MONKEYS, KANGAROOS, AND N†

E. T. Jaynes‡
St. John's College and Cavendish Laboratory
Cambridge CB2 1TP, England

Abstract: We examine some points of the rationale underlying the choice of priors for MAXENT image reconstruction. The original combinatorial (monkey) and exchangeability (kangaroo) approaches each contains important truth. Yet each also represents in a sense an extreme position which ignores the truth in the other. The models of W. E. Johnson, I. J. Good, and S. Zabell provide a continuous interpolation between them, in which the monkeys' entropy factor is always present in the prior, but becomes increasingly levelled out and disappears in the limit.

However, it appears that the class of interpolated priors is still too narrow. A fully satisfactory prior for image reconstruction, which expresses all our prior information, needs to be able to express the common–sense judgment that correlations vary with the distance between pixels. To do this, we must go outside the class of exchangeable priors, perhaps into an altogether deeper hypothesis space.

CONTENTS

---

† Presented at the Fourth Annual Workshop on Bayesian/Maximum Entropy Methods, University of Calgary, August 1984. In the Proceedings Volume, *Maximum Entropy and Bayesian Methods in Applied Statistics*, James H. Justice, Editor, Cambridge University Press (1986). The present version was revised and corrected, to clarify some arguments and include new understanding, in June 1994.

‡ Visiting Fellow, 1983–1984. Permanent Address: Dept. of Physics #1105, Washington University, St. Louis MO 63130, U.S.A.

INTRODUCTION

Image reconstruction is an excellent ground for illustrating the generalities in our Tutorial Introduction. Pedagogically, it is an instructive and nontrivial example of the open–ended problem of determining priors which reprepresent real states of knowledge in the real world. In addition, better understanding of this truly deep problem should lead in the future to better reconstructions and perhaps improvements in results for other problems of interest at this Workshop.

In discussing the various priors that might be used for image reconstruction, it should be emphasized that we are not dealing with an ideological problem, but a technical one. We should not think that any choice of prior hypothesis space and measure on that space is in itself either right or wrong. Presumably, any choice will be "right" in some circumstances, "wrong" in others. It is failure to relate the choices to the circumstances that gives the appearance of arbitrariness.

In a new problem, it is inevitable that different people have in the back of their minds different underlying hypothesis spaces, for several reasons:

(1) Different prior knowledge of the phenomenon.
(2) Different amounts or kinds of practical experience.
(3) Some have thought more deeply than others.
(4) Past training sets their minds in different channels.
(5) Psychological quirks that can't be accounted for.

Therefore, rather than taking a partisan stand for one choice against another, we want to make a start on better relating the choices to the circumstances.

This means that we must learn to define the problem much more carefully than in the past. If you examine the literature with this in mind, I think you will find that 90% of the past confusions and controversies in statistics have been caused, not by mathematical errors or even ideological differences; but by the technical difficulty that the two parties had different problems in mind, and failed to realize this. Thinking along different lines, each failed to perceive at all what the other considered too obvious to mention. If you fail to specify your sample space, sampling distribution, prior hypothesis space, and prior information, you may expect to be misunderstood – as I have learned the hard way.

We are still caught up to some degree in the bad habits of orthodox statistics, taught almost exclusively for decades. For example, denoting the unknown true scene by $\{p(i), 1 \leq i \leq n\}$, we specify the mock data

$$M_k = \sum_i A(k,i)\, p(i)\,, \qquad 1 \leq k \leq m \tag{1}$$

confidently, as if the point–spread function $A(k,i)$ were known exactly, and pretend it is a known, "objectively real" fact that the measurement errors were "independent gaussian" with known standard deviation.

But we say nothing about the prior information we propose to use – not even the underlying hypothesis space on which the prior probabilities are to exist. Then in applying Bayes' theorem with I = prior information:

$$p(\mathrm{Scene}|\mathrm{Data}, I) = p(\mathrm{Scene}|I)\, \frac{p(\mathrm{Data}|\mathrm{Scene}, I)}{p(\mathrm{Data}|I)} \tag{2}$$

the likelihood of a scene,

$$p(\text{Data}|\text{Scene}, I) = \exp\left[-\frac{1}{2\sigma^2}\sum_k (d_k - M_k)^2\right] = \exp(-\chi^2/2) \qquad (3)$$

has been fully specified in the statement of the problem; while its prior probability $p(\text{Scene}|I)$ is left unspecified by failure to complete the statement of the problem.

In effect, we are claiming more knowledge than we really have for the likelihood, and less than we really have for the prior; just the error that orthodox statistics has always made. This makes it easy to say, "The data come first" and dismiss $p(\text{Scene}|I)$ by declaring it to be completely uninformative. Yet in generalized inverse problems we usually have prior information that is fully as cogent as the data.

We need a more balanced treatment. A major point of Bayesian analysis is that it combines the evidence of the data with the evidence of the prior information. Unless we use an informative prior probability, Bayes' theorem can add nothing to the evidence of the data, and its advantage over sampling theory methods lies only in its ability to deal with technical problems like nuisance parameters.

To repeat the platitudes: in image reconstruction the data alone, whether noisy or not, cannot point to any particular scene because the domain R of maximum likelihood, where $\chi^2 = 0$, is not a point but a manifold of high dimensionality, every point of which is in the "feasible set" R' (which we may think of as R enlarged by adding all points at which χ^2 is less than some specified value). An uninformative prior leaves us, inevitably, no wiser. So if entropy is denied a role in the prior probability, it must then be invoked in the end as a value judgment in addition to Bayes' theorem, to pick out one point in R'. This does not necessarily lead to a difference in the actual algorithm, for it is well known that in decision theory the optimal decision depends only on the product of the prior probability and the utility function, not on the functions separately. But it does leave the question of rationale rather up in the air.

We want, then, to reexamine the problem to see whether some of that deeper analysis might have helped us; however, the following could hardly be called an analysis in depth. For lack of time and space we can indicate only how big the problem is, and note a few places where more theoretical work is needed. This is, in turn, only one facet of the general program to develop that neglected half of probability theory. We are not about to run out of jobs needing to be done.

MONKEYS

In the pioneering work of Gull and Daniell (1978) the prior probability of a scene (map of the sky) with n pixels of equal area and N_i units of intensity in the i'th pixel, was taken proportional to its multiplicity:

$$p(\text{Scene}|I_0) \propto W = \frac{N!}{N_1! \cdots N_n!} \qquad (4)$$

One could visualize this by imagining the proverbial team of monkeys making test maps by strewing white dots at random, N_i being the number that happen to land in the i'th pixel. If the resulting map disagrees with the data it is rejected and the monkeys try again. Whenever they succeed in making a map that agrees with the data, it is saved. Clearly, the map most likely to result is the one that has maximum multiplicity W, or equally well maximum entropy per dot, $H = (\log W)/N$, while agreeing with the data. If the N_i are large, then as we have all noted countless times, H goes asymptotically into the "Shannon entropy":

$$H \to -\sum_i (N_i/N)\log(N_i/N) \qquad (5)$$

and by the entropy concentration theorem (Jaynes, 1982) we expect that virtually all the feasible scenes generated by the monkeys will be close to the one of maximum entropy.

Mathematically, this is just the combinatorial argument by which Boltzmann (1877) found his most probable distribution of molecules in a force field. But in Boltzmann's problem, $N = \sum_i N_i$ was the total number of molecules in the system, a determinate quantity.

In the image reconstruction problem, definition of the monkey hypothesis space stopped short of specifying enough about the strewing process to determine N. As long as the data were considered noiseless this did no harm, for then the boundary of the feasible set, or class C of logically possible scenes, was sharply defined (the likelihood was rectangular, so $C = R = R'$) and Bayes' theorem merely set the posterior probability of every scene outside C equal to zero, leaving the entire decision within C to the entropy factor. The value of N did not matter for the actual reconstruction.

But if we try to take into account the fact that real data are contaminated with noise, while using the same "monkey hypothesis space" H1 with n^N elements, the scene of greatest posterior probability is not the "pure MAXENT" one that maximizes H subject to hard constraints from the data; it maximizes the sum $(NH + \log L)$, where $L(\text{Scene}) = p(\text{Data}|\text{Scene}, I)$, the likelihood that allows for noise, is no longer rectangular but might, for example be given by (3). Then N matters, for it determines the relative weighting of the prior probability and the noise factors.

If L is nonzero for all scenes and we allow N to become arbitrarily large, the entropy factor $\exp(NH)$ will overwhelm the likelihood L and force the reconstruction to the uniform grey scene that ignores the data. So if we are to retain the hypothesis space H1, we must either introduce some cutoff in L that places an upper limit on the possible noise magnitude; or assign some definite finite value of N. Present practice – or some of it – chooses the former alternative by placing an upper limit on the allowable value of χ^2. Although this leads, as we all know, to very impressive results, it is clearly an *ad hoc* device, not a true Bayesian solution. Therefore we ought to be able to do still better – how much better, we do not know.

Of course, having found a solution by this cutoff procedure, one can always find a value of N for which Bayes' theorem would have given the same solution without the cutoff. It would be interesting, for diagnostic purposes, to know what these after–the–fact N values are, particularly the ratios N/n; but we do not have this information.

In different problems of image reconstruction (optics, radio astronomy, tomography, crystallography) the true scene may be generated by Nature in quite different ways, about which we know something in advance. In some circumstances, this prior information might make the whole monkey rationale and space H1 inappropriate from the start; in others it would be clearly "right".

In a large class of intermediate cases, H1 is at least a usable starting point from which we can build up to a realistic prior. In these cases, multiplicity factors are always cogent, in the sense that they always appear as a factor in the prior probability of a scene. Further considerations may "modulate" them by additional factors.

What are we trying to express by a choice of N? There can be various answers to this. One possible interpretation is that we are specifying something about the fineness of texture that we are asking for in the reconstruction. On this view, our choice of N would express our prior information about how much fineness the data are capable of giving. We discuss only this view here, and hope to consider some others elsewhere.

A preliminary attempt to analyze the monkey picture more deeply with this in mind was made by the writer at the 1981 Laramie Workshop. If the measurement errors σ are generated in the variability of the scene itself:

$$N_i \pm \sqrt{N_i}$$

there is a seemingly natural choice of N that makes $N\sigma^2 = \text{const}$. Varying N and σ then varies only the sharpness of the peak in the posterior probability space, not its location; with more accurate measurements giving smaller σ and larger N, we do not change our reconstruction but only become more confident of its accuracy, and so display it with a finer texture.

However, it appears that in the current applications we are closer to the truth if we suppose that the errors are generated independently in the measurement apparatus. Then there is a seemingly natural choice of N that expresses our prior information about the quality of the data by making the typical smallest increments in the mock data M_k due to changes in the scene, of the same order of magnitude as the smallest increment σ that the real data could detect.

If $p_i = N_i/N$, this increment is $dM_k \simeq A/N$, where A is a typical large element of A; and $N\sigma \simeq A$. Smaller values of N will yield an unnecessarily coarse reconstruction, lacking all the density gradations that the data give evidence for; while larger values in effect ask for finer gradations than the data can justify. The reconstruction depends on σ for the intuitive reason that if, for given data, we learned that the noise level is smaller than previously thought, then some details in the data that were below the noise level and ignored, now emerge above the noise and so are believed, and appear in the reconstruction.

At present, we have no actual reconstructions based on this idea, and so do not know whether there are unrecognized difficulties with it. In one highly oversimplified case, where the data give evidence only for p_1, John Skilling concludes that the $N\sigma \simeq A$ choice leads to absurd conclusions about $(p_2 - p_3)$. Yet there are at least conceivable, and clearly definable, circumstances in which they are not absurd. If the true scene is composed of N_i quanta of intensity in the i'th pixel (whether placed there by monkeys or not) then p_1 cannot be measured more accurately – because it is not even defined more accurately – than $dM_k/A = N^{-1}$. It is not possible to 'measure' p_1 to one part in 100 unless N_1 is at least 100.

Then if we specify that p_1 is measured more and more accurately without limit, we are not considering a single problem with fixed N and a sequence of smaller and smaller values of a parameter σ. We are considering a sequence of problems with different N, in which we are drawing larger and larger samples, of size $N_1 = Np_1$. From this one expects to estimate other quantities to a relative accuracy improving like $N^{-1/2}$. This is not to say that we are "measuring" $(p_2 - p_3)$ more and more accurately; we are not measuring it at all. In a sequence of different states of knowledge we are *inferring* it more and more confidently, because the statement of the problem – that p_1 was measured very accurately – implies that N must have been large.

Doubtless, there are other conceivable circumstances (*i.e.* other states of knowledge about how Nature has generated the scene) in which our conclusion about $(p_2 - p_3)$ would indeed be absurd. Any new information which could make our old estimate seem absurd would be, to put it mildly, highly cogent; and it would seem important that we state explicitly what this information is so we can take full advantage of it. But at present, not having seen this information specified, we do not know how to use it to correct our estimate of $(p_2 - p_3)$; no alternative estimate was proposed.

This situation of unspecified information – intuition feels it but does not define it – is not anomalous, but the usual situation in exploring this neglected part of probability theory. It is not an occasion for dispute, but for harder thinking on a technical problem that is qualitatively different from the ones scientists are used to thinking about. One more step of that harder thinking, in a case very similar to this, appears in our discussion of the kangaroo problem below.

In any event, as was stressed at the 1981 Laramie Workshop and needs to be stressed again, the question of the choice of N cannot be separated from the choices of m and n, the number of pixels into which we resolve the blurred image and the reconstruction, and u, v, the quantizing increments that we use to represent the data $d(k)$ and the reconstruction $p(i)$ for calculational purposes.

In most problems the real and blurred scenes are continuous, and the binning and digitization are done by us. Presumably, our choices of (N, m, n, u, v) all express something about the fineness of texture that the data are capable of supporting; and also some compromises with computation cost. Although computer programmers must necessarily have made decisions on this, we are not aware of any discussion of the problem in the literature, and the writer's thinking about it thus far has been very informal and sketchy. More work on these questions seems much needed.

In this connection, we found it amusing to contemplate going to the "Fermi statistics" limit where n is very large and we decree that each pixel can hold only one dot or none, as in the halftone method for printing photographs.

Also one may wonder whether there would be advantages in working in a different space, expanding the scene in an orthogonal basis and estimating the expansion coefficients instead of the pixel intensities. A particular orthogonal basis recommends itself; that generated by the singular-value decomposition of the smearing matrix A_{ki}. Our data comprise an $(m \times 1)$ vector: $d = Ap + e$, where e is the vector of "random errors". Supposing the $(m \times n)$ matrix A to be of rank m, it can be factored:

$$A = VDU^T. \tag{7}$$

where U and V are $(n \times n)$ and $(m \times m)$ orthogonal matrices that diagonalize $A^T A$ and AA^T, and $D^2 = V^T AA^T V$ is the positive definite $(m \times m)$ diagonalized matrix. $D = V^T AU$ is its square root, padded with $(n - m)$ extra columns of zeroes. Label its rows and columns so that $D_{11}^2 \geq D_{22}^2 \geq \cdots$. Then if we use the columns of U as our basis:

$$p_i = \sum_{j=1}^{n} U_{ij}\, a_j\,, \qquad 1 \leq i \leq n \tag{8}$$

our data equation $d = Ap + e$ collapses to

$$d_k - e_k = \sum_{j=1}^{m} V_{kj}\, D_{jj}\, a_j\,, \qquad 1 \leq k \leq m \tag{9}$$

Only the first m expansion coefficients $(a_1 \cdots a_m)$ appear; in this coordinate system the relevance of the data is, so to speak, not spread all over the scene, but cleanly separated off into a known m-dimensional region. The likelihood (3) of a scene becomes

$$L(\text{Scene}) = L(a_1 \cdots a_m) = \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{m} \left[D_{jj}^2 (a_j - b_j)^2 \right] \right\} \tag{10}$$

where $b = dVD^{-1}$ is the data vector in the new coordinates. The expansion coefficients a_j belonging to large eigenvalues of AA^T are determined quite accurately by the data (to $\pm\sigma/D_{jj}$). But the data give no evidence at all about the last $(n - m)$ coordinates $(a_{m+1} \cdots a_n)$.

There might be advantages in a computational scheme that, by working in these coordinates, is able to deal differently with those a_j that are well determined by the data, and those that are undetermined. Perhaps we might decree that for the former "the data come first". But for the latter, the data never come at all.

In any event, whatever our philosophy of image reconstruction, the coordinates $(a_{m+1} \cdots a_n)$ must be chosen solely on grounds of prior information because the data give no evidence about them. If $(a_1 \cdots a_m)$ are specified first, the problem reverts to a pure generalized inverse problem (*i.e.* one with hard constraints). The scene which has maximum entropy subject to prescribed

$(a_1 \cdots a_m)$ is determined without any reference to N. Computational algorithms for carrying out the decomposition (7) are of course readily available (Chambers, 1977).

As we see from this list of unfinished projects, there is room for much more theoretical effort, which might be quite pretty analytically and worthy of a Ph.D. thesis or two; even the specialized monkey approach is open–ended.

KANGAROOS

A different rationale for maximizing entropy was illustrated by Steve Gull, on the occasion of a talk in Australia in 1983, by supposing it established by observation that 3/4 of the kangaroos are left–handed, and 3/4 drink Foster's; from which we are to infer what fraction of them are both right–handed and Foster's drinkers, *etc.*; that is, to reconstruct the (2×2) table of proportions p_{ij}

$$\begin{array}{rcl} & \begin{array}{cc} \mathrm{L} & \mathrm{R} \end{array} & \\ \begin{array}{r} \mathrm{F} \\ \mathrm{no\ F} \end{array} & \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} & \begin{array}{l} 3/4 \\ 1/4 \end{array} \\ & \begin{array}{cc} 3/4 & 1/4 \end{array} & \end{array} \tag{11}$$

from the specified marginal row and column totals given to the right and below the table.

It is interesting to compare the solutions of this problem given by various algorithms that have been proposed. Gull and Skilling (1984), applying the work of Shore and Johnson, find the remarkable result that if the solution is to be found by maximizing some quantity, entropy is uniquely determined as the only choice that will not introduce spurious correlations in the matrix (11), for which there is no evidence in the data. The maximum entropy solution is then advocated on grounds of logical consistency rather than multiplicity.

I want to give an analysis of the kangaroo problem, with an apology in advance to Steve Gull for taking his little scenario far more literally and seriously than he ever expected or wanted anybody to do. My only excuse is that it is a conceivable real problem, so it provides a specific example of constructing priors for real problems, exemplifying some of our Tutorial remarks about deeper hypothesis spaces and measures. And, of course, the principles are relevant to more serious real problems – else the kangaroo scenario would never have been invented.

What bits of prior information do we all have about kangaroos, that are relevant to Gull's question? Our intuition does not tell us this immediately, but a little pump priming analysis will make us aware of it. In the first place, it is clear from (11) that the solution must be of the form:

$$\begin{pmatrix} (.50+q) & (.25-q) \\ (.25-q) & q \end{pmatrix} , \qquad 0 \le q \le .25 \tag{12}$$

But, kangaroos being indivisible, it is required also that the entries have the form $p_{ij} = N_{ij}/N$ with N_{ij} integers, where N is the number of kangaroos. So for any finite N there are a finite number of integer solutions N_{ij}. Any particular solution will have a multiplicity

$$W = \frac{N!}{N_{11}!\, N_{12}!\, N_{21}!\, N_{22}!} \tag{13}$$

This seems rather different from the image reconstruction problem; for there it was at least arguable whether N makes any sense at all. The maximum entropy scene was undeniably the one the monkeys would make; but the monkeys were themselves only figments of our imagination.

Now, it is given to us in the statement of the problem that we are counting and estimating attributes of kangaroos, which are not figments of our imagination; their number N is a determinate quantity. Therefore the multiplicities W are now quite real, concrete things; they are exactly equal to the number of possibilities in the real world, compatible with the data. It appears that, far from abandoning monkeys, if there is any place where the monkey (combinatorial) rationale seems clearly called for, it is in the kangaroo problem!

Let us see some exact numerical solutions. Suppose $N = 4$; then there are only two solutions:

$$N_{ij} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \tag{14}$$

with multiplicities $W = 12, 4$ respectively. The solution with greater entropy comprises 75% of the feasible set of possibilities consistent with the data. If $N = 16$, there are five integer solutions:

$$\begin{array}{rccccc} N_{ij} = & \begin{pmatrix} 8 & 4 \\ 4 & 0 \end{pmatrix}, & \begin{pmatrix} 9 & 3 \\ 3 & 1 \end{pmatrix}, & \begin{pmatrix} 10 & 2 \\ 2 & 2 \end{pmatrix}, & \begin{pmatrix} 11 & 1 \\ 1 & 3 \end{pmatrix}, & \begin{pmatrix} 12 & 0 \\ 0 & 4 \end{pmatrix} \\ W = & 900900, & 1601600, & 720720, & 87360, & 1820 \\ & 36\%, & 64\%, & 29\%, & 3.5\%, & .07\% \end{array} \tag{15}$$

The single maximum entropy solution comprises nearly two–thirds of the feasible set.

But there are many kangaroos; when $N >> 1$ the multiplicities go asymptotically into $W \to \exp(NH)$ where from (12), the entropy is

$$H = -(.5 + q)\log(.5 + q) - 2(.25 - q)\log(.25 - q) - q\log q \tag{16}$$

This reaches its peak at $q = 1/16$, corresponding as noted to no correlations between the attributes of kangaroos. For $q < 1/16$ we have negative correlations (drinkers tend to be right handed, *etc.*); while the solutions with $q > 1/16$ give positive correlations. Near the peak, a power series expansion yields the asymptotic formula

$$W \sim \exp\left[-\frac{128N}{9}(q - 1/16)^2\right] \tag{17}$$

which would lead us to the (mean $\pm$ standard deviation) estimate of q:

$$q_{est} = \frac{1}{16}\left(1 \pm 3/\sqrt{N}\right) \tag{18}$$

Thus if there are $N = 900$ kangaroos the last factor in (18) is (1 ± 0.1); if $N = 90,000$ it is (1 ± 0.01); and if there are $N = 9,000,000$ kangaroos it becomes (1 ± 0.001). These are the predictions made by uniform weighting on our first (monkey) hypothesis space H1.

Here we can start to discover our own hidden prior information by introspection; at what value of N do you begin feeling unhappy at this result? Most of us are probably willing to believe that the data reported by Steve Gull could justify an estimate of q for which we could reasonably claim 10% accuracy; but we may be reluctant to believe that they could determine it to one part in 1000, however many kangaroos there are.

Eq. (18) is essentially the same kind of result discussed above, that John Skilling called "absurd"; but he could dismiss it before on the grounds that N was only an imagined quantity. Now that argument is unavailable; for N is a real, determinate quantity. So what has gone wrong this time? I feel another Sermon coming on.

SERMON ON THE MULTIPLICITY

However large N, it is a combinatorial theorem that most of the possibilities allowed by the data are within that shrinking interval (18). But at some point someone says: "This conclusion is absurd; I don't believe it!" What is he really saying?

It is well established by many different arguments that Bayesian inference yields the unique consistent conclusions that follow from the model, the data, and the prior information *that was actually used* in the calculation. Therefore, if anyone accepts the model and the data but rejects the estimate (18), there are two possibilities: either he is reasoning inconsistently and his intuition needs educating; or else he has extra prior information.

We have met nobody who claims the first distinction for himself, although we all have it to some degree. Many times, the writer has been disconcerted by a Bayesian result on first finding it, but realized on deeper thought that it was correct after all; his intuition thereby advanced a little more.

The same policy – entertain the possibility that your intuition may need educating, and think hard before rejecting a Bayesian result – is recommended most earnestly to others. As noted in our Tutorial, intuition is good at perceiving the relevance of information, but bad at judging the relative cogency of different pieces of information. If our intuition was always trustworthy, we would have no need for probability theory.

Over the past 15 years many psychological tests have shown that in various problems of plausible inference with two different pieces of evidence to consider, intuition can err – sometimes violently and in opposite directions – depending on how the information is received. Some examples are noted in Appendix A.

This unreliability of intuition is particularly to be stressed in our present case, for it is not limited to the untrained subjects of psychological tests. Throughout the history of probability theory, the intuition of those familiar with the mathematics has remained notoriously bad at perceiving the cogency of multiplicity factors. Some expositions of probability theory start by pointing to the fact that observed frequencies tend to remain within the $\pm n^{-1/2}$. "random error" bounds. This observed property of frequencies, to become increasingly stable with increasing number of observations, is seen as a kind of Miracle of Nature – the empirical fact underlying probability theory – showing that probabilities are physically real things.

Yet as Laplace noted, those frequencies are only staying within the interval of high multiplicity; far from being a Miracle of Nature, the great majority of all things that *could have* happened correspond to frequencies remaining in that interval. If one fails to recognize the cogency of multiplicity factors, then virtually every "random experiment" does indeed appear to be a Miracle of Nature, even more miraculous than (18).

In most of the useful applications of direct probability calculations – the standard queueing, random walk, and stochastic relaxation problems – the real function of probability theory is to correct our faulty intuition about multiplicities, and restore them to their proper strength in our predictions. In particular, the Central Limit Theorem expresses how multiplicities tend to pile up into a Gaussian under repeated convolution.

Present orthodox statistics takes multiplicity into account correctly in sampling distributions, but takes no note of multiplicity on parameter spaces. This can lead to very bad estimates of a parameter whose multiplicity varies greatly within the region of high likelihood. It behooves us to be sure that we are not committing a similar error here.

Bear in mind, therefore, that in this problem the entire population of kangaroos is being sampled; as N increases, so does the amount of data that is generating that estimate (18). Estimates which improve as the square root of the number of observations are ubiquitous in all statistical theory. But if, taking note of all this, you still cannot reconcile (18) to your intuition, then realize the implications. Anyone who adamantly refuses to accept (18) is really saying: "I have extra prior information *about kangaroos* that was not taken into account in the calculation leading to (18)."

More generally, having done any Bayesian calculation, if you can look at the result and know it is "wrong"; *i.e.* the conclusion does not follow reasonably from your information, then you must have extra information that was not used in the calculation. You should have used it.

Indeed, unless you can put your finger on the specific piece of information that was left out of the calculation, and show that the revised calculation corrects the difficulty, how can you be sure that the fault is in the calculation and not in your intuition?

HIDDEN PRIOR INFORMATION

The moral of that Sermon was that, if we react to (18) by casting out the whole monkey picture and calculation, and starting over from the beginning without asking what that extra information is, we are losing the whole value and point of the calculation. The monkey calculation on H1 has only primed the mental pump; at this point, the deep thought leading us down to H2 is just ready to begin:

What do we know about kangaroos,
that our common sense suddenly warns us was relevant,
but we didn't think to use at first?

There are various possibilities; again, intuition feels them but does not define them. Consider first an extreme but conceivable state of prior knowledge:

(H2a): If we knew that the left–handed gene and the Foster's gene were linked together on the same chromosome, we would know in advance that these attributes are perfectly correlated and the data are redundant: $q = 1/4$. In the presence of this kind of prior information the "logical consistency" argument pointing to $q = 1/16$ would be inapplicable.

Indeed, any prior information that establishes a logical link between these two attributes of kangaroos will make that argument inapplicable in our problem. Had our data or prior information been different, in almost any way, they would have given evidence for correlations and MAXENT would exhibit it. The "no correlations" phenomenon emphasized by the kangaroo rationale is a good illustration of the "honesty" of MAXENT (*i.e.* it does not draw conclusions for which there is no evidence in the data) in one particular case. But it seems to us a useful but isolated result – a reward for virtue – rather than a basic desideratum for all MAXENT.

Of course, if we agree in advance that our probabilities are always to be found by maximizing the same quantity whatever the data, then a single compelling case like this is sufficient to determine that quantity, and the kangaroo argument does pick out entropy in preference to any proposed alternative. This seems to have been Steve Gull's purpose, and it served that purpose well.

The H2a case is rather unrealistic, but as we shall see it is nevertheless a kind of caricature of the image reconstruction problem; it has, in grossly exaggerated form, a feature that was missing from the pure monkey picture.

(H2b): More realistically, although there are several species of kangaroos with size varying from man to mouse, we assume that Gull intended his problem to refer to the man–sized species (who else could stand up at a bar and drink Foster's?). The species has a common genetic pool and environment; one is much like another. But we did not have any prior information about left/right–handedness or drinking habits. In this state of prior knowledge, learning that one kangaroo is left–handed makes it more likely that the next one is also left–handed. This positive correlation (not between attributes, but between kangaroos) was left out of the monkey picture.

The same problem arises in survey sampling. Given that, in a sample of only 1000 kangaroos, 750 were left–handed, we would probably infer at once that about 3/4 of the millions of unsampled kangaroos are also left–handed. But as we demonstrate below, this would not follow from Bayes' theorem with the monkey prior (13), proportional only to multiplicities. In that state of prior knowledge (call it I_0), every kangaroo is a separate, independent thing; whatever we learn about one specified individual is irrelevant to inference about any other.

Statisticians involved in survey sampling theory noticed this long ago and reacted in the usual way: if your first Bayesian calculation contradicts intuition, do not think more deeply about what prior information your intuition was using but your calculation was not; just throw out all Bayesian notions. Thus was progress limited to the bits of Bayesian analysis (stratification) that intuition could perceive without any theory, and could be expressed in non–Bayesian terms by putting it into the model instead of the prior probability.

Following Harold Jeffreys instead, we elect to think more deeply. Our state of knowledge anticipates some positive correlation between kangaroos, but for purposes of defining H2, suppose that we have no information distinguishing one kangaroo from another. Then whatever prior we assign over the 4^N. possibilities, it will be invariant under permutations of kangaroos.

This reduces the problem at once; our neglected prior information about kangaroos must be all contained in a single function $g(x_1, x_2, x_3)$ of three variables (the number of attributes minus one) rather than N (the number of kangaroos). For it is a well–known theorem that a discrete distribution over exchangeable kangaroos (or exchangeable anything else) is a de Finetti mixture of multinomial distributions, and the problem reduces to finding the weighting function of that mixture.

For easier notation and generality, let us now label the four mutually exclusive attributes of kangaroos by (1, 2, 3, 4) instead of (11, 12, 21, 22), and consider instead of just four of them, any number n of mutually exclusive attributes, one of which kangaroos must have. Then de Finetti's famous theorem (Kyburg & Smokler, 1981) says that there exists a generating function $G(x_1 \cdots x_n)$ such that the probability that N_1 of them have attribute 1, and so on, is

$$p(N_1 \cdots N_n|I) = W(N) \int x_1^{N_1} \cdots x_n^{N_n} \, G(x_1 \cdots x_n) \, dx_1 \cdots dx_n \tag{19}$$

where $W(N)$ is the monkey multiplicity factor (4). Normalization for all N requires that G contain a delta–function:

$$G(x_1 \cdots x_n) = \delta(\Sigma x_i - 1)/, g(x_1 \cdots x_n) \,. \tag{20}$$

Since g need be defined only when $\Sigma x_i = 1$, it really depends only on $(n-1)$ variables, but it is better for formal reasons to preserve symmetry by writing it as in (20).

As it stands, (19) expresses simply a mathematical fact, which holds independently of whatever meaning you or I choose to attach to it. But it can be given a natural Bayesian interpretation if we think of $(x_1 \cdots x_n)$ as a set of "real" parameters which define a class of hypotheses about what is generating our data. Then the factor

$$p(N_1 \cdots N_n|x_1 \cdots x_n) = W(N) \, x_1^{N_1} \cdots x_n^{N_2} \tag{21}$$

is the multinomial sampling distribution conditional on those parameters; the hypothesis indexed by $(x_1 \cdots x_n)$ assigns a probability numerically equal to x_1 that any specified kangaroo has attribute 1, and so on. This suggests that we interpret the generating function as

$$G(x_1 \cdots x_n) = p(x_1 \cdots x_n|I) \,, \tag{22}$$

the prior probability density for those parameters, following from some prior information I. Note, to head off a common misconception, that this is in no way to introduce a "probability of a probability". It is simply convenient to index our hypotheses by parameters x_i chosen to be numerically equal to the probabilities assigned by those hypotheses; this avoids a doubling of our notation. We could easily restate everything so that the misconception could not arise; it would only be rather clumsy notationally and tedious verbally.

However, this is a slightly dangerous step for a different reason; the interpretation (21), (22) has a mass of inevitable consequences that we might or might not like. So before taking this road, let us note that we are here choosing, voluntarily, one particular interpretation of the theorem (19). But the choice we are making is not forced on us, and after seeing its consequences we are free to return to this point and make a different choice.

That this choice is a serious one conceptually is clear when we note that (22) implies that we had some prior knowledge about the x_i. But if the x_i are merely auxiliary mathematical quantities defined from $p(N_1 \cdots N_n|I)$ through (19), then they are, so to speak, not real at all, only figments of our imagination. They are, moreover, not necessary to solve the problem, but created on the spot for mathematical convenience; it would not make sense to speak of having prior knowledge about them. They would be rather like normalization constants or MAXENT Lagrange multipliers, which are also created on the spot only for mathematical convenience, so one would not think of assigning prior probabilities to them.

But if we do consider the x_i as "real" enough to have some independent existence justifying a prior probability assignment, (19) becomes a standard relation of probability theory:

$$p(N_1 \cdots N_n|I) = \int d^n x p(N_1 \cdots N_n|x_1 \cdots x_n)\, p(x_1 \cdots x_n|I) \tag{23}$$

in which the left–hand side has now become the joint *predictive* prior probability that exactly N_i kangaroos have attribute i, $1 \le i \le n$.

This choice is also serious functionally, because it opens up a long avenue of mathematical development. We can now invoke the Bayesian apparatus to calculate the joint *posterior* probability distribution for the parameters and the *posterior predictive distribution* for $(N_1 \cdots N_n)$ given some data D. Without the choice (22) of interpretation it would hardly make sense to do this, and we would not see how (19) could lead us to any such notion as a posterior predictive distribution. Any modification of (19) to take account of new data would have to be done in some other way.

But let us see the Bayesian solution. Suppose our data consist of sampling M kangaroos, $M < N$, and finding that M_1 have attribute 1, and so on. Then its sampling distribution is

$$p(D|x_1 \cdots x_n) = W(M)\, x_1^{M_1} \cdots x_n^{M_n} \tag{24}$$

where $W(M)$ is the multiplicity factor (4) with N's replaced by M's everywhere. The posterior distribution is

$$p(x_1 \cdots x_n|DI) = p(x_1 \cdots x_n|I)\, \frac{p(D|x_1 \cdots x_n)}{p(D|I)} = A/, G(x_1 \cdots x_n)\, x_1^{M_1} \cdots x_n^{M_n} \tag{25}$$

where A is a normalizing constant, independent of the x_i, and by G we always mean g with the delta function as in (20). This leads to a predictive posterior distribution for future observations; if we sample K more kangaroos, the probability that we shall find exactly K_1 ith attribute 1, and so on, is

$$p(K_1 \cdots K_n|DI) = A\, W(K) \int G(x_1 \cdots x_n)\, [x_1^{M_1+K_1} \cdots x_n^{M_n+K_n}]\, d^n x \tag{26}$$

These generalities will hold in any exchangeable situation where it makes sense to think of G as a prior probability.

Now, our aim being to relate the choices to the circumstances, we need to think about specific choices of g to represent various kinds of prior information. Some suggestions are before us; a generating function of the form

$$g = A\, x_1^{k-1} x_2^{k-1} \cdots x_n^{k-1} \tag{27}$$

is often called a "Dirichlet prior", although I do not know what Dirichlet had to do with it. For the case $k = 1$ it was given by Laplace (1778) and for general k by Hardy (1889). However, they gave only the choices, not the circumstances; intuitively, just what prior information is being expressed by (27)?

A circumstance was given by the Cambridge philosopher W. E. Johnson (1924); he showed, generalizing an argument that was in the original work of Bayes, that if in (19) all choices of $(N_1 \cdots N_n)$ satisfying $N_i \geq 0$, $\Sigma N_i = N$ are considered equally likely for all N, this uniquely determines Laplace's prior. In a posthumously published work (Johnson, 1932) he gave a much more cogent circumstance, which in effect asked just John Skilling's question: "Where would the next photon come from?". Defining the variables: $y_k = i$ if the k'th kangaroo has attribute i, $(1 \leq k \leq N,\ 1 \leq i \leq n)$, Johnson's "sufficientness postulate" is that

$$p(y_{N+1} = i | y_1 \cdots y_N, I) = f(N, N_i) \tag{28}$$

Let us state what this means intuitively in several different ways: (A) The probability that the next kangaroo has attribute i should depend only on how many have been sampled thus far, and how many had attribute i; (B) If a sampled kangaroo did not have attribute i, then it is irrelevant what attribute it had; (C) A binary breakdown into (i)/(not i) captures everything in the data that is relevant to the question being asked; (D) Before analyzing the data, it is permissible to pool the data that did not yield (i).

Johnson showed that if (28) is to hold for all (N, N_i), this requires that the prior must have the Dirichlet–Hardy form (27) for some value of k. For recent discussions of this result, with extensions and more rigorous proofs, see Good (1965), Zabell (1982).

In particular, an extension we need is that the function $f(N, N_i)$ need not be the same for all i; we may express prior information that is not symmetric over all attributes, without losing either Johnson's basic idea or the symmetry over kangaroos, by using n different functions $f_i(N, N_i)$, which leads to n different values $(k_1 \cdots . k_n)$ of k in the factors of (27). This intuitive insight of Johnson still does not reveal the meaning of the parameter k. Most discussions have favored small values, in $(0 \leq k \leq 1)$, on the grounds of being uninformative. Let us look at the specific details leading to the function $f(N, N_i)$ in (28). Analytically, everything follows from the generalized Beta function integral

$$\int_0^\infty dx_1 \cdots \int_0^\infty dx_n\, x_1^{k_1-1} \cdots x_n^{k_n-1} \delta(\Sigma x_i - a) = \frac{\Gamma(k_1) \cdots \Gamma(k_n)}{\Gamma(k_1 + \cdots + k_n)}\, a^{K-1}. \tag{29}$$

where $K = \Sigma k_i$. Thus a properly normalized generating function is

$$g(x_1 \cdots x_n) = \frac{\Gamma(K)}{\Gamma(k_1) \cdots \Gamma(k_n)}\, x_1^{k_1-1} \cdots x_n^{k_n-1}. \tag{30}$$

Denote by I_D the prior information leading to (30). Conditional on I_D, the probability of obtaining the data $(N_1 \cdots N_n)$ in N observations is given by (19); using (29) and rearranging, we have

$$p(N_1 \cdots N_n | I_D) = \frac{N!\,\Gamma(K)}{\Gamma(N+K)} \cdot \frac{\Gamma(N_1+k_1)}{N_1!\,\Gamma(k_1)} \cdots \frac{\Gamma(N_n+k_n)}{N_n!\,\Gamma(k_n)} . \tag{31}$$

Note that the monkey multiplicity factor $W(N)$ is still contained in (31). For Laplace's prior (all $k_i = 1$) it reduces to

$$p(N_1 \cdots N_n | I_D) = \frac{N!\,(n-1)!}{(N+n-1)!} \tag{32}$$

independent of the N_i, in accordance with Johnson's 1924 circumstance.

This is the reciprocal of the familiar Bose–Einstein multiplicity factor (number of linearly independent quantum states that can be made by putting N bosons into n single–particle states). Indeed, the number of different scenes that can be made by putting N dots into n pixels or N kangaroos into n categories, is combinatorially the same problem; one should not jump to the conclusion that we are invoking "quantum statistics" for photons. Note that the monkey multiplicity factor $W(N)$ is the solution to a very different combinatorial problem, namely the number of ways in which a given scene can be made by putting N dots into n pixels.

In the "uninformative" limit where one or more of the $k_i \to 0$, the integral (29) becomes singular. However, the relevant quantity (31) is a ratio of such integrals, which does not become singular. In the limit it remains a proper (*i.e.* normalized) distribution, for a typical factor of (31) behaves as follows: as $k \to 0$,

$$\frac{\Gamma(N+k)}{N!\,\Gamma(k)} \to \left\{ \begin{array}{ll} k/N\,, & N > 0 \\ 1\,, & N = 0 \end{array} \right\} \tag{33}$$

Therefore, for example, as $k_1 \to 0$, (31) goes into

$$p(N_1 \cdots N_n | I_D) \to \left\{ \begin{array}{ll} 0\,, & N_1 > 0 \\ p(N_2 \cdots N_n | I_D)\,, & N_1 = 0 \end{array} \right\} \tag{34}$$

The probability is concentrated on the subclass of cases where $N_1 = 0$. In effect, attribute #1 is removed from the menu available to the kangaroos (or pixel #1 is removed from the set of possible scenes). Then if any other $k_i \to 0$, attribute i is removed, and so on.

But if all k_i tend to zero simultaneously in a fixed proportion; for example, if we set

$$k_i = k a_i\,, \qquad a_i > 0\,, \qquad 1 \le i \le n \tag{35}$$

and let $k \to 0+$, (31) goes into

$$p(N_1 \cdots N_n | I_D) \to \left\{ \begin{array}{ll} a_i/\Sigma a_i & \text{if } N_i = N, \text{ all other } N_j = 0 \\ 0\,, & \text{otherwise} \end{array} \right\} \tag{36}$$

and the probability is concentrated entirely on those cases where all kangaroos have the same attribute (or those scenes with all the intensity in a single pixel); *i.e.* the extreme points of the sample space which have the minimum possible multiplicity $W = 1$.

But these results seem even more disconcerting to intuition than the one (18) which led us to question the pure monkey rationale. There we felt intuitively that the parameter q should not be determined by the data to an accuracy of 1 part in 1000. Does it seem reasonable that merely admitting the possibility of a positive correlation between kangaroos, should totally wipe out multiplicity ratios of $10^{100} : 1$, as it appears to be doing in (32), and even more strongly in (36)?

In the inference called for, relative multiplicities are cogent factors. We expect them to be moderated somewhat by the knowledge that kangaroos are a homogeneous species; but surely multiplicities must still retain much of their cogency. Common sense tells us that there should be a smooth, continuous change in our results starting from the pure monkey case to a more realistic one as, starting from the uncorrelated case, we allow the possibility of stronger and stronger correlations. Instead, (32) represents a discontinuous jump to the opposite extreme, which denies entropy any role at all in the prior probability. Eq. (36) goes even further and violently reverses the entropy judgments, placing all the prior probability on the situations of zero entropy!

In what sense, then, can we consider small values of k to be "uninformative"? In view of (34), (36) they are certainly not uninformative about the N_i.

A major thing to be learned in developing this neglected half of probability theory is that the mere unqualified epithet "uninformative" is meaningless. A distribution which is uninformative about variables in one space need not be in any sense uninformative about related variables in some other space. As we learn in quantum theory, the more sharply peaked one function, the broader is its Fourier transform; yet both are held to be probability amplitudes for related variables.

Our present problem exhibits a similar "uncertainty relation". The monkey multiplicity prior is completely uninformative on the sample space S of n^N. possibilities. But on the parameter space X of the x_i it corresponds to an infinitely sharply peaked generating function G, a product of delta functions $\delta(x_i - n^{-1})$. Conversely, small values of k are uninformative about the x_i but highly informative about the different points in S, in the limit (36) tying the sample numbers N_i rigidly together. It is for us to say which, if either, of these limits represents our state of knowledge. This depends, among other things, on the meaning we attach to the variables.

In the present problem the x_j are only provisionally "real" quantities. They were introduced for mathematical convenience, the integral representation (19) being easy to calculate with. But we have avoided saying anything about what they really mean.

We now see one of those inevitable consequences of assigning priors to the x_i, that the reader was warned he might or might not like.

In the kangaroo problem it is the N_i that are the truly, unquestionably real things about which we are drawing inferences. Prior to de Finetti, nobody's intuition had perceived that exchangeability alone, without knowledge of the x_i, is such a strong condition that a broad generating function can force such correlations between all the N_i.

If our prior information were that the x_j are themselves the "real physical quantities" of interest and the N_i only auxiliary quantities representing the exigencies of real data, then a prior that is uninformative about the x_i might be just what we need. This observation opens up another interpretive question about the meaning of a de Finetti mixture, that we hope to consider elsewhere. Now let us examine the opposite limit of (31). As $k \to \infty$, the LHS of (33) goes into $k^N/N!$. Thus as $k_1 \to \infty$, (31) goes into

$$p(N_1 \cdots N_n | I_D) \to \left\{ \begin{array}{ll} 1\,, & N_i = N, \text{ all other } N_i = 0 \\ 0\,, & \text{otherwise} \end{array} \right\} \tag{37}$$

All categories except the first are removed from the menu. But if all k_i increase in a fixed proportion by letting $k \to \infty$ in (35), the limiting form of (31) is

$$p(N_1 \cdots N_n | I_D) \to W(N)\,(k_1/K)^{N_1} \cdots (k_n/K)^{N_n} \tag{38}$$

just the multinomial distribution with selection probabilities k_i/K. If the k_i are all equal, this reverts to a constant times the pure monkey multiplicity from whence we started. So it is the region of large k, not small, that provides the smooth, continuous transition from the "too good" prediction (18).

One way to define an intuitive meaning for the parameters k_i is to calculate Johnson's predictive function $f(N, N_i)$ in (28) or its generalization $f_i(N, N_i)$. With any initial generating function G, (26) shows that, having observed M kangaroos and finding sample numbers $(M_1 \cdots M_n)$, the probability that the next kangaroo sampled will be found to have attribute i is proportional to

$$\int G\,[x_1^{M_1} \cdots x_i^{M_i+1} \cdots x_n^{M_n}]\, d^n x \tag{39}$$

but for the particular generating function (30) the result is given, with the correct normalization factor, by the RHS of (31) after the appropriate changes of notation:

$$k_i \to k_i + M_i; \quad N_i = 1, \text{ all other } N_j = 0.$$

We find

$$f_i(M, M_i) = \frac{\Gamma(M+K)}{\Gamma(M+K+1)} \cdot \frac{\Gamma(M_i+k_i+1)}{\Gamma(M_i+k_i)} = \frac{M_i+k_i}{M+K} \tag{40}$$

a generalized form of Laplace's famous Rule of Succession; it has a strange history.

THE RULE OF SUCCESSION

Given by Laplace in the 18'th Century, this rule came under scathing attack in the 19'th Century from the philosopher John Venn (here in Cambridge, where his portrait can be seen in the Caius College Hall). Although the incident happened a long time ago, some comments about it are still needed because the thinking of Venn persists in much of the recent statistical literature. With today's hindsight we can see that Venn suffered from a massive confusion over "What is the Problem?"

Laplace derived the mathematical result as the solution of one problem. Venn (1866), not a mathematician, ignored his derivation – which might have provided a clue as to what the problem is – and tried to interpret the result as the solution to a variety of very different problems. Of course, he chose his problems so that Laplace's solution was indeed an absurd answer to every one of them. Apparently, it never occurred to Venn that he himself might have misunderstood the circumstances in which the solution applies. R. A. Fisher (1956), pointed this out and expressed doubt as to whether Venn was even aware that Laplace's Rule had a mathematical basis and like other mathematical theorems had "stipulations specific for its validity".

Fisher's testimony is particularly cogent here, for he was an undergraduate in Caius College when Venn was still alive (Venn eventually became the President of Caius College), and they must have known each other. Furthermore, Fisher was himself an opponent of Laplace's methods; yet he is here driven to defending Laplace against Venn. Indeed, it apparently never occurred to Venn that no single result – Laplace's or anybody else's – could possibly have provided the solution to all of the great variety of problems where he tried to use it. Yet we still find Venn's arguments repeated uncritically in some recent "orthodox" textbooks; so let the reader beware.

Now in the 1910's and 1920's Laplace's result became better understood by many: C. D. Broad, H. Jeffreys, D. Wrinch, and W. E. Johnson (all here in Cambridge also). Their work being ignored, it was rediscovered again in the 1930's by de Finetti, who added the important observation that the results apply to all exchangeable sequences. de Finetti's work being in turn ignored, it was partly rediscovered still another time by Carnap and Kemeny, whose work was in turn ignored by almost everybody in statistics, still under the influence of Venn.

It was only through the evangelistic efforts of I. J. Good and L. J. Savage in the 1950's and 1960's, and D. V. Lindley in the 1960's and 1970's, that this exchangeability analysis finally became recognized as a respectable and necessary working part of statistics. Today, exchangeability is a large and active area of research in probability theory, much as Markov chains were thirty years ago.

We think, however, that the autoregressive models, in a sense intermediate between exchangeable and markoffian ones, that were introduced in the 1920's by G. Udny Yule (also here in Cambridge, and living in the same rooms that John Skilling now occupies), offer even greater promise for future applications.

Today, more than 200 years after Laplace started it, great mathematical generalizations are known but we are still far from understanding the useful range of application of exchangeability theory, because the problem of relating the choices to the circumstances is only now being taken seriously and studied as a technical problem of statistics, rather than a debating point for philosophers. Indeed, our present problem calls for better technical understanding than we really have at the moment. But at least the mathematics flows on easily for some distance more.

Thinking of the x_i as "real" parameters, we have a simple intuitive meaning of the hyperparameters $(k_1 \cdots k_n)$ if we denote the observed proportion of attribute i in the sampled population by $p_i = M_i/M$, and define a fictious prior proportion by $g_i = k_i/K$. Then (40) can be written

$$f_i(M, M_i) = \frac{Mp_i + Kg_i}{M + K}\,, \tag{41}$$

a weighted average of the observed proportion and an initial estimate of it. Thus we may regard $K = \Sigma k_i$ as the "weight" we attach to our prior information, measured in equivalent number of observations; *i.e.* the prior information I_D that leads to (30) has the same cogency as would K observations yielding the proportions $g_i = k_i/K$, starting from a state of complete ignorance about the x_i.

We may interpret the k's also in terms of the survey sampling problem. Starting from the prior information I_D and considering the data $(M_1 \cdots M_n)$ to be the result of a survey of $M < N$ kangaroos as in (24)–(26), what estimate should we now make of the proportion of kangaroos with attribute i? What accuracy are we entitled to claim for this estimate?

The answer is given by (26) with $L = N - M$, $L_i = N_i - M_i$. Substituting (30) into (26), sum out $(L_2 \cdots L_n)$ before doing the integrations using (29). The probability that exactly L_1 unsampled kangaroos have attribute 1 is found to be a mixture of binomial distributions:

$$p(L_1|DI) = \int_0^1 p(L_1|x)\, g(x)\, dx \tag{42}$$

where

$$p(L_1|x) = \frac{L!}{L_1!\,(L - L_1)!}\, x^{L_1}\, (1 - x)^{L - L_1} \tag{43}$$

and a generating function

$$g(x) = \frac{\Gamma(b)}{\Gamma(a)\,\Gamma(b - a)}\, x^{a-1}\, (1 - x)^{b-a-1} \tag{44}$$

where $a \equiv (M_1 + k_1)$, $b \equiv (M + K)$. The first two factorial moments of (42) are then

$$\langle L_1 \rangle = L \int_0^1 x\, g(x)\, dx = L\, \frac{a}{b} \tag{45}$$

$$\langle L_1(L_1 - 1) \rangle = L(L - 1) \int_0^1 x^2\, g(x)\, dx = L(L - 1)\, \frac{a(a + 1)}{b(b + 1)} \tag{46}$$

from which the (mean) $\pm$ (standard deviation) estimate of the number L_1 of unsampled kangaroos with attribute #1 is

$$(L_1)_{est} = L \left[p \pm \sqrt{\frac{p(1-p)}{M+K+1} \cdot \frac{M+K+L}{N^2}} \right] \tag{47}$$

where $p = a/b = (M_1 + k_1)/(M+K)$. Comparing with (40) we see that the Rule of Succession has two different meanings; this estimated *proportion* p of unsampled kangaroos with attribute #1 is numerically equal to the *probability* that the next kangaroo sampled will have attribute #1. As we have stressed repeatedly, such connections between probability and frequency always appear automatically, as a consequence of probability theory, whenever they are justified. Generally, the results of survey samplings are reported as estimated fractions of the total population N, rather than of the unsampled part $L = N - M$. Since $(N_1)_{est} = (L_1)_{est} + M_1$, we find from (17), after a little algebra, the estimated fraction of all kangaroos with attribute #1:

$$(N_1/N)_{est} = p + p' \pm \sqrt{\frac{p(1-p)}{M+K+1} \cdot \frac{N+K}{N} \cdot \frac{N-M}{N}} \tag{48}$$

where $p' = (Kp - k_1)/N$.

Examining the dependence of (48) on each of its factors, we see what Bayes' theorem tells us about the interplay of the size of the population, the prior information, and the amount of data. Suppose we have sampled only a small fraction of the population, $M << N$. If we also have relatively little prior information about the x_i, $K << N$, the accuracy of the estimate depends basically on $(M + K + 1)$, the number of actual observations plus the effective number implied by the weight of prior information; and depends little on N. Thus the "too good" estimates implied by (18) as $N \to 0$ are now corrected.

But if $K >> N$ (the monkey multiplicity factor limit), the accuracy goes into the limiting form $p(1-p)/N$ and a result like (18) is recovered. The changeover point from one regime to the other is at about $K = N$. Note, however, that (48) is not directly comparable to (18) because in (18) we used Steve Gull's data on kangaroos to restrict the sample space before introducing probabilities.

Now suppose we have sampled an appreciable fraction of the entire population. Our estimates must perforce become more accurate, and the $(N-M)/N = 1 - (M/N)$ factor so indicates. When we have sampled the entire population, $M = N$, then we know the exact N_1, so the error vanishes, the prior information becomes irrelevant, and the RHS of (48) reduces to $M_1/M \pm 0$, as it should. Thus if we admit the x_i as real quantities, so that it makes sense to apply Bayes' theorem in the way we have been doing, then Bayes' theorem tells us in quantitative detail – just as it always does – what our common sense might have perceived if our intuition was powerful enough.

THE NEW KANGAROO SOLUTION

We started considering Steve Gull's kangaroo problem on the original monkey hypothesis space H1, were somewhat unhappy at the result (18), and have now seen some of the general consequences of going down into H2. How does this affect the answer to the original kangaroo problem, particularly in the region of large N where we were unhappy before?

When the N_i and k_i are large enough to use the Stirling approximation for all terms, a typical term in the exchangeable prior (31) goes into the form

$$L = \log \left[\frac{\Gamma(N+k)}{/N!\,\Gamma(k)} \right] \simeq \log \left[\frac{((N+h)^{N+h}}{N!\,h^h} \right] + const. \tag{49}$$

where $h = k - (1/2)$. Thus, when N and k are quite different we have for all practical purposes

$$L \simeq \left\{ \begin{array}{ll} N\log(ke/N), & N << k \\ k\log(Ne/k), & k << N \end{array} \right\} \tag{50}$$

So if $N_i << k_i$, call it prior information I_2, the prior (31) is given by

$$\log p(N_1 \cdots N_n | I_2) \simeq -\Sigma N_i \log(N_i/k_i) + const. \tag{51}$$

and the most probable sample numbers $(N_1 \cdots N_n)$ subject to any data D that imposes a "hard" constraint on them, are the ones that maximize the entropy relative to the "prior prejudice" (k_i/K). With no prior prejudice, $k_i = k$, this will just lead us back to the original solution (18) from the pure monkey multiplicity factors, confirming again that the region of large k is the one that connects smoothly to the previous solution. When $N >> k$, call it prior information I_2', instead of (51) we have the limiting form

$$\log p(N_1 \cdots N_n | I_2') \simeq \Sigma k_i \log N_i + const. \tag{52}$$

and the solution will be the one that maximizes this expression, which resembles the "Burg entropy" of spectrum analysis. So applying Bayes' theorem with $n = 4$, the exchangeable prior (52) and Steve Gull's hard constraint data

$$D: \quad N_1 + N_2 = N_1 + N_3 = 3N/4$$

the posterior probability for the parameter $q = N_4/N$ can be read off from (12):

$$p(q!DI_2') \propto (0.5 + q)^{k_1} /, (0.25 - q)^{k_2+k_3}\, q^{k_4} \tag{53}$$

When all $k_i = k$, this is proportional to

$$p(q|DI_2') \propto (q - 6q^2 + 32q^4)^k \tag{54}$$

This reaches its peak at $q = 0.915$, and yields the (mean) $\pm$ (standard deviation) estimate

$$q_{est} = 0.915(1 \pm 0.77/\sqrt{k})\,. \tag{55}$$

The "too good" estimate (18) where we had the accuracy factor $(1 \pm 3/\sqrt{N})$, is indeed corrected by this prior information on H2; however large N, the accuracy cannot exceed that corresponding to an effective value

$$N_{eff} = (3/.77)^2\, k = 15.2k = 3.8K \tag{56}$$

These comparisons have been quite educational; we had from the start the theorem that maximizing any quantity other than entropy will introduce correlations in the (2×2) table (12) for which there is no evidence in the data D. That is, starting from the pure monkey solution with $q = 1/16$, learning that one kangaroo is left handed makes no difference; the odds on his being a Foster's drinker remains 3:1. But now, admitting the possibility of a positive correlation between kangaroos must, from the theorem, induce some correlation between their attributes. With the new solution q is increased to about $1/11$; so learning that a kangaroo is left–handed increases the odds on his being a drinker to 3.73:1; while learning that he is right–handed reduces them to only 1.73:1.

At this point our intuition can again pass judgment; we might or might not be happy to see such correlations. Our first analysis of the monkey rationale on H1 was a mental pump–priming that made us aware of relevant information (correlations between kangaroos) that the monkey rationale did not recognize, and led us down into H2. Now the analysis on H2 has become a second mental pump–priming that suddenly makes us aware of still further pertinent prior information that we had not thought to use, and leads us down into H3. When we see the consequences just noted, we may feel that we have overcorrected by ignoring a nearness effect; it is relevant that correlations between kangaroos living close together must be stronger than between those at opposite ends of the Austral contintent. In the U.S.A. there are very marked differences in the songs and other behavior of birds of the same species, living in New Hampshire and Iowa. But an exchangeable model insists on placing the same correlations between all individuals.

In image reconstruction, we feel intuitively that this nearness effect must be more important than it is for kangaroos; in most cases we surely know in advance that correlations are to be expected between nearby pixels, but not between pixels far apart. But in this survey we have only managed to convey some idea of the size of the problem. To find the explicit hypothesis space H3 on which we can express this prior information, add the features that the data are noisy and N is unknown; and work out the quantitative conequences, are tasks for the future.

CONCLUSION: RESEMBLANCE TO THE TRUTH

However far we may go into deeper spaces, we can never escape entirely from the original monkey multiplicity factors, because counting the possibilities is always relevant to the problem, whatever other considerations may also be relevant. Therefore, however you go at it, when you finally arrive a satisfactory prior, you are going to find that monkey multiplicity factor sitting there, waiting for you. This is more than a mere philosophical observation, for the following reason.

In image reconstruction or spectrum analysis, if entropy were not a factor at all in the prior probability of a scene, then we would expect that MAXENT reconstructions from sparse data, although they might be "preferred" on other grounds, would seldom resemble the true scene or the true spectrum.

This would not be an argument against MAXENT in favor of any alternative method, for it is a theorem that no alternative using the same information could have done better. Resemblance to the truth is only a reward for having used good and sufficient information, whether it comes from the data or the prior. If the requisite information is lacking, neither MAXENT nor any other method can give something for nothing. But if the MAXENT reconstruction seldom resembled the truth, neither would we have a very good argument for MAXENT in preference to alternatives; there would be small comfort in the admittedly correct value judgment that MAXENT was the only consistent thing we could have done.

More important, the moral of our Sermons on this in the Tutorial was that if such a discrepancy should occur, far from being a calamity, it might enable us to repeat the Gibbs scenario and find a better hypothesis space. In many cases, empirical evidence on this resemblance to the truth or lack of it for image reconstruction can be obtained. It might be thought that there is no way to do this with astronomical sources, since there is no other independent evidence. For an object of a previously uncharted kind, this is of course true, but we already know pretty well what galaxies look like. If Roy Frieden's MAXENT reconstruction of a galaxy was no more likely to be true than any other, then would we not expect it to display any one of a variety of weird structures different from spiral arms?

We need hardly ask whether MAXENT reconstructions of blurred auto license plates do or do not resemble the true plates, or whether MAXENT tomographic or crystal structure reconstructions do or do not resemble the true objects. If they did not, nobody would have any interest in them.

The clear message is this: if we hold that entropy has no role in the prior probability of a scene, but find that nevertheless the MAXENT reconstructions consistently resemble the true scene, does it not follow that MAXENT was unnecessary? Put differently, if any of the feasible scenes is as likely to be true as the MAXENT one, then we should expect any feasible scene to resemble the truth as much as does the MAXENT one; resemblance to the truth would not be ascribable to the use of MAXENT at all.

It seems to us that there is only one way this could happen. As the amount of data increases, the feasible set contracts about the true scene, and we might conjecture (by analogy with John Parker Burg's shrinking circles for reflection coefficients in spectrum analysis) that eventually all the feasible scenes would resemble the true one very closely, making MAXENT indeed superfluous; any old inversion algorithm, such as the canonical generalized inverse matrix $R = A^T(AA^T)^{-1}$ for Eq. (1), would do as well. If so, how much data would we need to approach this condition?

In March 1984 the writer found, in a computer study of a one–dimensional image reconstruction problem, that when the number of constraints was half the number of pixels the feasible set had not contracted very much; it still contained a variety of wildly different scenes, having almost no resemblance to the true one. The canonical inverse (which picks out the feasible scene of minimum Σf_i^2) was about the wildest of all, grossly underestimating every pixel intensity that was not forced to be large by the data, and having no aversion to negative estimates had the program allowed them. So this amount of data still seems "sparse" and in need of MAXENT; any old algorithm would have given any old result, seldom resembling the truth. Perhaps the conjecture is wrong; more ambitious computer studies and analytical work will be needed to understand this.

To say: "The MAXENT reconstruction is *no more likely* to be true than any other" can be misleading to many, including this writer, because it invites us to interpret "likely" in the colloquial sense of the word. After months of puzzlement over this statement, I finally learned what John Skilling meant by it, through some close interrogation just before leaving Cambridge. Indeed, it requires only a slight rephrasing to convert it into a technically correct statement: "The MAXENT reconstruction has *no more likelihood* than any other with equal or smaller Chisquared." Then it finally made sense.

The point is that "likelihood" is a well-defined technical term of statistics. What is being said can be rendered, colloquially, as "The MAXENT reconstruction is not indicated *by the data alone* any more strongly than any other with equal or smaller Chisquared." But that is just the statement that we are concerned with a generalized inverse problem, from whence we started. In any such problem, a specific choice within the feasible set must be made on other considerations than the data; prior information or value judgments. Procedurally, it is possible to put the entropy factor in either. The difference is that if we consider entropy only a value judgment, it is still "preferred" on logical consistency grounds, but we have less reason to expect that our reconstruction resembles the true scene because we have invoked only our wishes, not any actual information, beyond the data.

In my view, the MAXENT reconstruction is far more "likely" (in the colloquial sense of that word) to be true than any other consistent with the data, precisely because it *does* take into account some highly cogent prior information in addition to the data. MAXENT images and spectrum estimates should become still better in the future, as we learn how to take into account other prior information not now being used. Indeed, John Skilling's noting that bare MAXENT is surprised to find isolated stars, but astronomers are not; and choosing "prior prejudice" weighting factors accordingly, has already demonstrated this improvement.

Pragmatically, all views about the role of entropy seem to lead to the same actual class of algorithms for the current problems; different views have different implications for the future. For diagnostic purposes in judging future possibilities it would be a useful research project to explore the full feasible set very carefully to see just how great a variety of different scenes it holds, how it contracts with increasing data, and whether it ever contracts enough to make MAXENT unnecessary as far as resemblance to the truth is concerned. We conjecture that it will not, because as long as $m < n$ it has not contracted in all directions; *i.e.*, the coordinates $(a_{m+1} \cdots a_n)$ of Eq. (8) remain undetermined by the data, and confined only by nonnegativity.

In the meantime, we think there is still some merit in monkeys, and no one needs to be apologetic for invoking them. If they are not the whole story, they are still relevant and useful, providing a natural starting point from which to construct a realistic prior. For very fundamental reasons they will continue to be so.

APPENDIX A. PSYCHOLOGICAL TESTS

Kahneman and Tversky (1973) report on tests in which subjects were given the prior information: $I \equiv$ "In a certain city, 85% of the taxicabs are blue, 15% green"; and then the data: $D \equiv$ "A witness to a crash who is 80% reliable (*i.e.* who in the lighting conditions prevailing can distinguish correctly between green and blue 80% of the time) reports that the cab involved was green." The subjects are then asked to judge the probability that the cab was actually blue. From Bayes' theorem, the correct answer is

$$p(B|DI) = .85 \times .2/(.85 \times .2 + .15 \times .8) = 17/29 = .59$$

This is easiest to reason out in one's head in terms of odds; since the statement of the problem told us that the witness was equally likely to err in either direction ($G \to B$ or $B \to G$), Bayes' theorem reduces to simple multiplication of odds. The prior odds in favor of blue are 85:15, or nearly 6:1; but the odds on the witness being right are only 80:20 = 4:1, so the posterior odds on blue are 85:60 = 17:12. Yet the subjects in the test tended to guess $p(B|DI)$ as about .2, corresponding to odds of 4:1 in favor of green, thus ignoring the prior information. For them, "the data come first" with a vengeance, even though the prior information implies many more observations than the single datum.

The opposite error – clinging irrationally to prior opinions in the face of massive contrary evidence – is equally familiar to us; that is the stuff of which fundamentalist religious/political stances are made. The field is reviewed by Donmell and Du Charme (1975). It is perhaps not surprising that the intuitive force of prior opinions depends on how long we have held them.

Persons untrained in inference are observed to commit wild irrationalities of judgment in other respects. Slovic et al (1977) report experiments in which subjects, given certain personality profile information, judged the probability that a person is a Republican lawyer to be greater than the probability that he is a lawyer. Hacking (1984) surveys the history of the judicial problem and notes that the Bayesian probability models of jury behavior given by Laplace and long ignored, account very well for the performance of modern English juries. L. J. Cohen (1984) reports on controversy in the medical profession over whether one should, in defiance of Bayesian principles, test first for rare diseases before common ones.

Such findings not only confirm our worst fears about the soundness of jury decisions, but engender new ones about medical decisions. These studies have led to proposals – doubtless 100 years overdue – to modify current jury systems. The services of some trained Bayesians are much needed wherever important decisions are being made.

REFERENCES

L. Boltzmann (1877), Wiener Berichte **76**, p. 373.

J. M. Chambers (1977), *Computational Methods for Data Analysis*, J. Wiley & Sons, Inc., New York.

L. J. Cohen (1984), Epistemologia, VII, Special Issue on Probability, Statistics, and Inductive Logic, pp. 68-71 and 213-222.

M. L. Donmell & W. M. du Charme (1975), "The Effect of Bayesian Feedback on Learning in an 0dds Estimation Task", *0rganizational Behavior and Human Performance*, **14**, 305–313.

H. J. Einhorn & R. M. Hogarth (1981), "Behavioral Decision Theory: Processes of Judgment and Choice", Annual Review of Psychology, **32**, pp. 53–88.

R. A. Fisher (1956), *Statistical Methods and Scientific Inference*, Hafner Publishing Co., New York.

I. J. Good (1965), *The Estimation of Probabilities*, Research Monographs #30, MIT Press, Cambridge Mass.

S. F. Gull & G. J. Daniell (1978) "Image Reconstruction with Incomplete and Noisy Data", Nature, **272**, 686.

S. F. Gull & J. Skilling (1984), "The Maximum Entropy Method" in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Universtiy Press, U. K.

Ian Hacking (1984), "Historical Models for Justice", Epistemologia, VII, Special Issue on Probability, Statistics, and Inductive Logic, pp. 191-212.

G. F. Hardy (1889), Letter in Insurance Record, p. 457. Reprinted in Trans. Fac. Actuaries, **8**, 180 (1920).

E. T. Jaynes (1982) "On the Rationale of Maximum–Entropy Methods", Proc. IEEE, **70**, pp. 939–952.

W. E. Johnson (1924) *Logic, Part III: The Logical Foundations of Science*, Cambridge University Press. Reprinted by Dover Publishing Co., 1964.

W. E. Johnson (1932), "Probability, the Deduction and Induction Problems", Mind **44**, pp. 409–413.

D. Kahneman & A. Tversky (1973), "On the Psychology of Prediction", Psychological Review **80**, pp. 237–251.

H. E. Kyburg & H. I. Smokler (1981), *Studies in Subjective Probability*, 2nd Edition, J. Wiley & Sons, Inc., New York.

P. S. Laplace (1778), Mem. Acad. Sci. Paris, 227–332; Oeuvres Completes **9**, 383–485.

P. Slovic, B. Fischhoff, & S. Lichtenstein (1977), "Behavioral Decision Theory", Annual Review of Psychology **28**, pp. 1–39.

John Venn (1866), *The Logic of Chance*, MacMillan & Co., London.

S. L. Zabell (1982), "W. E. Johnson's 'Sufficientness Postulate'", Ann. Sci., **10**, 1091–1099.