probability of final annihilation can be written in either of the forms:

$$
p(A|nI) = \sum_{x \epsilon X} p(A|xnI)\, p(x|nI) = \sum_{\theta \epsilon S} p(A|\theta nI)\, p(\theta|nI) \tag{15–15}
$$

and the general theorem on weighted averages guarantees that nonconglomerability cannot occur in either partition for any finite n, or for an infinite set generated as the limit of a sequence of these finite sets.

A few things about the actual range of variability of the conditional probabilities $p(A|nxI)$ can be seen at once without any calculation. For any n, there are possible records of length n for which we know that no annihilation occurred; the lower bound is always reached for some x, and it is $p(A|nxI) = 0$, not 3/4. The lower bound in statement (b) could never have been found for any prior information, had the infinite set been approached as a limit of a sequence of finite sets. Furthermore, for any even n there are possible records of length zero for which we know that the final toss was annihilated; the upper bound is always reached for some x and it is $p(A|nxI) = 1$.

Likewise, for even n it is not possible for θ to be blank, so from (15–15) we have $p(A|nI) = p(A|\theta nI) = 1/4$ for all $\theta \epsilon S$. Therefore, if n is even, there is no need to invoke even the weighted average theorem; there is no possibility for nonconglomerability in either the partition $\{x\}$ or $\{\theta\}$.

At this point it is clear that the issue of nonconglomerability is disposed of in the same way as in our first example; it is an artifact of trying to calculate probabilities directly on an infinite set without considering any limit from a finite set. Then it is not surprising that KSS never found any specific answer to their original question: "What we can infer about final annihilation from the final record x?" But we would still like to see the answer (particularly since it reveals an even more startling consequence of jumping directly into the infinite set).

The Solution for Finite Number of Tosses

If n is known, we can get the exact analytical solution easily from valid application of our rules. It is a straightforward Bayesian inference in which we are asking only for the posterior probability of final annihilation A. But this enables us to simplify the problem; there is no need to draw inferences about every detail of the previous record θ.

If there is annihilation at the n'th toss, then the length of the record decreases by one: $y(n) = y(n-1) - 1$. If there is no annihilation at the n'th toss, the length increases by one: $y(n) = y(n-1) + 1$. The only exception is that $y(n)$ is not permitted to become negative; if $y(n-1) = 0$ then the n'th toss cannot give annihilation. Therefore, since the available record x tells us the length $y(n)$ but not $y(n-1)$, any reasoning about final annihilation may be replaced immediately by reasoning about $\alpha \equiv y(n-1)$, which is the sole parameter needed in the problem.

Likewise, any permutations of the symbols $\{e^{\pm}, \mu^{\pm}\}$ in $x(n)$ which keep the same $y(n)$ will lead to just the same inferences about A. But then n and $y \equiv y(n)$ are sufficient statistics; all other details of the record x are irrelevant to the question being asked. Thus the scenario of the tetrahedrons is more complicated than it needs to be in order to define the mathematical problem (in fact, so complicated that it seems to have prevented recognition that it is a standard textbook random walk problem).

At each (n'th) toss we have the sampling probability $1/4$ of annihilating, independently of what happened earlier (with a trivial exception if $y(n-1) = 0$). Therefore if we plot n horizontally, $y(n)$ vertically, we have the simplest random walk problem in one dimension, with a perfectly reflecting boundary on the horizontal axis $y = 0$. At each horizontal step, if $y > 0$ there is probability $3/4$ of moving up one unit, $1/4$ of moving down one unit; if $y = 0$, we can move only up. Starting with $y(0) = 0$, annihilation cannot occur on step 1, and immediately after the n'th step, if there have been m annihilations, the length of the record is $y(n) = n - 2m$.

After the n'th step we have a prior probability distribution for $y(n)$ to have the value i:

$$p_i^{(n)} \equiv p(i|nI) , \qquad 0 \le i \le n \tag{15–16}$$

with the initial vector

$$p_i^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \tag{15–17}$$

and successive distributions are connected by the Markov chain relation

$$p_i^{(n)} = \sum_{j=0}^{n-1} M_{ij}\, p_j^{(n-1)} , \qquad \begin{matrix} 0 \le i \le n \\ 1 \le n < \infty \end{matrix} \tag{15–18}$$

with the transition matrix (number the rows and columns starting with zero):

$$M \equiv \begin{pmatrix} 0 & 1/4 & 0 & 0 & 0 & \dots \\ 1 & 0 & 1/4 & 0 & 0 & \dots \\ 0 & 3/4 & 0 & 1/4 & 0 & \dots \\ 0 & 0 & 3/4 & 0 & 1/4 & \dots \\ 0 & 0 & 0 & 3/4 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{15–19}$$

The reflecting boundary at $y = 0$ is indicated by the element $M_{10} = 1$, which would be $3/4$ without the reflection.

The matrix M is in principle infinite dimensional, but for the n'th step only the first $n + 1$ rows and columns are needed. The vector $p^{(n)}$ is also in principle infinite dimensional, but $p_i^{(n)} = 0$ when $i > n$. Then the exact solution for the prior probabilities $p_i^{(n)}$ is the first column of M^n:

$$p_i^{(n)} = M_{i,0}^n \tag{15–20}$$

[note that this is intended to represent $(M^n)_{i,0}$, not $(M_{i,0})^n$]. The short computer program RANDWALK in Appendix I prints out this solution.

Now let us see how this prior is to be used in our Bayesian inference problem. Denote the data and the hypothesis being tested by

$$D \equiv \text{“}y(n) = i\text{”}\,, \qquad\qquad H \equiv \text{“}y(n-1) = \alpha\text{”}\,,$$

which are the only parts of the data x and the parameter θ that are relevant to our problem. From the above their prior probabilities are

$$p(D|I) = M_{i,0}^n, \qquad\qquad p(H|I) = M_{\alpha,0}^{n-1} \tag{15–21}$$

The sampling distribution is

$$p(D|H,I) = \begin{Bmatrix} 3/4\,\delta(i,\alpha+1) + 1/4\,\delta(i,\alpha-1)\,, & \alpha > 0 \\ \delta(i,1)\,, & \alpha = 0 \end{Bmatrix} \tag{15–22}$$

So Bayes' theorem gives the posterior probability for α:

$$p(H|D,I) = p(H|I)\,\frac{p(D|H,I)}{p(D|I)} = \frac{M_{\alpha,0}^{n-1}}{M_{i,0}^n}\begin{Bmatrix} 3/4\,\delta(i,\alpha+1) + 1/4\,\delta(i,\alpha-1)\,, & \alpha > 0 \\ \delta(i,1)\,, & \alpha = 0 \end{Bmatrix} \tag{15–23}$$

Now final annihilation A occurs if and only if $\alpha = i+1$, so the exact solution for finite n is

$$p(A|D,n,I) = \frac{M_{i+1,0}^{n-1}}{4\,M_{i,0}^n} \tag{15–24}$$

in which $i = y(n)$ is a sufficient statistic. Another way of writing this is to note that the denominator of (15–24) is

$$4\,M_{i.0}^n = 4\sum_j M_{i,j}\,M_{j,0}^{n-1} = 3\,M_{i-1,0}^{n-1} + M_{i+1,0}^{n-1}$$

and so the posterior odds on A are

$$o(A|DnI) \equiv \frac{p(A|xnI)}{p(\overline{A}|xnI)} = \frac{1}{3}\,\frac{M_{i+1,0}^{n-1}}{M_{i-1,0}^{n-1}}\,, \tag{15–25}$$

and it would appear, from their remarks, that the exact solution to the problem that KSS had in mind is the limit of (15–24) or (15–25) as $n \to \infty$.

This solution for finite n is complicated because of the reflecting boundary. Without it, the aforementioned matrix element $M_{1,0}$ would be 3/4 and the problem would reduce to the simplest of all random walk problems. That solution gives us a very good approximation to (15–24), which actually yields the exact solution to our problem in the limit. Let us examine this alternative formulation because its final result is very simple and the derivation is instructive about a point that is not evident from the above exact solution.

The problem where at each step there is probability p to move up one unit, $q = 1-p$ to move down one unit, is defined by the recursion relation in which $f(i|n)$ is the probability to move a total distance i in n steps:

$$f(i|n+1) = p\,f(i-1|n) + q\,f(i+1|n) \tag{15–26}$$

With initial conditions $f(i|n=0) = \delta(i,0)$, the standard textbook solution is the binomial for r successes in n trials; $f_0(i|n) = b(r|n,p)$ with $r = (n+i)/2$. In our problem we know that on the first step we necessarily move up: $y(1) = 1$, so our initial conditions are $f(i|n=1) = \delta(i,1)$, and using the binomial recursion (15–26) after that the solution would be $f(i|n) = f_0(i-1|n-1) = b(r|n-1,p)$ with again $r = (n+i)/2$.

But with $p = 3/4$, this is not exactly the same as (15–18) because it neglects the reflecting boundary. If too many 'failures' (*i.e.*, annihilations) occur early in the sequence, this could reduce the length of the record to zero, forcing the upward probability for the next step to be 1 rather than 3/4; and (15–18) is taking all that into account. Put differently, in the solution to (15–26), when n is small some probability drifts into the region $y < 0$; but if $p = 3/4$ the amount is almost negligibly small and it all returns eventually to $y > 0$.

But when n is very large the solution drifts arbitrarily far away from the reflecting boundary, putting practically all the probability into the region $(-\sqrt{n} < y-\hat{y} < \sqrt{n})$ where $\hat{y} \equiv (p-q)n = n/2$, so conclusions drawn from (15–26) become highly accurate (in the limit, exact).

The sampling distribution (15–22) is unchanged, but we need binomial approximations to the priors for i and α. The latter is the length of the record after $n-1$ steps, or tosses. No annihilation is possible at the first toss, so after $n-1$ tosses we know that there were $n-2$ tosses at which annihilation could have occurred, with probability 1/4 at each, so the prior probability for m annihilations in the first $n-1$ tosses is the binomial $b(m|n-2, 1/4)$:

$$f(m) \equiv p(m|n) = \binom{n-2}{m}\left(\frac{1}{4}\right)^m \left(\frac{3}{4}\right)^{n-2-m}, \qquad 0 \le m \le n-2 \tag{15–27}$$

Then the prior probability for α, replacing the numerator in (15–25), is

$$p(\alpha|n) = f\left(\frac{n-1-\alpha}{2}\right) \tag{15–28}$$

from which we find the prior expectation $E(\alpha|I) = n/2$. Likewise in the denominator we want the prior for $y(n) = i$. This is just (15–28) with the replacements $n-1 \to n, \quad \alpha \to i$.

Given y, the possible values of α are $\alpha = y \pm 1$, so the posterior odds on final annihilation are, writing $m \equiv (n-y)/2$,

$$o = \frac{p(A|y,n)}{p(\overline{A}|y,n)} = \frac{p(\alpha = y+1|y,n)}{p(\alpha = y-1|y,n)} = \frac{\frac{1}{4}\cdot\binom{n-2}{m-1}\left(\frac{1}{4}\right)^{m-1}\left(\frac{3}{4}\right)^{n-1-m}}{\frac{3}{4}\cdot\binom{n-2}{m}\left(\frac{1}{4}\right)^{m}\left(\frac{3}{4}\right)^{n-2-m}}. \tag{15–29}$$

But, at first sight astonishing, the factors (1/4), (3/4) cancel out, so the result depends only on the factorials:

$$o = \frac{m!\,(n-2-m)!}{(m-1)!\,(n-1-m)!} = \frac{n-y}{n-2+y} \tag{15–30}$$

and the posterior probability of final annihilation reduces simply to

$$p(A|y,n) = \frac{o}{1+o} = \frac{n-y}{2(n-1)}, \tag{15–31}$$

which does not bear any resemblance to any of the solutions proposed by those who tried to solve the problem by reasoning directly on infinite sets. The sampling probabilities $p = 3/4$, $q = 1/4$ that figured so prominently in previous discussions, do not appear at all in this solution.

But now *think* about it: Given n and $y(n)$, we know that annihilation might have occurred in any of $n-1$ tosses, but that in fact it did occur in exactly $(n-y)/2$ tosses. But we have no information about which tosses, so the posterior probability for annihilation at the final toss (or at any toss after the first) is, of course,

$$\frac{n-y}{2(n-1)} \,. \tag{15–32}$$

We derived (15–31) directly from the principles of probability theory by a rather long calculation; but with a modicum of intuitive understanding of the problem, we could have reasoned it out in our heads without any calculation at all!

In Fig. 15.1 we compare the exact solution (15–24) with the asymptotic solution (15–31). The difference is negligible numerically when $n > 20$.

But then, why did so many people think the answer should be 1/4? Perhaps it helps to note that the prior expectation for y is $E(y|I) = (n+1)/2$, so the *predictive probability* of final annihilation is

$$p(A|nI) = \frac{n - E(y|I)}{2(n-1)} = \frac{1}{4} \,. \tag{15–33}$$

Then the posterior probability of final annihilation is indeed 1/4, *if the observed record length y is the expected value.* Quite generally in probability theory, if our new information is only what we already expected, that does not change any of our estimates; it only makes us more confident of them. But if y is observed to be different from its prior expectation, this tells us the *actual number* of annihilations, and of course this information takes precedence over whatever initial probability assignments $(1/4, 3/4)$ we might have made. That is why they cancelled out in the posterior odds.† In spite of our initial surprise, then, Bayes' theorem is doing exactly the right thing here; and the exact solution of the problem originally posed is given also by the limit of (15–32) as $n \to \infty$:

$$p(A|xI) = \frac{1}{2}(1 - z) \tag{15–34}$$

where $z \equiv \lim \; y(n)/n$.

In summary, the common feature of these two claims of nonconglomerability is now apparent. In the first scenario, there was no mention of the existence of the finite numbers M, N whose ratio M/N is the crucial quantity on which the solution depends. In the second scenario, essentially the same thing was done; failure to introduce the length n of the sequence – and, incredibly, even the length $y(n)$ of the observed record – likewise causes one to lose the crucial thing (in this case, the sufficient statistic y/n) on which the solution depends. In both cases, by supposing the infinite limit as something already accomplished at the start, *one is throwing away the very information required to find the solution.*

This has been a very long discussion, but it is hard to imagine a more instructive lesson in how and why one must carry out probability calculations where infinite sets are involved, or a more horrible example of what can happen if we fail to heed the advice of Gauss.

Finite vs. Countable Additivity

At this point, the reader will be puzzled and asking, "Why should anybody care about nonconglomerabiity? What difference does it make?" Nonconglomerability is, indeed, of little interest in itself; it is only a kind of red herring that conceals the real issue. A follower of de Finetti would say that the underlying issue is the technical one of finite additivity. To which we would reply that 'finite additivity' is also a red herring, because it is used for a purpose almost the opposite of what it sounds like.

In Chapter 2 we derived the sum rule (2–64) for mutually exclusive propositions: if as a statement of Boolean algebra, $A \equiv A_1 + A_2 + \ldots + A_n$ is a disjunction of a finite number of mutually exclusive propositions, then

$$p(A|C) = \sum_{i=1}^{n} p(A_i|C)$$

Then it is a trivial remark that our probabilities have "finite additivity". As $n \to \infty$ it seems rather innocuous to suppose that the sum rule goes in the limit into a sum over a countable number of terms, forming a convergent series; whereupon our probabilities would be called countably additive. Indeed (although we do not see how it could happen in a real problem), if this should ever fail to yield a convergent series we would conclude that the infinite limit does not make sense, and we would refuse to pass to the limit at all. In our formulation of probability theory, it is difficult to see how one could make any substantive issue out of this perfectly straightforward situation.

However, the conventional formulations, reversing our policy, suppose the infinite limit already accomplished at the beginning, before such questions as additivity are raised; and then are concerned with additivity over propositions about intervals on infinite sets. To quote Feller (1971, p. 107):

> Let F be a function assigning to each interval I a finite value $F\{I\}$. Such a function is called (finitely) *additive* if for every partition of an interval I into finitely many non-overlapping intervals $I_1, \ldots, I_n$, $\quad F\{I\} = F\{I_1\} + \ldots + F\{I_n\}$.

† This cancellation is the thing that is not evident at all in the exact solution (15–24), although it is still taking place out of sight.

Then (p. 108) he gives an example showing why he wishes to replace finite additivity by countable additivity:

> In R^1 put $F\{I\} = 0$ for any interval $I = (a, b)$ with $b < \infty$ and $F\{I\} = 1$ when $I = (a, \infty)$. This interval function is additive but weird because it violates the natural continuity requirement that $F\{(a, b)\}$ should tend to $F\{(a, \infty)\}$ as $b \to \infty$.
>
> This last example shows the desirability of strengthening the requirement of finite additivity. We shall say that *an interval function F is countably additive, or σ-additive, if for every partitioning of an interval I into countably many intervals $I_1, I_n, \ldots,$ $F\{I\} = \Sigma F\{I_k\}$.*

Then he adds that the condition of countable additivity is "manifestly violated" in the above weird example (let it be an exercise for the reader to explain clearly *why* this is manifest).

What is happening in that weird example? Surely, the weirdness does not lie in lack of continuity (since continuity is quite unnecessary in any event), but in something far worse. Supposing those intervals occupied by some variable x and the interval function $F\{I\}$ to be the probability $p(x \epsilon I)$, one is assigning zero probability to any finite range of x, but unit probability to the infinite range. This is almost impossible to comprehend when we suppose the infinite interval already accomplished, but we can understand what is happening if we heed the advice of Gauss and think in terms of passage to a limit. Suppose we have a properly normalized *pdf*:

$$p(x|r) = \left\{ \begin{array}{ll} 1/r, & 0 \le x < r \\ 0 & r \le x < \infty \end{array} \right\} \tag{15–35}$$

As long as $0 < r < \infty$ there is nothing strange, and we could describe this by an interval function

$$F(a, b) \equiv \int_a^b p(x|r)\, dx = \left\{ \begin{array}{ll} (b-a)/r, & 0 \le a \le b \le r < \infty \\ (r-a)/r, & 0 \le a \le r \le b < \infty \\ 0, & 0 \le r \le a \le b < \infty \end{array} \right\} \tag{15–36}$$

which is, rather trivially, countably additive and *a fortiori* finitely additive. As r increases, the density function becomes smaller and spread over a wider interval; but as long as $r < \infty$ we have a well–defined and non–paradoxical mathematical situation.

But if we try to describe the limit of $p(x|r)$ as something already accomplished *before* discussing additivity, then we have created Feller's weird example. We are trying to make a probability density that is everywhere zero, but which integrates to unity. But *there is no such thing*, according not only to all the warnings of classical mathematicians from Gauss on, but according to our own elementary common sense.

Invoking finite additivity is a sneaky way of approaching the real issue. To see why the kind of additivity matters in the conventional formulation, let us note what happens when one carries out the order of operations corresponding to our advice above. We assign a continuous monotonic increasing cumulative probability function $G(x)$ on the real line, with the natural continuity property that

$$G(x) \to \left\{ \begin{array}{ll} 1, & x \to +\infty \\ 0, & x \to -\infty \end{array} \right\} \tag{15–37}$$

then the interval function F for the interval $I = (a, b)$ may be taken as $F\{I\} = G(b) - G(a)$, and it is 'manifest' that this interval function is countably additive in the sense defined. That is, we can choose x_k satisfying $a < x_1 < x_2 < \ldots < b$ so as to break the interval (a, b) into as many nonoverlapping subintervals $\{I_0, I_1, \ldots, I_n\} = \{(a, x_1), (x_1, x_2), \ldots (x_n, b)\}$ as we please, and it will be true that $F\{I\} = \sum F\{I_k\}$. If $G(x)$ is differentiable, then its derivative $f(x) \equiv G'(x)$ may be interpreted as a normalized probability density: $\int f(x)\, dx = 1$.

We see, finally, what the point of all this is: "finite additivity" is a euphemism for "reversing the proper order of approaching limits, and thereby getting into trouble with non–normalizable probability distributions". Feller saw this instantly, warned the reader against it, and proceeded to develop his own theory in a way that avoids the many useless and unnecessary paradoxes that arise from it.†

As we saw in Chapter 6, passage to the limit $r \to \infty$ at the end of a calculation can yield useful results; some other probability derived from $p(x|r)$ might approach a definite, finite, and simple limiting value. We have now seen that trying to pass to the limit at the beginning of a calculation can generate nonsense because crucial information is lost before we have a chance to use it.

The real issue here is: do we admit such things as uniform probability distributions on infinite sets into probability theory as legitimate mathematical objects? Do we believe that an infinite number of zeroes can add up to one? In the strange language in which these things are discussed, to advocate 'finite additivity' as de Finetti and his followers do, is a devious way of answering 'yes' without seeming to do so. To advocate 'countable additivity' as Kolmogorov and Feller did, is an equally devious way to answer 'no' in the spirit of Gauss.

The terms are red herrings because 'finite additivity' sounds colloquially as if were a cautious assumption, 'countable additivity' a bit more adventurous. de Finetti does indeed seem to think that finite additivity is the weaker assumption; and he rails against those who, as he sees it, are intellectually dishonest when they invoke countable additivity only for "mathematical convenience", instead of for a compelling reason. As we see it, jumping directly into an infinite set at the very beginning of a problem is a vastly greater error of judgment, which has far worse consequences for probability theory; there is a little more than just 'mathematical convenience' at stake here.

We noted the same psychological phenomenon in Chapter 3, when we introduced the binomial distribution for sampling with replacement; those who commit the sin of throwing away relevant information, invented the term 'randomization' to conceal that fact and make it sound like they were doing something respectable. Those who commit the sin of doing reckless, irresponsible things with infinity often invoke the term 'finite additivity' to make it sound as if they are being *more* careful than others with their mathematics.

The Borel–Kolmogorov Paradox

For the most part, the transition from discrete to continuous probabilities is uneventful, proceeding in the obvious way with no surprises. However, there is one tricky point concerning continuous densities that is not at all obvious, but can lead to erroneous calculations unless we understand it. The following example continues to trap many unwary minds.

Suppose I is prior information according to which $(x,\ y)$ are assigned a bivariate normal *pdf* with variance unity and correlation coefficient ρ:

$$p(dx\, dy|I) = \frac{\sqrt{1-\rho^2}}{2\pi}\ \exp\left[-\frac{1}{2}(x^2 + y^2 - 2\rho x y)\right]\ dx\, dy \qquad (15\text{–}38)$$

† Since we disagree with Feller so often on conceptual issues, we are glad to be able to agree with him on nearly all technical ones. He was, after all, a very great contributor to the technical means for solving sampling theory problems, and practically everything he did is useful to us in our wider endeavors.

We can integrate out either x or y to obtain the marginal *pdf*'s [to prepare for integrating out x, write $x^2 + y^2 - 2\rho xy = (x - \rho y)^2 + (1 - \rho^2)\, y^2$, *etc.*]:

$$p(dx|I) = \left(\frac{1-\rho^2}{2\pi}\right)^{1/2} \exp\left[-\frac{1}{2}(1-\rho^2)\, x^2\right] dx \tag{15–39}$$

$$p(dy|I) = \left(\frac{1-\rho^2}{2\pi}\right)^{1/2} \exp\left[-\frac{1}{2}(1-\rho^2)\, y^2\right] dy \tag{15–40}$$

Thus far, all is routine. But now, what is the conditional *pdf* for x, given that $y = y_0$? We might think that we need only set $y = y_0$ in (15–38) and renormalize:

$$p(dx|y = y_0, I) = A \exp\left[-\frac{1}{2}(x^2 + y_0^2 - 2\rho x y_0)\right] dx \tag{15–41}$$

where A is a normalizing constant. But there is no guarantee that this is valid, because we have obtained (15–41) by an intuitive *ad hoc* device; we did not derive it from (15–38) by applying the basic rules of probability theory, which we derived in Chapter 2 for the discrete case:

$$p(AB|X) = p(A|BX)\, p(B|X) \tag{15–42}$$

from which a discrete conditional probability is given by the usual rule

$$p(A|BX) = \frac{p(AB|X)}{p(B|X)} \tag{15–43}$$

often taken as the definition of a conditional probability. But we can do the calculation by strict application of our rules if we define the discrete propositions:

$$A \equiv \text{“}x \text{ in } dx\text{”}$$
$$B \equiv \text{“}y \text{ in } (y_0 < y < y_0 + dy)\text{”}$$

Then we should write instead of (15–41), using (15–38) and (15–40),

$$p(A|BI) = p(dx|dy\, I) = \frac{p(dx\, dy|I)}{p(dy|I)} = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x - \rho y_0)^2\right] dx \tag{15–44}$$

Since dy cancels out, taking the limit $dy \to 0$ does nothing.

Now on working out the normalizing constant in (15–41) we find that (15–41) and (15–44) are in fact identical. So, why all this agony? Didn't the quick argument leading to (15–41) give us the right answer?

This is a good example of our opening remarks that a fallacious argument may lead to correct or incorrect results. The reasoning that led us to (15–41) happened to give a correct result here; but it can equally well yield any result we please instead of (15–41). It depends on the particular form in which you or I choose to write our equations. To show this, and therefore generate a paradox, suppose that we had used instead of $(x,\ y)$ the variables $(x,\ u)$, where

$$u \equiv \frac{y}{f(x)} \tag{15–45}$$

with $0 < f(x) < \infty$ [for example, $f(x) = 1 + x^2$ or $f(x) = \cosh x$, *etc.*]. The Jacobian is

$$\frac{\partial(x,u)}{\partial(x,y)} = \left(\frac{\partial u}{\partial y}\right)_x = \frac{1}{f(x)} \tag{15–46}$$

so the *pdf* (15–38), expressed in the new variables, is

$$p(dx\,du|I) = \frac{\sqrt{1-\rho^2}}{2\pi}\ \exp\left[-\frac{1}{2}\left(x^2 + u^2 f^2(x) - 2\rho u f(x)\right)\right]\ f(x)dx\,du\,. \tag{15–47}$$

Again, we can integrate out u or x, leading to a marginal distribution $p(dx|I)$ which is easily seen to be identical with (15–39); and $p(du|I)$ which is found to be identical with (15–40) transformed to the variable u, as it should be; so far, so good.

But now, what is the conditional *pdf* for x, given that $u = 0$? If we follow the reasoning that led us to (15–41) [*i.e.*, simply set $u = 0$ in (15–47) and renormalize], we find

$$p(dx|\,u = 0,\, I) = A\ \exp\left(-\frac{1}{2}\,x^2\right)\ f(x)dx \tag{15–48}$$

Now from (15–45) the condition $u = 0$ is the same as $y = 0$; so it appears that this should be the same as (15–41) with $y_0 = 0$. But (15–48) differs from that by an extra factor $f(x)$ which could be arbitrary!

Many find this astonishing and unbelievable; they repeat over and over: "But the condition $u = 0$ is *exactly the same condition* as $y = 0$; how can there be a different result?" We warned against this phenomenon briefly, and perhaps too cryptically, in Chapter 4; but there it did not actually cause error because we had only one parameter in the problem. Now we need to examine it carefully to see the error and the solution.

We noted already in Chapter 1 that we shall make no attempt to define any probability conditional on contradictory premises; there could be no unique solution to such a problem. We start each problem by defining a 'sample space' or 'hypothesis space' which sets forth the range of conditions we shall consider *in that problem.* In the present problem our discrete hypotheses were of the form '$a \le y \le b$', placing y in an interval of positive measure $b - a$. Then what could we mean by the proposition "$y = 0$", which has measure zero? We could mean only the limit of some sequence of propositions referring to positive measure, such as

$$A_\epsilon \equiv \text{“}|y| < \epsilon\text{”}$$

as $\epsilon \to 0$. The propositions A_ϵ confine the point (x, y) to successively narrower horizontal strips, but for any $\epsilon > 0$, A_ϵ is a discrete proposition with a definite positive probability, so by the product rule the conditional probability of any hypothesis $H \equiv$ "$x \in dx$",

$$p(H|A_\epsilon\, I) = \frac{p(H, A_\epsilon|I)}{p(A_\epsilon|I)} \tag{15–49}$$

is well–defined, and the limit of this as $\epsilon \to 0$ is also a well–defined quantity. Perhaps that limit is what one meant by $p(H|y = 0, I)$.†

† Note again what we belabor constantly: the rules of probability theory tell us unambiguously that it is the limit of the ratio, not the ratio of the limits, that is to be taken in (15–49). The former quantity remains finite and well–behaved in conditions where the latter does not exist.

But the proposition "$y = 0$" may be defined equally well as the limit of the sequence

$$B_\epsilon \equiv \text{“}|y| < \epsilon |x|\text{”}$$

of successively thinner wedges, and $p(H|B_\epsilon I)$ is also unambiguously defined as in (15–49) for all $\epsilon > 0$. Yet although the sequences $\{A_\epsilon\}$, $\{B_\epsilon\}$ tend to the same limit $y = 0$, the conditional densities tend to different limits:

$$\begin{aligned} &\lim p(H|A_\epsilon) \propto g(x), \\ &\lim p(H|B_\epsilon) \propto |x|\, g(x) \end{aligned} \qquad (15\text{–}50)$$

and in place of $|x|$ we could put an arbitrary non–negative function $f(x)$. As we see from this, merely to specify "$y = 0$" without any qualifications is ambiguous; it tells us to pass to a measure–zero limit, but does not tell us which of any number of limits is intended.

We have here one more example showing why the rules of inference derived in Chapter 2 must be obeyed *strictly, in every detail*. Intuitive shortcuts have a potential for disaster, which is particularly dangerous just because of the fact that it strikes only intermittently. An intuitive *ad hockery* that violates those rules will probably lead to a correct result in some cases; but it will surely lead to disaster in others. Whenever we have a probability density on one space and we wish to generate from it one on a subspace of measure zero, the only safe procedure is to pass to an explicitly defined limit by a process like (15–49). In general, the final result will and must depend on which limiting operation was specified. This is extremely counter–intuitive at first hearing; yet it becomes obvious when the reason for it is understood.

A famous puzzle based on this paradox concerns passing from the surface of a sphere to a great circle on it. Given a uniform probability density over the surface area, what is the corresponding conditional density on any great circle? Intuitively, everyone says immediately that, from geometrical symmetry, it must be uniform also. But if we specify points by latitude $(-\pi/2 \le \theta \le \pi/2)$ and longitude $(-\pi < \phi \le \pi)$, we do not seem to get this result. If that great circle is the equator, defined by $(|\theta| < \epsilon,\ \epsilon \to 0)$, we have the expected uniform distribution $[p(\phi) = (2\pi)^{-1},\ -\pi < \phi \le \pi]$; but if it is the meridian of Greenwich defined by $(|\phi| < \epsilon,\ \epsilon \to 0)$, we have $[p(\theta) = (1/2)\cos\theta,\ -\pi/2 \le \theta \le \pi/2]$ with density reaching a maximum on the equator and zero at the poles.

Many quite futile arguments have raged – between otherwise competent probabilists – over which of these results is 'correct'. The writer has witnessed this more than once at professional meetings of scientists and statisticians. Nearly everybody feels that he knows perfectly well what a great circle is; so it is difficult to get people to see that the term 'great circle' is ambiguous until we specify what limiting operation is to produce it. The intuitive symmetry argument presupposes unconsciously the equatorial limit; yet one eating slices of an orange might presuppose the other.

The Marginalization Paradox

The 'Strong Inconsistency' problem (Stone, 1970) flared up into an even more spectacular case of probability theory gone crazy, with the work of Dawid, Stone, & Zidek (1973), hereafter denoted by DSZ, which for a time seemed to threaten the consistency of all probability theory. The marginalization paradox is more complicated than the ones discussed above, because it arises not from a single error, but from a combination of errors of logic and intuition, insidious because they happened to support each other. When first propounded it seems to have fooled every expert in the field, with the single exception of D. A. S. Fraser, who as discussant of the DSZ paper saw that the conclusions were erroneous and put his finger correctly on the cause of this; but was not listened to.

The marginalization paradox also differs from the others in that it has done far more damage to the cause of Scientific Inference than any other; yet when properly understood, the phenomenon has useful applications in Scientific Inference. We have seen the role of marginalization in clarifying and extending the general theory of sufficiency in Chapter 8, but there developed without contradictory assumptions and therefore without any appearance of paradox. Marginalization as a potentially useful means of constructing uninformative priors is discussed later; this rather deep subject still has the status of ongoing research, in which the main theorems are probably not yet known.

In the present Chapter we are concerned with the marginalization story only as a weird incident of history which forced us to revise some easy, shortcut ways of dealing with infinite sets and limits. Its resolution requires us to recognize not only technical errors (failure to obey the rules of probability theory), but also some conceptual errors that Harold Jeffreys had tried repeatedly to point out long before – and again was not listened to.

We illustrate the original paradox by the scenario of DSZ, again following their notation until we see why we must not. It starts as a conventional, and seemingly harmless, nuisance parameter problem; a conscientious Bayesian B_1 studies a problem with parameters θ which he partitions into two components, $\theta = (\eta, \zeta)$, being interested only in inferences about ζ. Thus η is a nuisance parameter to be integrated out of the problem, and with a prior $\pi(\eta, \zeta)$ and data x, B_1 thus obtains the marginal posterior *pdf* for ζ:

$$p(\zeta|x) = \int p(\eta, \zeta|x)\, d\eta = \frac{\int p(x|\eta, \zeta)\, \pi(\eta, \zeta)\, d\eta}{\int\int p(x|\eta, \zeta)\, \pi(\eta, \zeta)\, d\eta\, d\zeta}\,, \qquad (15\text{–}51)$$

the standard result. The issue now turns on what class of priors $\pi(\eta, \zeta)$ we may assign for this purpose. Our answer is, of course:

> "Any proper prior, any prior for which the integrals in (15–51) converge, or any limit of a sequence of such priors such that the ratio of integrals in (15–51) converges to yield a proper *pdf* for ζ, may be admitted into our theory as representing a conceivable state of prior knowledge about the parameters. Eq. (15–51) will then yield the correct conclusions that follow from that state of knowledge."

This need not be qualified by any special circumstances of the particular problem; we believe that this policy, followed strictly, cannot generate ambiguities or contradictions. But *failure* to follow it can lead to almost anything.

However, DSZ did not see it that way at all. They concentrate on a special circumstance, noting that in many cases the data x may be partitioned into two components: $x = (y, z)$ in such a way that the sampling distribution for z is independent of the nuisance parameter η:

$$p(z|\eta, \zeta) = p(z|\zeta)\,. \qquad (15\text{–}52)$$

which, by itself, would appear rather generally possible and in many different ways, but without any particular significance. For example, if η is a location parameter, then any function $z(x)$ of the data that is invariant under rigid translations will have a sampling distribution independent of η. If η is a scale parameter, then any function $z(x)$ invariant under scale changes will have this property. If η is a rotation angle, then any component of the data that is invariant under those rotations will qualify.

DSZ proceed to discover cases in which, when (15–52) holds and B_1 assigns an improper prior to η, he finds that his marginal posterior *pdf* for ζ "is a function of z only", which property they write as

$$p(\zeta|y, z) = p(\zeta|z)\,. \qquad (15\text{–}53)$$

and the combination of (15–52) and (15–53) sets off a curious train of thought.

At this point there enters a lazy Bayesian B_2, who "always arrives late on the scene of inference" and proposes to simplify the calculation. The appearance of (15–52) suggests that η might be removed from the model as irrelevant; and from (15–53) as written it appears that the component y of the data can also be discarded as irrelevant to inferences about ζ. So his intuitive judgment is that, given (15–52) and (15–53), we should be able to derive the marginal *pdf* for ζ more easily by direct application of Bayes' theorem in a reduced model $p(z|\zeta)$ in which (y,η) do not appear at all. Thus if B_2 assigns the prior $\pi(\zeta)$, he obtains the posterior distribution

$$p(\zeta|z) = \frac{p(z|\zeta)\,\pi(\zeta)}{\int p(z|\zeta)\,\pi(\zeta)\,d\zeta}\,. \tag{15–54}$$

But he finds to his dismay that he cannot reproduce B_1's result (15–51) whatever prior he assigns to ζ. What conclusions should we draw from this?

For DSZ, the reasoning of B_2 seemed compelling; on grounds of this intuitive 'reduction principle' they considered it obvious that B_1 and B_2 ought to get the same results, and therefore that one of them must be guilty of some transgression. They point the accusing finger at B_1 thus: "B_2's intervention has revealed the paradoxical unBayesianity of B_1's posterior distribution for ζ". They place the blame on his use of an improper prior for η.

For us, the situation appears very different; B_2's result was not derived by application of our rules. The step from (15–51) to (15–54) was only an intuitive guess, and as the reader may verify, it does not follow mathematically from (15–52) and (15–53). Therefore (15–54) *is not a valid application of probability theory to* B_1's *problem.* If intuition suggests otherwise, then that intuition needs educating – just as it did in the other paradoxes.

The crucial point that is concealed by the notation of (15–53) is that while the result is "a function of z only" in the sense that y does not appear explicitly, it is a *different* function of z for different η–priors. That is, it is still a functional of the η–prior, as is clear from a glance at (15–51); prior information about η still matters. In other words, B_2 comes to a different conclusion than B_1 not because B_1 is committing a transgression, but for just the opposite reason: B_1 is taking into account cogent prior information about η that B_2 is ignoring.

There is a better way of looking at this; we may take a more charitable view toward B_2 if we reinterpret this scenario. He appeared first to be a lazy fellow who invents an easy short–cut method that violates the rules of probability theory. But we may suppose equally well that, through no fault of his own, he is only an uninformed fellow who was given only the reduced model $p(z|\zeta)$ and is unaware of the existence of (η, y). Then (15–54) is a valid inference for the *different state of knowledge* that B_2 has. Viewed in this way, B_1 and B_2 are both making valid inferences, but about two different problems. The above notation conceals this fact by failing to indicate their different prior information I_1, I_2 explicitly.

Exercise 15.2. Consider the intermediate case of a third Bayesian B_3, who has the same prior information as B_1 about η but is not given the data component y. Then y does not appear in B_3's equations at all. Show that, nevertheless, if (15–52) still holds, then B_2 and B_3 are always in agreement; $p(\zeta|z, I_3) = p(\zeta|z, I_2)$. Withholding the datum y automatically makes any prior knowledge about η irrelevant. Ponder this until you can explain why it is, after all, intuitively obvious.

If we follow the example of Jeffreys and write the left–hand sides of (15–51) and (15–54) correctly as $p(\zeta|y, z, I_1)$ and $p(\zeta|z, I_2)$, thereby making this difference in the problems clear, there can be no appearance of paradox. For, in inference our conclusions always depend necessarily on both the prior information and the data; that B_1 and B_2 came to different conclusions from different prior information is no more strange than if they had come to different conclusions from different data.

Up to this point, we had only a misreading of equations through inadequate notation; but now a mutual reinforcement of errors commenced. In support of their contention that B_1 is the guilty party, DSZ offered a proof that this paradox (*i.e.*, the discrepancy in the results of B_1 and B_2) "could not have arisen if B_1 had employed proper prior distributions". Let us examine their proof of this. With a general joint proper prior $\pi(\eta, \zeta)$ the integrals in (15–51) are separately convergent and positive, so if we multiply through by the denominator, we are not dividing by zero. Then

$$p(x|\eta,\zeta) = p(y,z|\eta,\zeta) = p(y|z,\eta,\zeta)\,p(z|\eta,\zeta) = p(y|z,\eta,\zeta)\,p(z|\zeta) \tag{15–55}$$

where we used the product rule and (15–52). Then (15–51) becomes

$$p(\zeta|y,z)\int\int p(y|z,\eta,\zeta)\,p(z|\zeta)\,\pi(\eta,\zeta)\,d\eta d\zeta = \int p(y|z,\eta,\zeta)\,p(z|\zeta)\,\pi(\eta,\zeta)\,d\eta \tag{15–56}$$

But now we assume that (15–53) still holds; then we may [since the integrals are absolutely convergent] integrate out y from both sides of (15–56), whereupon $\int \pi(\eta,\zeta)d\eta = \pi(\zeta)$ and (15–56) reduces to

$$p(\zeta|z)\int p(z|\zeta)\pi(\zeta)\,d\zeta = p(z|\zeta)\,\pi(\zeta) \tag{15–57}$$

which is identical with (15–54). DSZ concluded that, if B_1 uses a proper prior, then B_1 and B_2 are necessarily in agreement – from which it would follow again, in agreement with their intuition, that the paradox must be caused by B_1's use of improper priors.

But this proof of (15–57) has used mutually contradictory assumptions. As Fraser recognized, if B_1 uses a proper prior, then in general (15–53) cannot be true and (15–57) does not follow; it is no accident that DSZ had found (15–53) only with improper priors. This is easiest to see in terms of a specific example, after which it will be clear that it is true in general. In the following we use the full notation of Jeffreys so that we always distinguish between the two problems.

Example 1: The Change–Point Problem. Observations have been made of n successive, independent, positive real, 'exponentially distributed' [to use orthodox jargon] quantities $\{x_1 \cdots x_n\}$. It is known (definition of the model) that the first ζ of these have expectations $1/\eta$ and the remaining $n-\zeta$ have expectation $1/(c\eta)$, where c is known and $c \neq 1$, while η and ζ are unknown. From the data we want to estimate at what point in the sequence the change occurred. The sampling density for $x \equiv (x_1 \cdots x_n)$ is

$$p(x|\eta,\zeta,I_1) = c^{n-\zeta}\,\eta^n \exp\left\{-\eta\left(\sum_{i=1}^{\zeta} x_i + c\sum_{i=\zeta+1}^{n} x_i\right)\right\}, \qquad 1 \le \zeta \le n \tag{15–58}$$

If $\zeta = n$, then there is no change, the last sum in (15–58) is absent, and c disappears from the model. Since η is a scale parameter, the sampling distribution for ratios of observations $z_i \equiv x_i/x_1$ should be independent of η. Indeed, separating the data $x = (y,z)$ into $y \equiv x_1$ which sets the scale and the ratios $(z_2 \cdots z_n)$ and noting that the volume element transforms as $dx_1 \cdots dx_n = y^{n-1} dy dz_2 \cdots dz_n$, we find that the joint sampling distribution for $\{z_2 \cdots z_n\}$ depends only on ζ:

$$p(z|\eta,\zeta,I_1) = \int p(x|\eta,\zeta,I_1)\prod_{i=2}^{n}\delta\left(\frac{x_i}{x_1} - z_i\right)\,dx_1 \cdots dx_n = \frac{c^{n-\zeta}\,(n-1)!}{Q(\zeta,z)^n} = p(z|\zeta,I_1) \tag{15–59}$$

where $z_1 \equiv 1$ and

$$Q(\zeta, z) \equiv \sum_{1}^{\zeta} z_i + c \sum_{\zeta+1}^{n} z_i \tag{15–60}$$

is a function that is known from the data. Let B_1 choose a properly normalized discrete prior $\pi(\zeta)$ in $(1 \le \zeta \le n)$, and independently a prior $\pi(\eta)\, d\eta$ in $(0 < \eta < \infty)$. Then B_1's posterior *pdf* for the change point ζ is

$$p(\zeta|y, z, I_1) \propto \pi(\zeta)\, c^{n-\zeta} \int_0^\infty \exp(-\eta y Q)\, \pi(\eta)\, \eta^n\, d\eta \tag{15–61}$$

As expected, not only does our knowledge about ζ depend on both y and z; it depends just as strongly on what prior $\pi(\eta)$ is assigned to the nuisance parameter.

On meditation we see that this result is indeed correct. For if we know absolutely nothing about η except that it is positive, then the only evidence we can have about the change point ζ must come from noting the relative values of the x_i; for example, at which i does the ratio x_i/x_1 appear to change? On the other hand, suppose that we knew η exactly; then clearly not only the ratios x_i/x_1, but also the absolute values of the x_i would be relevant to inference about ζ [because then, whether x_i is closer to $1/\eta$ or to $1/(c\eta)$ tells us something about whether $(i < \zeta)$ or $(i > \zeta)$ that the ratio x_i/x_1 does not tell us], and this extra information would enable us to make better estimates of ζ. If we had only partial prior knowledge of η, then knowledge of the absolute values of the x_i would be less helpful, but still relevant.

In general, then, $p(\zeta|y, z, I_1)$ depends properly and necessarily on both y and z and on our prior information about η. But now B_1 discovers that use of the improper prior

$$\pi(\eta) = \eta^{-k}\,, \qquad 0 < \eta < \infty \tag{15–62}$$

where k is any real number for which the integral (15–83) converges, leads to the separation property (15–53), and to the posterior *pdf*

$$p(\zeta|z, I_1) \propto \frac{\pi(\zeta)\, c^{-\zeta}}{[Q(\zeta, z)]^{n+1-k}}\,. \tag{15–63}$$

which still depends, through k, on the prior assigned to η. But B_2's posterior *pdf* for ζ is

$$p(\zeta|z, I_2) \propto \pi(\zeta)\, p(z|\zeta) = \frac{\pi(\zeta)\, c^{-\zeta}}{[Q(\zeta, z)]^{n}} \tag{15–64}$$

which takes no note of $\pi(\eta)$. We see that for no prior $\pi(\zeta)$ can B_2 agree with B_1, except when $k = 1$, in which case B_2 and B_1 find themselves in agreement after all, and with the same prior $\pi(\zeta)$. But this result is not peculiar to the change–point model; it holds quite generally, as the following Exercise shows.

Exercise 15.3. Prove that the $k = 1$ prior is always uninformative in this sense whenever η is a scale parameter for y. That is, if the sampling distribution has the functional form

$$p(y, z|\eta, \zeta) = \eta^{-1}\, h(z, \zeta; y/\eta)\,,$$

then the prior $\pi(\eta) = \eta^{-1}$ will lead B_1 and B_2 to agree.

It seems to us that this is an eminently satisfactory result without any trace of paradox. For the case $k = 1$ is just the Jeffreys prior, which we have already seen to be 'completely uninformative' about any scale parameter η, by several different criteria. Then of course, if y is irrelevant and $k = 1$ then B_1 has no extra information after all and they should, indeed, find themselves in agreement.

But again, DSZ did not see it that way at all, and persisted in their intuitive judgment that there is a serious paradox and B_1 was at fault for using an improper prior; so the story continues. DSZ proceed to exhibit many more examples in which this 'paradox' appears – invariably when an improper prior was used. The totality of all these demonstrations appeared to mount up into overwhelming evidence that to use any improper prior is to generate inconsistencies. But, in the belief that their proof of (15–57) had already dealt with it, they failed to examine what happens in those examples in the case of proper priors, and so they managed to get through a long string of examples without discovering the error in that proof.

To correct this omission, and reveal the error in (15–57) clearly, we need only to examine any of the DSZ examples, to see what happens in the case of proper priors $\pi(\eta)$. In the change–point problem, whatever this prior, B_1's result (15–61) depends on y and z through a function of the product $yQ(\zeta, z)$. Then for what functions $f(yQ)$ will the separation property (15–53) hold? Evidently, the necessary and sufficient condition for this is that y and ζ appear in separate factors: in the case where the integrals in (15–51) converge, we require the integral to have the functional form

$$\int_0^\infty \exp(-\eta y Q)\, \pi(\eta)\, \eta^n \, d\eta = f(yQ) = g(y, z)\, h(\zeta, z)\,, \qquad (15\text{–}65)$$

for then and only then will y cancel out upon normalization of $p(\zeta|y, z)$. The answer is obvious: if a function of $(\log y + \log Q(\zeta))$ has the form $(\log g(y) + \log h(\zeta))$, the only possibility is a linear function: $\log f(yQ) = a[\log y + \log Q]$ or $f(yQ) = (yQ)^a$, where $a(z, n)$ may depend on z and n. But then, noting that the Laplace transform is unique and that

$$\int_0^\infty \exp(-\eta y Q)\, \eta^{a-1}\, d\eta = \frac{(a-1)!}{(yQ)^a} \qquad (15\text{–}66)$$

we see that (15–53) cannot hold unless the prior is of the improper form $\pi(\eta) = \eta^{-k}$, $0 < \eta < \infty$.

> **Exercise 15.4.** Show that this result is also general; that is, not only in the change–point problem, but in any problem like that of Exercise 15.3 where η is a scale parameter for y, a prior of the form $\pi(\eta) = \eta^{-k}$ will lead to a factorization of the form $\int p(y, z|\eta, \zeta)\, \pi(\eta)\, d\eta = g(y, z)\, h(\zeta, z)$ for some functions g, h, whereupon (15–53) will hold. For this reason, the many later examples of DSZ are essentially repetitious; as long as η is a scale parameter, they are only making the same point over and over again.

Evidently, any value of k which makes the integral (15–65) converge will lead to a well–behaved posterior distribution for ζ; but a still wider class of values of k will do so if the improper prior is approached, as it should be, as the limit of a sequence of proper priors; for fundamentally it is again the limit of the ratio (15–51), not the ratio of the limits, that is to be calculated.

But use of a proper prior $\pi(\eta)$ necessarily means that the separation property (15–53) cannot hold. For example, choose the prior $\pi(\eta) \propto \eta^a e^{-b\eta}$. Then (15–61) becomes

$$p(\zeta|y, z, I_1) \propto \frac{\pi(\zeta) c^{-\zeta}}{(b + yQ)^{n+a+1}} \qquad (15\text{–}67)$$

and as long as the prior is proper (that is, $b > 0$), the datum y cannot be disentangled, but remains relevant; and so (15–53) does not hold; another counter–example to the DSZ proof. The 'paradox' disappears, not because B_1 and B_2 agree, but because B_2 cannot play his game at all; just what Fraser stated. From the standpoint of information content this is rather obvious; for any proper prior represents cogent information about η, determining what are for all practical purposes upper and lower bounds to the values it could reasonably have; such information cannot be irrelevant to the inference about ζ. Indeed, in any of the DSZ examples, inserting any proper prior for which we can do the integrals will yield an equally good counter–example; how could this have gone undetected? Actually, much more was unnoticed.

Discussion

Some people have denied that there is any such thing as 'complete ignorance', much less any 'completely uninformative' prior. From their introductory remarks, it appears that to demonstrate this was the original motivation of DSZ; and several discussants continue to emphasize the point in agreement with them. But their arguments were invalid, and our results confirm the sense of the idea of 'complete ignorance' after all. The Jeffreys prior is doing here what we should naturally suppose an uninformative prior ought to do; and it does this quite generally (whenever η is a scale parameter).

This result of Exercise (15.3) shows another sense in which our previous conclusion (that the prior $d\eta/\eta$ is uninformative about a scale parameter η) is quite literally true; not as an intuitive judgment, but now as a definite theorem that follows from the rules of probability theory. Many of the participants in the discussion following the DSZ paper endorsed the suggestion that a Bayesian should not try to characterize 'prior ignorance', but should instead strive to represent prior knowledge. But this is a contradiction; it is impossible to do the latter without the former. Of course, our ultimate goal is always to represent honestly the prior information that we actually have. But both conceptually and mathematically, the notion of 'complete ignorance' is a valid and necessary part of this program, as the starting point from which all inference proceeds; just as the notion of zero is a necessary part of arithmetic.

Technically, the concurrence of many different results like that of Exercise (15.3) shows us that the notion of complete ignorance is consistent and useful; the fact that the same Jeffreys prior emerges uniquely from many different and independent lines of reasoning shows how impossible it would be to modify it or abandon it. As is almost invariably the case in this field, past difficulties with the ideas of Jeffreys signified not any defects in his ideas, but only misapplications of probability theory on the part of his critics.

In the discussion following the DSZ paper, nobody noticed that there was a counter–example to their proof of (15–57) already in plain sight in the DSZ article (their Example #5, where it is evident that B_1 and B_2 remain in disagreement for all priors, proper or improper), and only Fraser expressed any doubts about the DSZ conclusions. He noted that DSZ

> " - - - propose that the confusion can be avoided by a restriction to *proper* priors. This is a strange proposal as a resolution of the difficulties—for it means in the interesting cases that one cannot eliminate a variable, and hence cannot go to the marginal likelihood."

But it seems that these words were, like the prophecies of Nostradamus, too cryptic for anyone to understand until he had first located the error for himself. Fraser's point – and ours above – is that when B_1 uses a proper prior, then in general B_2 simply ceases to exist, because (15–53) ceases to be true. But as we shall see, in some exceptional cases (15–53) remains true; and then B_1 and B_2 remain in disagreement, contradicting (15–57). In other words, when B_1 uses proper priors, this *never* brings B_1 and B_2 into agreement. In (15–65), (15–66) we have demonstrated that in the change point problem, agreement of B_1 and B_2 *requires* that B_1 uses an improper prior, just the opposite of the DSZ conclusion.

In any event, what happened was that nearly everybody accepted the DSZ conclusions uncritically, without careful examination of their argument. Anti–Bayesians, who very much wanted the DSZ conclusion to be true, seized upon it eagerly as sounding the death–knell of all Bayesianity. Under this pressure the prominent Bayesian D. V. Lindley broke down and confessed to sins of which he was not guilty, and the Royal Society bestowed a warm vote of thanks upon DSZ for this major contribution to our understanding of inference.

As a result, since 1973 a flood of articles has appeared, rejecting the use of improper priors under any and all circumstances, on the grounds that they have been proved by DSZ to generate inconsistencies. Incredibly, the fact that proper priors never 'correct' the supposed inconsistencies never came out in all this discussion. Thus the marginalization paradox became, like nonconglomerability, quite literally institutionalized in the literature of this field, and taught as truth. Scientific Inference thus suffered a setback from which it will require decades to recover.

Nobody noted that this same 'paradox' had been found and interpreted correctly long before by Harold Jeffreys (1939, §3.8) in connection with estimating the correlation coefficient ρ in a bivariate normal distribution, in which the location parameters are the uninteresting nuisance parameters. He gives two examples of B_1's result, corresponding to different prior information about the nuisance parameters, in his equations (10) and (24), their difference indicating the effect of that prior information. Then he gives B_2's result in (28), the agreement with (24) indicating that a uniform prior for the location parameters is uninformative about ρ.

This was seen again independently by Geisser & Cornfield (1963) and Geisser (1965) in connection with priors for multivariate normal distributions. They perceived that the difference between the results of B_1 and B_2 [their equations (3.10) and (3.26)] was not a paradox, because B_2's result was not a valid solution to the problem; they termed it, very properly, a "pseudoposterior distribution." DSZ refer to this work, but still place more confidence in the 'reduction principle' than in the rules of probability theory, and conclude that Geisser and Cornfield "do not appear to have noticed its significance." So things that had been understood correctly many years before, now became non–understood; thus do we make progress in this field.

In all these examples except one – that DSZ Example #5 again – an interesting phenomenon occurred. While the paradox was present for general improper priors in some infinite class C, there was always one particular improper prior in that class for which the paradox disappeared; B_1 and B_2 found themselves in agreement after all. DSZ noted this curious fact, but "do not appear to have noticed its significance." We suggest that this was by far the most important fact unearthed in all the marginalization work.

The prior which leaves B_1 and B_2 in agreement must be *completely uninformative*, firstly about η and, *a fortiori*, about ζ. This means that, far from casting doubt on the notion of complete ignorance, in the marginalization phenomena *we have for the first time a purely objective definition of complete ignorance, that springs directly out of the product and sum rules of probability theory without appeal to any extraneous notions like symmetry or entropy.*

This is, again, an eminently satisfactory result; but why does it seem not to be true in Example #5? There is still something new and important to be learned here.

An Even More Startling Example

Now we examine the DSZ Example #5, from which we can learn far more about what it means to be "uninformative". Suppose we have data $D = \{x_1, \ldots, x_n\}$ consisting of n observations from the standard normal sampling distribution $N(\mu, \sigma)$. With prior information I described by the proper prior *pdf*

$$p(d\mu d\sigma|I) = f(\mu, \sigma) d\mu d\sigma \tag{15–68}$$

we have the usual posterior *pdf* for the parameters:

$$p(d\mu d\sigma|DI) = g(\mu,\sigma)\, d\mu d\sigma \tag{15–69}$$

with

$$g(\mu,\sigma) = \frac{f(\mu,\sigma)\cdot\sigma^{-n}\exp\left\{-\frac{n}{2\sigma^2}\left[s^2+(\mu-\bar{x})^2\right]\right\}}{\int\int d\mu d\sigma f(\mu,\sigma)\cdot\sigma^{-n}\exp\left\{-\frac{n}{2\sigma^2}\left[s^2+(\mu-\bar{x})^2\right]\right\}} \tag{15–70}$$

in which, as usual, $\bar{x} \equiv n^{-1}\sum x_i$ and $s^2 \equiv n^{-1}\sum(x_i-\bar{x})^2$ are the sufficient statistics. Although we suppose the prior $f(\mu,\sigma)$ normalizable, it need not be actually normalized in (15–70), because any normalization constant appears in both numerator and denominator, and cancels out.

As long as $s^2 > 0$, the likelihood is bounded throughout the region of integration $-\infty < \mu < \infty$, $0 \le \sigma < \infty$, and therefore with a proper prior the integral in (15–70) is guaranteed to converge, leading to a proper posterior *pdf*. Furthermore, if the prior has moments of order m, k:

$$\int_{-\infty}^{\infty} d\mu \int_0^{\infty} d\sigma \mu^m \sigma^k f(\mu,\sigma) < \infty \tag{15–71}$$

the posterior distribution is guaranteed to have moments of higher order (in fact, all orders for μ and at least as high as order $k+n$ for σ). The solution is therefore very well–behaved mathematically.

But now we throw the proverbial monkey–wrench into this by declaring that we are interested only in the quantity

$$\theta \equiv \frac{\mu}{\sigma}. \tag{15–72}$$

Making the change of variables $(\mu\sigma) \to (\theta,\sigma)$, the volume element transforms as

$$d\mu d\sigma = \sigma d\theta d\sigma\,, \tag{15–73}$$

so writing $p(d\theta|DI) = h_1(\theta)\, d\theta$, the posterior *pdf* is

$$h_1(\theta) = \int_0^{\infty} g(\sigma\theta,\sigma)\,\sigma d\sigma\,, \tag{15–74}$$

and in view of the high moments of g we do not expect any convergence problems here, as long as $n > 1$. Thus far, there is no hint of trouble.

But now we examine the solution for a specific proper prior that can approach an improper prior. Consider the conjugate prior probability element

$$f(\mu,\sigma)\, d\mu\, d\sigma \propto \sigma^{-\gamma-1}\exp(-\beta/\sigma - \alpha\mu^2)\, d\mu\, d\sigma \tag{15–75}$$

which is proper when $(\alpha,\beta,\gamma) > 0$, and tends to the Jeffreys uninformative prior $d\mu d\sigma/\sigma$ as $(\alpha,\beta,\gamma) \to 0$. This leads to the joint posterior *pdf*, $p(d\mu\, d\sigma|D,I) = g(\mu,\sigma)\, d\mu d\sigma$ with density function

$$g(\mu,\sigma) \propto \sigma^{-n-\gamma-1}\cdot\exp\left\{-\frac{\beta}{\sigma} - \alpha\mu^2 - \frac{n}{2\sigma^2}\left[s^2+(\mu-\bar{x})^2\right]\right\} \tag{15–76}$$

from which we are to calculate the marginal posterior *pdf* for θ alone by the integration (15–74). The result depends on both sufficient statistics $(\bar{x}, s)$, but is most easily written in terms of a different set. The quantities R, r, where

$$R^2 \equiv n(\bar{x}^2+s^2) = \Sigma x_i^2\,, \qquad r \equiv n\bar{x}/R = \Sigma x_i \Big/ \sqrt{\Sigma x_i^2} \tag{15–77}$$

also form a set of jointly sufficient statistics, and from (15–76), (15–74), we find the functional form $p(d\theta|D,I) = h_1(\theta|r,R)\,d\theta$, where

$$h_1(\theta|r,R) \propto \exp\left(-\frac{n\theta^2}{2}\right)\int_0^\infty d\omega\,\omega^{n+\gamma-1}\,\exp(-\frac{1}{2}\omega^2 + r\theta\omega - \beta R^{-1}\omega - \alpha\theta^2 R^2\omega^{-2}) \qquad (15\text{–}78)$$

As long as α or β is positive, the result depends on both sufficient statistics, as Fraser predicted; but as α, β tend to zero and we approach an improper prior, the statistic R becomes less and less informative about θ, and when α, β both vanish the dependence on R drops out altogether:

$$h_1(\theta|r,R) \to h_1(\theta|r) \propto \exp\left(-\frac{n\theta^2}{2}\right)\int_0^\infty d\omega\,\omega^{n+\gamma-1}\,\exp(-\frac{1}{2}\omega^2 + r\theta\omega) \qquad (15\text{–}79)$$

If then one were to look only at the limiting case $\alpha = \beta = 0$ and not at the limiting process, it might appear that just r alone is a sufficient statistic for θ, as it did in (15–53). This supposition is encouraged by noting that the sampling distribution for r in turn depends only on θ, not on μ and σ separately:

$$p(r|\theta) \propto (n - r^2)^{(n-3)/2}\int_0^\infty d\omega\,\omega^{n-1}\,\exp(-\frac{1}{2}\omega^2 + r\theta\omega) \qquad (15\text{–}80)$$

It might then seem that, in view of (15–79) and (15–80), we should be able to derive the same result by applying Bayes' theorem to the reduced sampling distribution (15–80). But one who supposes this finds, to his dismay, that (15–80) is not a factor of (15–79); that is, the ratio $h_1(\theta|r)/p(r|\theta)$ depends on r as well as θ. The Jeffreys uninformative prior $\gamma = 0$ does indeed make the two integrals equal, but there remains an uncompensated factor with $(n - r^2)$, and so even the uninformative Jeffreys prior for (μ, σ) cannot bring about agreement of B_1 and B_2. There is no prior $p(\theta|I)$ that can yield the posterior distribution (15–79) from the sampling distribution (15–80).

Since the paradox is still present for a proper prior, this is another counter–example to (15–57); but it has a deeper meaning for us. What is now the information being used by B_1 but ignored by B_2? It is not the prior probability for the nuisance parameter; the new feature is that in this model the mere qualitative fact of the existence of the nuisance parameter in the model *already constitutes prior information relevant to the inference*, which B_2 is ignoring.

Discussion: Looking at the above equations with this in mind, we now see that there is no inconsistency after all; one should not have expected (15–79) to be derivable from (15–80) by Bayes' theorem because they are the posterior distribution and sampling distribution for two different problems. Equation (15–80) can indeed be interpreted as a sampling distribution; but only for a 'pseudoproblem' in which the model has different parameters. Eq. (15–79) is the correct marginal posterior *pdf* for θ in a problem P_1 with two parameters (θ, σ); but although σ is integrated out to form the marginal *pdf*, the result still depends on what prior we have assigned to σ – as it should, since if σ is known, it is highly relevant to the inference; if it is unknown, any prior information we have about it must still be relevant.

In contrast, (15–80) can be interpreted as a valid sampling distribution for a problem P_2 in which θ is the only parameter present; the prior information does not even include the existence of the parameter σ which was integrated out in P_1. With a prior density $f_2(\theta)$ it would yield a posterior *pdf*

$$h_2(\theta) \propto f_2(\theta)\int d\omega\,\omega^{n-1}\,\exp(-\frac{1}{2}\omega^2 + r\theta\omega) \qquad (15\text{–}81)$$

of a different functional form, making no reference to any prior for σ. In view of the earlier work of Jeffreys and of Geisser & Cornfield, one could hardly claim that the situation was new and startling; much less paradoxical. As was noted later (Jaynes, 1980, 1983), we had here a double confusion; improper priors were blamed for causing a paradox which they did not cause and which was not a paradox. Yet a counter–example to this conclusion was in plain sight, in the DSZ article.

Harold Jeffreys was immune from such errors because (1) he perceived that the product and sum rules of probability theory are adequate to conduct inference and they take precedence over the intuitive *ad hoc* devices that dominated orthodox statistics; (2) he had recognized from the start that all probabilities are necessarily conditional not only on the parameters or the data, but also on the prior information; therefore his formal probability symbols $P(A|BI)$ always indicated the prior information I explicitly.

Today, it seems to us incredible that anyone could have examined even one problem of inference without perceiving this necessary role of prior information; what kind of logic could they have been using? Nevertheless, those trained in the 'orthodox' tradition of probability theory did not recognize it. They did not have a term for prior information in their vocabulary, much less a symbol for it in their equations; and *a fortiori* no way of indicating when two probabilities are conditional on different prior information.† So they were helpless when prior information matters.

A Useful Result After All?

In most paradoxes there is something of value to be salvaged from the debris, and we think (Jaynes, *loc cit*) that the marginalization paradox may have made an important and useful contribution to the old problem of 'complete ignorance'. How is the notion to be defined, and how is one to construct priors expressing complete ignorance? We have discussed this from the standpoint of entropy and symmetry (transformation groups) in previous Chapters; now marginalization suggests still another principle for constructing uninformative priors.

Many cases are known, of which we have seen examples in DSZ, where a problem has a parameter of interest θ and an uninteresting nuisance parameter η. Then the marginal posterior *pdf* for θ will depend on the prior assigned to η as well as on the sufficient statistics. Now for certain particular priors $p(\eta|I)$ one of the sufficient statistics may drop out of the marginal distribution $p(\theta|D,I)$, as R did in (15–79). It is at first glance surprising that the sampling distribution for the remaining sufficient statistics may in turn depend only on θ as in (15–80).

Put differently, suppose a problem has a set of sufficient statistics (t_1, t_2) for the parameters (θ, η). Now if there is some function $r(t_1, t_2)$ whose sampling distribution depends only on θ, so that $p(r|\theta,\eta,I) = p(r|\theta,I)$, this defines a pseudoproblem with different prior information I_2, in which η is never present at all. Then there may be a prior $p(\eta|I)$ for which the posterior marginal distribution $p(\theta|D,I) = p(\theta|r,I)$ depends only on the component r of the sufficient statistic. This happened in the example studied above; but now, more may be true. It may be that for that prior on η the pseudoposterior *pdf* for θ is identical with the marginal *pdf* in the original problem. If a prior brings about agreement between the marginal posterior and the pseudoposterior distributions, how should we interpret this?

† Indeed, in the period 1930 – 1960 nearly all orthodoxians scorned Jeffreys' work and some taught their students that it is not only intellectually foolish, but also morally reprehensible – a deliberate breach of 'scientific objectivity' – to allow one's self to be influenced by prior information at all! This did relatively little damage in physical science, because scientists ignored orthodox teaching and persisted in doing, qualitatively, the Bayesian reasoning that their own common sense told them was the right thing to do. But we think it was a disaster for fields such as Econometrics and Artificial Intelligence, where adoption of the orthodox 'frequentist' view of probability had the automatic consequence that the significant problems could not even be formulated meaningfully, much less solved.

Suppose we start from the pseudoproblem. It seems that if introducing a new parameter η and using the prior $p(\eta|I)$ makes no difference; it leads to the same inferences about θ as before, then it has conveyed no *information* at all about θ. Then that prior must express 'complete ignorance' of η in a rather fundamental sense that makes no use of extraneous notions such as entropy and symmetry, but springs directly out of the principles of probability theory. Now in all cases yet found the prior $p(\eta|I)$ which does this is improper; this lends support to that conclusion because our intuitive judgment has already told us that only an improper prior can be completely uninformative on an infinite domain.

Can this observation be extended to a general technique for constructing uninformative priors? This is at present an ongoing research project rather than a finished part of probability theory, so we consider it in the later 'Advanced Applications' section of this work.

How to Mass–Produce Paradoxes

Having examined a few paradoxes, we can recognize their common feature. Fundamentally, the error was always failure to obey the product and sum rules of probability theory. Almost always, the mechanism of this was careless handling of infinite sets and limits, sometimes accompanied also by attempts to replace the rules of probability theory by intuitive *ad hoc* devices like B_2's 'reduction principle'. Indeed, paradoxes caused by careless dealing with infinite sets or limits can be mass–produced by the following simple procedure:

(1) Start from a mathematically well–defined situation, such as a finite set or a normalized probability distribution or a convergent integral, where everything is well behaved and there is no question about what is the correct solution.

(2) Pass to a limit – infinite magnitude, infinite set, zero measure, improper *pdf*, or some other kind – without specifying how the limit is approached.

(3) Ask a question whose answer depends on how the limit was approached.

This is guaranteed to produce a paradox in which a seemingly well–posed question has more than one seemingly right answer, with nothing to choose between them. The insidious thing about it is that, as long as we look only at the limit, and not the limiting process, the source of the error is concealed from view.

Thus it is not surprising that those who persist in trying to evaluate probabilities directly on infinite sets have been able to study finite additivity and nonconglomerability for decades – and write dozens of papers of impressive scholarly appearance about it. Likewise, those who persist in trying to calculate probabilities conditional on propositions of probability zero, have before them an unlimited field of opportunities for scholarly looking research and publication – without hope of any meaningful or useful results.

In our opening quotation, Gauss had a situation much like this in mind; and we think it is in the spirit of Gauss to suggest that, if the elements of our sets are to carry the attribute of probability, then infinite sets can have no definite meaning except as well–defined limits of finite sets. Whenever we find a belief that such infinite sets possess some kind of "existence" and mathematical properties in their own right, independent of any such limiting process, we can expect to see paradoxes of the above type.

Note that this does not in any way prohibit us from using infinite sets to define propositions. Thus the proposition

$$G \equiv \text{“}1 \le x \le 2\text{”}$$

invokes an uncountable set, but it is still a single discrete proposition, to which we may assign a probability $P(G|I)$ defined on a sample space of a finite number of such propositions without violating our "probabilities on finite sets" policy. We are not assigning any probability directly on an infinite set.

But then if we replace the upper limit 2 by a variable quantity z, we may (and usually do) find that this defines a well–behaved function, $f(z) \equiv P(G|z, I)$. In calculations, we are then free to make use of whatever analytic properties this function may have. Even if $f(z)$ is not an analytic function, we may be able to define other analytic functions from it, for example, by integral transforms. In this way, we are able to deal with any real application that we have been able to imagine, by continuum analytical methods, without losing the protection of Cox's theorems.

COMMENTS

Back in Chapter 2, we saw that the rules of probability theory can be derived as necessary conditions for consistency, as expressed by Cox's functional equations. The proofs applied to finite sets of propositions, but when the results of a finite set calculation can be extended to an infinite set by a mathematically well–behaved passage to a limit, we also accept that limit.

It might be thought that it would be possible, and more elegant, to generalize Cox's proofs so that they would apply directly to infinite sets; and indeed that is what the writer believed and tried to carry out for many years. However, since at least the work of Bertrand (1889), the literature has been turning up paradoxes that result from attempts to apply the rules of probability theory directly and indiscriminately on infinite sets; we have just seen some representative examples and their consequences. Since in recent years there has been a sharp increase in this paradoxing, one must take a more cautious view of infinite sets.

Our conclusion is that, at least in probability theory, an infinite set should be thought of only as the limit of a specific (*i.e.* unambiguously specified) sequence of finite sets. Likewise, an improper *pdf* has meaning only as the limit of a well–defined sequence proper *pdf*'s. The paradoxes have been found only when we tried to depart from this policy by treating an infinite limit as something already accomplished, without regard to any limiting operation. Indeed, experience to date shows that almost any attempt to depart from our recommended 'finite sets' policy has the potentiality for generating a paradox, in which two equally valid methods of reasoning lead to contradictory results.

The paradoxes studied here stand as counter–examples to any hope that we can ever work with full freedom on infinite sets. We propose as a conjecture: *All correct results in probability theory are either combinatorial theorems on finite sets, generated by the product and sum rules of probability theory, or well–behaved limits of them.*

It may be that this judgment is too harsh, and future analysis may reveal general preconditions under which it is safe to apply the rules of probability theory directly on infinite sets. However, the variety of seemingly innocuous arguments which led us to the paradoxes above teaches us that such preconditions, if they exist, must be quite subtle. For the present, therefore, we think that the only safe procedure is to follow our 'finite sets' policy at the beginning of a calculation.

Of course, there are constantly recurring problems of a standard kind (such as estimation of the mean and variance μ, σ^2 of a Gaussian distribution from a finite sample $\{x_1 \ldots x_n\}$), for which undoubtedly safe rules of conduct can be formulated once and for all. Having applied this cautious policy once we become confident, very quickly, that in future problems of this type we can think of all these quantities as continuously variable from the start. But then we should show a "higher level" form of caution and start from normalizable priors until more experience has been gained.

Finally, we may recognize a class of problems in which passage to the limit of an uninformative improper prior is always well behaved. In the Gaussian case it turns out that if $n > 1$, the likelihood function falls off so rapidly away from the peak that it is almost impossible to get into trouble. Then we can behave in a seemingly irresponsible way and still get the correct results that a valid argument would yield. For this reason, in practice our cautious policy is not a major obstacle to obtaining useful results quickly. Much of the existing literature on estimation with continuous sampling distributions can be justified in a similar way.

Unfortunately, the Borel–Kolmogorov and marginalization paradoxes turn up so seldom as to encourage overconfidence in the inexperienced. As long as one works on problems where they do not cause trouble, the psychological phenomenon: "You can't argue with success!" noted at the beginning of this Chapter, controls the situation.

Mea Culpa For many years, the present writer was caught in this error just as badly as anybody else, because Bayesian calculations with improper priors continued to give just the reasonable and clearly correct results that common sense demanded; so warnings about improper priors went unheeded. But this was only because the problems considered had highly convergent likelihood functions $L(\alpha)$.Then in such a calculation as making an interval estimate of a parameter from the posterior moments

$$\langle \alpha^n \rangle = E(\alpha^n | DI) = \frac{\int d\alpha\, \alpha^n\, p(\alpha|I)\, L(\alpha)}{\int d\alpha\, p(\alpha|I)\, L(\alpha)} \tag{15–82}$$

the correct calculation would be to insert a proper prior $p(\alpha|I)$ in numerator and denominator (which guarantees convergence of both integrals whenever the likelihood function is bounded). Indeed, any problem with a proper prior and bounded likelihood function is guaranteed to yield a proper posterior *pdf*, and therefore the posterior median and quartile estimates of α will always exist. Then if one wishes to approach an improper prior, an easy shortcut mode of reasoning would be to insert the improper prior directly into (15–82). If the likelihood $L(\alpha)$ falls off rapidly enough, the integrals in numerator and denominator still converge separately, leading to a well–behaved result in spite of the faulty procedure.

Finally, it was the marginalization paradox that forced recognition that we had only been lucky in our choice of problems. If we wish to approach an improper prior, the correct calculation must take not the ratio of the limits, which do not in general exist, but the limit of the ratio, which does exist in all cases that are sufficiently well–posed for probability theory to be usable. That is, if the correct limiting procedure should yield an improper posterior *pdf*, then probability theory is telling us that the prior information and data are too vague to determine any estimates of α. Then the only remedy is to seek more data or more prior information.

Generally, the posterior *pdf* is better behaved than the prior because of the extra information in the likelihood function, and the correct limiting procedure yields a useful posterior *pdf* that is analytically simpler than any from a proper prior. The most universally useful results of Bayesian analysis obtained in the past are of this type, because they tended to be rather simple problems, in which the data were indeed so much more informative than the prior information that an improper prior gave a reasonable approximation to the strictly correct solution. In the future, however, we cannot expect this to continue because the field is turning to more complex problems in which the prior information is essential, and it would be quite wrong to think of passing to an improper prior. Further discussion of these issues, and mathematical pathology in general, is continued in Appendix B.