Dice, Entropy, and Likelihood

B. ROY FRIEDEN

We show that a famous die experiment used by E. T. Jaynes as intuitive justification of the need for maximum entropy (ME) estimation admits, in fact, of solutions by classical, Bayesian estimation. The Bayesian answers are the maximum probable (m.a.p.) and posterior mean solutions to the problem. These depart radically from the ME solution, and are also much more probable answers.

INTRODUCTION

Maximum entropy (ME) algorithms currently exist for the estimation of power spectra [1], [2], optical objects [3]–[5], CT images [6], and many other phenomena. (A word on terminology: some workers [4] use "maximum entropy" to mean *maximum probability*, so that the mathematical form of entropy would change with the physics of each problem. We do not do this. By ME we mean the *fixed* mathematical form due to Shannon [8] which is defined below.) By and large, the users of these algorithms are quite happy with their empirical results; this author is one such person.

However, the sense in which ME is an optimum estimator is another matter. For example, this author originally believed ME to provide a maximum probable answer [3]. However, at least for photon images, this is usually wrong [4], [5]. The photons would have to behave like classical particles and be radiated from a gray object [5]. Or, if it were required to estimate the most probable roll occurrences for an unknown die, the die *would have to be known a priori* to be fair [5], a rather restrictive assumption.

Other authors derive an ME principle from information-theoretic arguments: for example, as a special case of an estimator based upon the criterion of maximum information in an optical object given its image [7]. (This is perhaps not surprising, since entropy is an intrinsic part of information theory [8].) Or, heuristic arguments are made that appeal to the reader's sense of what is "reasonable" as an estimate under given conditions [9]–[11].[1]

Manuscript received July 16, 1984, revised January 2, 1985, April 25, 1985, and July 22, 1985. This work was partially supported by a visitor's stipend from the Z.W.O. (The Netherlands Organization for the Advancement of Pure Research) while the author was on sabbatical leave at the University of Groningen.

The author is with the Optical Sciences Center, University of Arizona, Tucson, AZ 85721, USA.

[1]Note that some people use the terms "entropy" and "cross-entropy" interchangeably (the latter concept defined in [11]). We do not do this. The concept of cross-entropy is not treated in this paper.

However, the most compelling "proof" of ME as being optimum in some sense is given by its inventor, E. T. Jaynes. In a thought experiment (described below) with an "unknown die," [12] he showed that an optimum estimate of its biases, given *maximum ignorance* about it, should obey ME. Here, "optimum" is taken in a certain unconventional sense, defined below.

The ME approach is remarkable in finding a way out of a miasma of indeterminacy. However, it has some serious drawbacks. First, maximum entropy makes no claim that its solution is *maximum probable*. Granted, *if* the die happens to be fair, then the solution is maximum probable [12]. However, the problem at hand was to establish the state of fairness of the die in the first place, so this proviso self-destructs the statement. (Many readers, by the way, miss this point and *assume* the ME solution to be maximum probable, regardless of the state of bias of the die.) What ME does claim is that its solution is maximally degenerate: it is the solution that can be combinatorily formed in the most number of ways. This is the sense in which it is optimum. However, from a utilitarian viewpoint, maximum degeneracy is a weaker statement than maximum probability. Usually an engineer wants to know how probable his answer is, not how degenerate it is. The two concepts differ in general, and only coincide when every outcome has the same probability (i.e., when the die is fair).

A second problem with the ME approach to the die problem grows out of the implication (probably unintended) that it was the only way out; that if one wanted a solution to this difficult problem, he would have to abandon all conventional approaches such as Bayes'. More seriously, in light of the preceding paragraph, this leads the reader to the following inescapable (but erroneous) conclusion: it is impossible *in principle* to find a solution (any solution) to the die problem whose probability can be ascertained. This would be a serious and confining statement, if true. For these reasons, we re-examined the die problem, searching for an alternative approach that *would* permit solution probabilities to be computed.

The aim of this paper is to show that the die experiment just spoken of has solutions by *classical*, Bayesian estimation; that the probability of these solutions may be computed, as with any Bayesian problem; that therefore, there is no need to introduce a new concept such as maximum entropy in this most basic of problems; and that maximum entropy is *not* coincident with these solutions. In fact,

0018-9219/85/1200-1764$01.00

maximum entropy not only gives the wrong answer, it gives an answer that is very far from right.

The Problem

The experiment in question is as follows [12]. A die of unknown biases $x_1, \cdots, x_6$ is rolled N times. *Note*: a bias x_i is the *a priori* probability of face value i. The user is supposed to known "nothing" *a priori* about what biases to expect. The arithmetic mean $\bar{n}$ is observed (although the individual occurrence numbers $n_1, \cdots, n_6$ are not). Hence, the user knows $\bar{n}$, where

$$\bar{n} = (n_1 + 2n_2 + \cdots + 6n_6)/N \tag{1}$$

and he knows N, where

$$N = n_1 + n_2 + \cdots + n_6. \tag{2}$$

The problem posed is, given merely $\bar{n}$ and N, can the user make a "good" or "reasonable" estimate of the biases $x_1, \cdots, x_6$ of the die? In particular, *what biases should be assumed present on the very next roll of the die*? We shall solve this problem in a purely classical way, without the need for recourse to any exotic estimator, such as ME. Then we shall solve the problem using ME and compare answers.

To simplify the situation and make for a clearer presentation, consider the allied problem of a 3-sided die. All the salient features of the original problem appear in this simpler one. An effectively 3-sided die may be attained, e.g., by the use of an ordinary die by making outcome 1 or 6 a 1, 2 or 5 a 2, and 3 or 4 a 3. In this way only outcomes 1, 2, or 3 may occur. Let x_1 represent the net probability of a 1 by this scheme (it would be the sum of the probabilities for a 1 and a 6 in the actual die), x_2 represent the probability of a 2 and x_3 that of a 3. For this case of course normalization is obeyed

$$x_1 + x_2 + x_3 = 1 \tag{3}$$

as well as the "3-sided" versions of (1) and (2)

$$\bar{n} = (n_1 + 2n_2 + 3n_3)/N \tag{4}$$

and

$$N = n_1 + n_2 + n_3 \tag{5}$$

due to the N rolls.

The unknowns of the problem are probabilities x_1, x_2, and x_3, which we shall call "biases" because later we shall have to consider the probability of these $\{x_i\}$ and it would be awkward to refer to a "probability of probabilities."

Bayesian Solution

Let us take a Bayesian approach and attempt to construct the posterior probability law $p(x_1, x_2, x_3|\bar{n})$. Once known, this law would enable any number of estimates of the $\{x_i\}$ to be formed based on different cost functions. For example, the maximum *a posteriori* probable (m.a.p.) estimate, posterior mean, etc. These are conventional estimates, of course, that have been widely used and accepted.

By the usual approach of Bayesian statistics use the identity

$$p(x_1, x_2, x_3|\bar{n}) = \frac{P(\bar{n}|x_1, x_2, x_3)\, p_0(x_1, x_2, x_3)}{P(\bar{n})}. \tag{6}$$

The right-hand probabilities define the probability law we seek. Consider these next.

Probability p_0 is the *a priori* probability of biases x_1, x_2, x_3 irrespective of knowing $\bar{n}$, usually called the "prior probability" law.[2] The answer to the problem will usually depend upon what form is assumed for this law. The law describes the user's state of expectation of "ignorance" as to what die *a priori* is present. As mentioned at the outset, the user wants to assume "nothing" about the probable die present. What does "nothing" mean vis-a-vis p_0?

By "nothing" the user usually means that *a priori* every possible set of numbers x_1, x_2, x_3 (obeying normalization equation (3)) may be present with equal probability or frequency. Such a flat or uniform law is widely used in estimation problems. For example: when x_1, x_2, x_3 are the spatial coordinates of a material object whose location in a finite box is completely unknown *a priori*. Or, when a uniformly glowing planar image emits photons from unknown positions $(x, y) = x_1, x_2$. Or, when a distant aircraft of unknown coordinates (x, y) is being tracked; etc. This is also MacQueen and Marschak's definition [13] of maximum ignorance, and we shall use it as well.

Hence, let

$$p_0(x_1, x_2, x_3) = 2!\, \mathrm{rect}\left(x_1 - \tfrac{1}{2}\right) \mathrm{rect}\left(x_2 - \tfrac{1}{2}\right) \cdot \delta(x_1 + x_2 + x_3 - 1) \tag{7}$$

where rect (x) is defined as

$$\mathrm{rect}(x) = \begin{cases} 1, & \text{for } |x| \leqslant \tfrac{1}{2} \\ 0, & \text{for } |x| \geqslant \tfrac{1}{2} \end{cases} \tag{8}$$

and $\delta(x)$ is the Dirac delta function. (The Dirac delta function enforces normalization in (7).)

A possible criticism of this form for p_0 is that "it has a definite form and therefore violates the premise of knowing nothing *a priori*." But recall that we were to know nothing about *the die*; this does not prevent us from *precisely knowing* (with definite form) our state of uncertainty (7) *about* it. To fault the use of (7) on this basis, is, e.g., to fault Heisenberg's uncertainty principle for having a definite form. (Note that the form (7) of $p_0(x)$ may be relaxed anyhow; see the Discussion Section and Appendix II.)

As we shall see, the most valid objection to the use of (7) is that, although it describes "maximum ignorance," it does not describe the user's state *for a die* in particular. The wrong experiment is being performed to model maximum ignorance.

The next probability in (6) is $P(\bar{n}|x_1, x_2, x_3)$. This describes the frequency with which $\bar{n}$ can occur in sequences of die rolls regardless of order. It is convenient to first consider $P(n_1, n_2, n_3|x_1, x_2, x_3)$ for *any* set of occurrences n_1, n_2, n_3, regardless of whether they have mean $\bar{n}$. As these occurrences may occur in any order, the multinomial law applies [14]

$$P(n_1, n_2, n_3|x_1, x_2, x_3) = \frac{N!}{n_1!n_2!n_3!} x_1^{n_1} x_2^{n_2} x_3^{n_3}. \tag{9}$$

But we are only interested in n_1, n_2, and n_3 having mean $\bar{n}$,

[2] Hence p_0 has the curious distinction of being a prior probability of prior probabilities (the xs). This is not merely a play on words. In most problems of Bayesian estimation, the xs are *known* entities defining the user's state of prior knowledge. Here they are the very unknowns of the problem, and instead it is p_0 which defines our state of prior knowledge about *them*. We are one step removed from the usual situation.

i.e., obeying (4) and (5). As these are two equations in three unknowns, we may choose one to be a free parameter and solve for the others in terms of it. Let n_1 be the parameter. Then

$$n_2 = N(3 - \bar{n}) - 2n_1$$

and

$$n_3 = N(\bar{n} - 2) + n_1 \tag{10}$$

satisfy both the datum $\bar{n}$ and normalization.

Event $\bar{n}$ occurs for any n_1 as long as n_2 and n_3 obey (10). Also, since each triplicate n_1, n_2, n_3 is disjoint, the probability of $\bar{n}$ is merely the sum over all such triplicates. Hence

$$P(\bar{n}|x_1, x_2, x_3) = \sum_{n_1=[0,\,N(2-\bar{n})]}^{(N/2)(3-\bar{n})} P(n_1, N(3-\bar{n}) - 2n_1, N(\bar{n}-2) + n_1). \tag{11}$$

Notation $[a, b]$ means the larger of a and b. Integer parts of the n_1 limits are to be taken. As a check, note that if $\bar{n} = 1$, the only term in the sum (11) is that for $n_1 = N$, $n_2 = n_3 = 0$, indicating that only rolls of 1 must have occurred.

Combining (9) and (11), the result is

$$P(\bar{n}|x_1, x_2, x_3) = \sum_{n_1=[0,\,N(2-\bar{n})]}^{[(N/2)(3-\bar{n})]} \cdot \frac{N! x_1^{n_1} x_2^{N(3-\bar{n})-2n_1} x_3^{N(\bar{n}-2)+n_1}}{n_1!\,[N(3-\bar{n}) - 2n_1]!\,[N(\bar{n}-2) + n_1]!}. \tag{12}$$

where $[a]$ means the largest integer not exceeding a.

The final right-hand probability to consider in (6) is $P(\bar{n})$. By the total law of probability [14]

$$P(\bar{n}) = \iiint_{x_1+x_2+x_3=1} dx_1\, dx_2\, dx_3 \cdot P(\bar{n}|x_1, x_2, x_3)\, p_0(x_1, x_2, x_3). \tag{13}$$

This integration can immediately be done since the integrand is already known. Using results (7) and (12), we observe that (13) becomes a sum of Dirichlet integrals [15]

$$\iiint_{x_1+x_2+x_3=1} dx_1\, dx_2\, dx_3\, x_1^{m_1} x_2^{m_2} x_3^{m_3} = \frac{m_1! m_2! m_3!}{(m_1 + m_2 + m_3 + 2)!}. \tag{14}$$

The result is that

$$P(\bar{n}) = \frac{2}{(N+2)(N+1)} \cdot \begin{cases} [(N/2)(\bar{n}-1) + 1], & \text{if } \bar{n} \leqslant 2 \\ [(N/2)(3-\bar{n}) + 1], & \text{if } \bar{n} > 2 \end{cases} \tag{15}$$

At this point, the general answer for the required law $p(x_1, x_2, x_3|\bar{n})$ may be formed by combining (6), (7), (12), and (15). As this would be somewhat cumbersome, let us specialize to an interesting test case.

Case of Sample Mean "In the Middle"

Suppose that the observable $\bar{n} = 2$ is obtained. This is directly in the middle of its range (1,3) of possibilities. By the preceding prescription

$$p(x_1, x_2, x_3|2) = 2(N+1)! \sum_{n_1=0}^{[N/2]} \frac{x_1^{n_1} x_2^{N-2n_1} x_3^{n_1}}{(n_1!)^2 (N-2n_1)!} \tag{16}$$

for N even, or times $(N+2)/(N+1)$ for N odd. This is our posterior probability law.

The first question we ask is, what x_1, x_2, x_3 maximize the probability [16]? Since there are liable to be multiple local maxima, we found the solution by sampling over a fine grid of values x_1 and x_2, making x_3, obey normalization each time. This is an easy problem because only a *two*-dimensional search is necessary; in fact, this is why we analyzed a 3-sided die and not a 6-sided one. Also, that search need only range over a finite range $0 \leqslant x_1 \leqslant 1$, $0 \leqslant x_2 \leqslant 1$. Results are shown in Fig. 1 for two cases—a low $N = 5$ and a moderately high $N = 20$. Contour curves (x_1, x_2) of equal probability p (values 0.05, 0.1, 0.2, etc.) are shown. The curves show the following trends: 1) A principal maximum in p for any N at the point

$$x_1 = x_3 = 0 \quad x_2 = 1 \tag{17}$$

this is the m.a.p. solution; 2) a weak subsidiary maximum at $x_1 = 0.5$, $x_2 = 0$ when N is even; and 3) the spread around the principal maximum becoming ever smaller as N increases. We shall further discuss these curves below, when treating the maximum entropy solution to the problem.

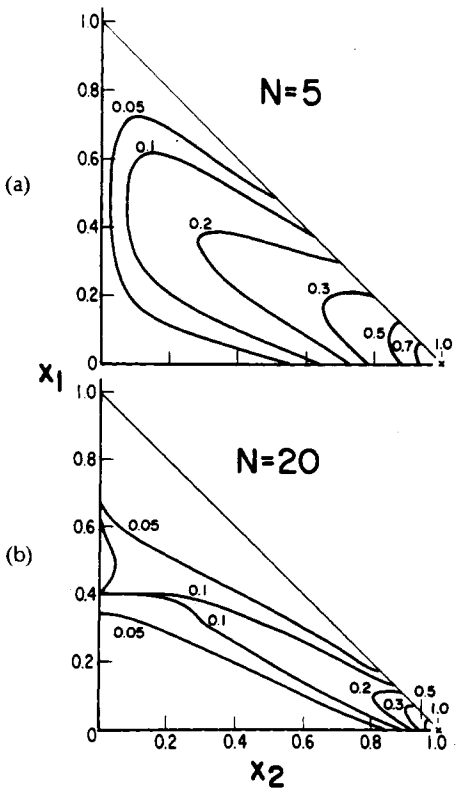

Fig. 1. Contour curves of constant $p(x_1, x_2|\bar{n} = 2)$ for cases (a) $N = 5$ trials and (b) $N = 20$ trials. There is always a single principal maximum, and it resides at the point $x_1 = 0$, $x_2 = 1$. The ME solution point $(\frac{1}{3}, \frac{1}{3})$ enjoys no special status on the plots. There is a weak subsidiary maximum for cases where N is even, but at the point $x_1 = 0.5$, $x_2 = 0$.

Hence, the m.a.p. prediction (17) is that given a sample mean midway in its range, the biases are completely at their *extreme values*! (This is, in fact, why a search was necessary, and why ordinary differentiation to find local maxima would not work.) In other words, the prediction is that *only* roll outcomes 2 occurred!

Actually this result can be explained in hindsight. Suppose we try to simulate the situation by repeatedly selecting sets of biases for a die, rolling the die, and only counting those biases which give rise to the required $\bar{n}$. In this way, $p(x_1, x_2, x_2|\bar{n})$ is built up as a histogram, event by event. Let the biases be selected on a fine grid so that "every" triplet x_1, x_2, x_3 is sampled only once. This accomplishes the flat prior probability law (7). Which such triplet will most often give rise to a value $\bar{n} = 2$? It is obvious that the triplet (17) can *only* give rise to value $n = 2$. So will triplets (x_1, x_2, x_3) a differential distance away. Hence, these triplets will have a peak in the histogram, and consequently be the maximum probable solution to the problem.

A Monte Carlo calculation actually carried through these operations and came up with the same solution. The calculation, in fact, preceded the analytic approach taken here and was motivation for it because of the strange answer.

Intuition, on the other hand, suggests that given a die whose $\bar{n}$ is in the middle, the "correct" estimate ought to be

$$x_1 = x_2 = x_3 = \tfrac{1}{3} \tag{18}$$

i.e., bias values *also* in the middle. Where, then, does intuition go wrong? I believe it traces to the following effect. We are used to dealing with "fair" dice, for which (18) is true *physically* (not as an estimate). Hence, any estimate that *predicts* (18) will tend to seem reasonable.

In fact, regarding real-world dice, given the user's expectations (18) the correct Bayesian approach *should not* assume a flat $p_0(x_1, x_2, x_3)$ present. Instead, p_0 should be made peaked around (18), mirroring the user's expectation that fair dice occur *a priori* more often than biased ones. The solution as in the preceding steps *would* then give a solution near (18) when $\bar{n} = 2$ is observed. This was verified by the use of a $p_0(x_1, x_2, x_3)$ of triangular shape, peaked about $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$. When (16) was modulated by this prior probability law, the maximum was found shifted toward $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$.

What this means is that we are *not* in a state of maximum ignorance when given an unknown die. We know what to expect *a priori* of its biases. For the particular case of a die, a real one, it would be wrong to assume maximum ignorance present. Hence, rolling a die is the *wrong* experiment to use when attempting to model "maximum ignorance" situations. No wonder the result (17) goes against intuition.

On the other hand, maximum ignorance (7) is indeed present in other related problems, as previously described. For problems such as these the preceding analysis holds strictly true, in particular with extreme results like (17) as the maximum probable solution.

In summary, we have found that in situations of true maximum ignorance, *classical* estimation theory is sufficient to provide an answer for the biases implied by a given sample mean; it is not necessary to introduce a new and exotic approach such as ME. Also, the reason the classical answer seems unreasonable when applied to a die in particular is that the user's expectations about a die *are not* so nebulous as to be maximally uncertain. Rolling a die does not represent a situation of maximum ignorance.

We next consider the maximum entropy answer to the same die problem.

Maximum Entropy Solution

The entropy H for a set of biases $\{q_i\}$ is defined as

$$H = -\sum_i q_i \ln q_i. \tag{19}$$

Jaynes' ME approach [11], [12] to the die problem is as follows. Assume that *N is large enough* that the law of large numbers [14] holds, so that the die biases can be well-approximated by values

$$q_i = n_i/N. \tag{20}$$

Note that for N finite, these $\{q_i\}$ *are not* precisely the true biases $\{x_i\}$. On the other hand, *N must be finite* in this approach, or else the ME principle (enunciated below) approaches an ordinary principle of least squares! (See Appendix I.). That is, it merely approaches another classical estimation method, as it approaches a state of being rigorously correct. This is but the first problem with the approach.

Next, the $\{q_i\}$ are sought[3] which simultaneously maximize H and satisfy the constraint equations, here (4) and (5), which in terms of $\{q_i\}$ become

$$q_1 + 2q_2 + 2q_3 = \bar{n} \tag{21}$$

$$q_1 + q_2 + q_3 = 1. \tag{22}$$

The ordinary use of Lagrange multipliers λ_1 and λ_2 results in the ME estimation principle

$$-\sum_{i=1}^{3} q_i \ln q_i + \lambda_1\left(\sum_{i=1}^{3} q_i - 1\right) + \lambda_2\left(\sum_{i=1}^{3} iq_i - \bar{n}\right) = \text{maximum}. \tag{23}$$

It can be shown [12] that the maximum point $\{q_i\}$ is obtainable merely by differentiating (23) in turn with respect to q_1, q_2, and q_3, setting each resulting expression equal to zero. This results in three equations, which are augmented by (21) and (22) to produce a system of 5 equations in 5 unknowns $\{q_i\}, \lambda_1, \lambda_2$. It is easily shown, e.g., by direct substitution, that the solution when $\bar{n} = 2$ is

$$q_1 = q_2 = q_3 = \tfrac{1}{3} \qquad \lambda_1 = 1 - \ln 3 \qquad \lambda_2 = 0. \tag{24}$$

We note with great interest, then, that the ME solution $\{q_i\}$ when the sample mean lies "in the middle" is also "in the middle." This, of course, is diametrically opposite from the true maximum probable solution (17) which lies at an extreme of solution space. A return to Fig. 1 helps to clarify the situation. The ME solution is the point $(\frac{1}{3}, \frac{1}{3})$ on the graphs. Note that there is nothing "special" about this point on either of the two graphs. It is *not* a point of local maximum, e.g., and is in a region of low probability, as indicated by nearby isoprobability lines. Moreover, as N grows (top plot to bottom) the point $(\frac{1}{3}, \frac{1}{3})$ becomes *less*

[3] This was also essentially the approach of H. Theil and K. Laitinen in [17].

probable an answer, not more probable. Hence ME is also not the asymptotic solution as the number of trials increases.

Interestingly, there *is* a local maximum point when N is even (found by one of the reviewers, whom we thank). This is the point (0.5, 0). Again, this is far from the ME point. It is also quite a weak maximum, with value 0.18 of the principal maximum when $N = 20$.

Hence, in a situation (7) of maximum ignorance, ME gives the diametrically wrong answer. As discussed below (18), it only *seems* to provide the right answer because it *is* about right for real-world dice, agreeing with intuition. However, it is wrong for maximum ignorance because real dice do not represent a state of maximum ignorance.

We suggest that in the past readers have been seduced into a belief in ME principally because of this confusion between what constitutes maximum ignorance on one hand, and what constitutes the state of ignorance in a real die experiment on the other. If you want maximum ignorance do not consider a die experiment!

How Probable is the ME Answer Compared with the True One?

But this discussion would be academic if, in fact, the maximum probable solution (17) turned out to be *only slightly* more likely than the ME answer (24). We next compute the ratio R of likelihoods for the two answers using (16)

$$R \equiv \frac{p(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}|2)}{p(0,1,0|2)} = 3^{-N} N! \sum_{n_1=0}^{N/2} \frac{1}{(n_1!)^2 (N-2n_1)!}. \quad (25)$$

As this is purely a function of N, the number of trials, it is easily evaluated. Results are plotted for a range of N values in Fig. 2. The ME answer is always 0.33, or less, probable

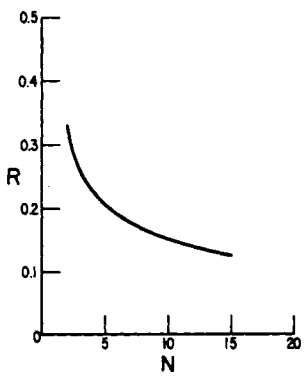

Fig. 2. Ratio R of probability of ME solution to m.a.p. solution. The ratio is quite low for all N.

than the maximum probable answer. With R already down to 0.12 at $N = 15$, it is evident that as N increases further the ME answer is approaching close to zero probability compared with the most probable answer. This is a compelling reason for not using ME in a situation of maximum ignorance. ME does not solve the problem of the completely unknown die.

Mean Biases on the Next Roll

Recall that the problem stated at the outset was to estimate what biases to assume present on the die on its next roll (the $N+1$st). We have been assuming the *maximum probable* set of biases as the solution to this problem. An alternative is to seek the *mean* biases, on the basis that if bias x_1 obeys law $p(x_1)$, there is a spread in possible x_1 values possible on the next roll, *only the most probable of which* has been considered above. The mean x_1 could also be used as representative of the bias value x_1 to use. Of course, the choice is arbitrary; justification for using one or the other statistic depends upon what penalty will be suffered by making a wrong estimate [16]. If the penalty is constant, irrespective of error size, then the maximum probable estimate is chosen; if the penalty goes as the square of the error, then the posterior mean estimate is chosen. Let us seek, then, the posterior mean set of biases $\bar{x}_1, \bar{x}_2, \bar{x}_3$ under the condition $\bar{n} = 2$.

Using (6), (12), (13), and the definition that

$$\bar{x}\, i_1 \equiv \frac{\iiint_{x_1+x_2+x_3=1} dx_1\, dx_2\, dx_3\, x_1\, p(x_1, x_2, x_3|2)}{\iiint_{x_1+x_2+x_3=1} dx_1\, dx_2\, dx_3\, p(x_1, x_2, x_3|2)} \quad (26)$$

leads to determination of $\bar{x}_1$ as

$$\bar{x}_1 = \frac{\iiint_{x_1+x_2+x_3=1} dx_1\, dx_2\, dx_3\, x_1 (x_2^5 + 20x_1x_2^3x_3 + 30x_1^2x_2x_3^2)}{\iiint_{x_1+x_2+x_3=1} dx_1\, dx_2\, dx_3\, (x_2^5 + 20x_1x_2^3x_3 + 30x_1^2x_2x_3^2)} \quad (27)$$

after use of (16). The particular case $N = 5$ was first assumed, for simplicity. The integrals are easily evaluated by means of identity (14). The result is a value $\bar{x}_1 = \frac{1}{4}$. By symmetry in the problem, it also must be that $\bar{x}_3 = \frac{1}{4}$. Hence by normalization $\bar{x}_2 = \frac{1}{2}$.

This solution $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ is still different from the ME one $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$, but closer than the maximum probable answer 0, 1, 0 found before. We are tempted to try higher N, on the basis that $p(x_1)$ might become less peaked at $x_1 = 0$ as N increases, thereby perhaps making $\bar{x}_1$ approach $\frac{1}{3}$. Hence, perhaps with a large enough amount of data the ME solution approaches the mean solution.

Accordingly, we redid steps (26) and (27) but now with general N. Contrary to expectations, it was found that

$$\bar{x}_1 = \bar{x}_3 = \begin{cases} \frac{1}{4}\dfrac{N+4}{N+3}, & \text{for } N \text{ even} \\ \frac{1}{4}, & \text{for } N \text{ odd.} \end{cases} \quad (28)$$

Regardless of N, $\bar{x}_1$ *is always* $\frac{1}{4}$ if N is odd. Also, this is *significantly far* from the ME value $\frac{1}{3}$. It was found by the approach (27) that $\sigma_{\bar{x}_1} \to (4\sqrt{3})^{-1} = 0.144$ with N. Hence a value $x_1 = \frac{1}{3}$ is about half of a standard deviation away from $\bar{x}_1 = \frac{1}{4}$.

If N is even $\bar{x}_1$ can be as large as 3/10, but this is for the extremely sparse data case $N = 2$. For $N \geqslant 10$, we find that $\bar{x}_1 \leqslant 0.27$ and rapidly approaching $\frac{1}{4}$.

The conclusion is that the posterior mean solution never equals the ME solution nor approaches it significantly close, regardless of how much data are present.

Discussion

These results are consistent with previous analysis of a related problem, maximum probable determination of number count distributions [5]. There the problem was to estimate the occurrences $\{n_i\}$ rather than their biases $\{x_i\}$. In particular, it was found that ME is maximum probable only when the die is *a priori known* to be fair! Thus the prior probability law $p_0(x_1, x_2, x_3)$ *must be assumed* to be sharply peaked about values $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$, diametrically opposite from a situation (7) of maximum ignorance. Here again, ME is inconsistent with maximum ignorance.

The foregoing analysis assumed a flat prior law (7) on $p_0(x_1, x_2, x_3)$. What if the form for $p_0(x_1, x_2, x_3)$ were *itself* maximally uncertain? Would the results (17) and (28) for the ML and posterior mean solutions still hold, or would the solutions *now* approach the ME results (24)? It is shown in Appendix II that even with this higher form of ignorance, the solutions are not ME, and remain as before.

Conclusions

The principal underlying reason for many (if not most) people using ME is a feeling that it gives a maximum probable answer in the presence of maximum ignorance. However, with maximum ignorance defined as a situation where every set of possibilities is equally present *a priori* (or more generally as in Appendix II), the most probable solution does not obey ME nor come close to it. The most probable solution is also more probable than the ME solution.

It may be unfair to compare our maximum probable solutions with ME, since ME does not make a claim to be maximum probable. However, each ME solution *does* have a definite probability of being correct under the given conditions of maximum ignorance, and we were curious to know just how large (or small) this was.

We have shown these results by analyzing basically the same die experiment considered by Jaynes [12]. This experiment is usually regarded as strong motivation for the use of ME. Surprisingly, the problem had a solution via ordinary Bayesian estimation, a solution which apparently has been overlooked in the past.

Although we have, for simplicity, worked with a 3-sided die, analogous answers follow for the usual 6-sided one. With $\bar{n} = 3.5$, "in the middle" now for the 6-sided case, the m.a.p. answer with maximum ignorance analogous to (7) present is

$$x_1 = x_2 = x_5 = x_6 = 0 \qquad x_3 = x_4 = 0.5. \tag{29}$$

The ME answer to the same problem is all $x_i = \frac{1}{6}$, once again far from the m.a.p. solution.

It follows that justification for use of ME lies only in intrinsically "fair" cases (where the state of prior knowledge is very strong); or, through the assumption of non-classical criteria such as the information-theoretic or heuristic ones mentioned at the outset.[4]

Appendix I
Limiting Form of the ME Principle as $N \to \infty$

In definition (19) of entropy, let each

$$q_i = x_i + \epsilon_i, \qquad i = 1, \cdots, M \tag{A1}$$

where M is the number of different possible outcomes for the experiment. As before, x_i is the true value of the ith bias. For example, $M = 6$ for an ordinary die. By (20) and normalization property (2)

$$\sum_{i=1}^{M} \epsilon_i = 0. \tag{A2}$$

By (A1)

$$\begin{aligned} \ln q_i &= \ln (x_i + \epsilon_i) = \ln (x_i)(1 + \epsilon_i/x_i) \\ &= \ln x_i + \ln (1 + \epsilon_i/x_i). \end{aligned} \tag{A3}$$

By the law of larger numbers [11]

$$n_i/N \equiv q_i \to x_i \text{ as } N \to \infty. \tag{A4}$$

Therefore by (A1) the ϵ_i must all be small. Accordingly, expand

$$\ln (1 + \epsilon_i/x_i) \to \epsilon_i/x_i - \epsilon_i^2/2x_i^2 \text{ as } N \to \infty. \tag{A5}$$

Then using (A1), (A3), and (A5) in (19)

$$\begin{aligned} H \to &- \sum_{i=1}^{M} (x_i + \epsilon_i)\left[\ln x_i + \epsilon_i/x_i - \epsilon_i^2/2x_i^2\right] \\ &= H_0 - \sum_i \epsilon_i + \sum_i \epsilon_i^2/2x_i - \sum_i \epsilon_i \ln x_i - \sum_i \epsilon_i^2/x_i \end{aligned} \tag{A6}$$

to second order in $\{\epsilon_i\}$, where

$$H_0 \equiv -\sum_i x_i \ln x_i. \tag{A7}$$

However, ME is only maximum probable when the experiment is known to be fair [5], i.e.,

$$x_i = 1/M = \text{constant}. \tag{A8}$$

Under this condition, and using identity (A2), (A6) becomes

$$H \to \text{constant} - (M/2)\sum_i \epsilon_i^2 \equiv \text{maximum} \tag{A9}$$

according to the ME principle. Then reusing (A1), (A9) becomes

$$-(M/2)\sum_i (q_i - 1/M)^2 = \text{maximum}$$

since the additive constant does not affect the solution. This is equivalent to

$$+\sum_i (q_i - 1/M)^2 = \text{minimum} \tag{A10}$$

a least squares principle! Note that any information constraints, such as normalization of the $\{q_i\}$ and a known $\bar{n}$, may be added to (A10) via Lagrange multipliers as in (23).

In summary, ME goes over into a principle of least squares *when* the ME solution is maximum probable and as the number N of observations become indefinitely large. We

[4] A recent derivation of ME rests on the novel premise of "consistency of solution," another non-Bayesian criterion; see [18].

found before (see (20)) that N *must be* indefinitely large in order for the ME solution $\{q_i\}$ to rigorously represent biases $\{x_i\}$. Hence, ME goes over into ordinary least squares when used in a rigorously correct way. Least squares is, of course, a classical approach to estimation.

Appendix II
Classical Estimates when the Prior is Itself Random

We previously assumed "maximum ignorance" to mean a situation where the unknowns x_1, x_2, x_3 were maximally random, obeying (7). This can be imagined to be the situation where the die is chosen from a barrel of dice containing every conceivable set of die biases x_1, x_2, x_3 ranging from completely biased, 0,1,0 to completely unbiased, $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ in equal proportions. The prior p_0 is unique and flat.

We now consider the higher state of ignorance where it is not known *a priori what barrel* will be chosen. There is an infinity of barrels present, where each barrel has dice proportions obeying a *different* law p_0. The shapes of p_0 range from flat (as above) to triangular, to spiked about 1,0,0, etc. Moreover, the state of ignorance is so high that every conceivable barrel (and, consequently, shape p_0) may be chosen with equal probability. Let $i = 1, 2, \cdots, I$ denote barrel number. (I may be indefinitely large.) Then the probability of randomly selecting barrel i is

$$P_i = 1/I, \qquad i = 1, 2, \cdots, I. \tag{A11}$$

Now what is the maximum probable set of biases x_1, x_2, x_3?

Here the $p_0(x_1, x_2, x_3)$ curve depends on the barrel i chosen, so that the former is effectively $p_0(x_1, x_2, x_3|i)$. Then Bayes' rule (6) must now be replaced by its average over all barrels

$$p(x_1, x_2, x_3|\bar{n}) = \frac{P(\bar{n}|x_1, x_2, x_3)}{P(\bar{n})}(1/I)\sum_{i=1}^{I} p(x_1, x_2, x_3|i). \tag{A12}$$

Every conceivable curve shape $p(x_1, x_2, x_3|i)$ is present. Then any one point (x_1, x_2, x_3) sees effectively the same average over i as any other. (See Fig. 3 for the analogous one-dimensional argument.) Hence

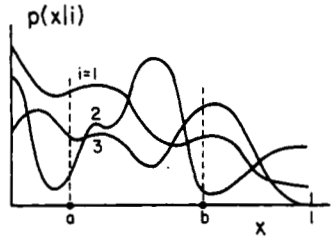

Fig. 3. Some typical curves $p(x|i)$ are shown, $i = 1, 2, 3, \cdots, I$. These curves are constructed to have every conceivable shape consistent with normalization and positivity. Specifically, the values $p(x|i)$, $i = 1, \cdots, N$, x fixed, are imagined to be N random samples from an arbitrary probability law (*on values* p). To keep maximum flexibility in curve shape, in general a different such probability law can hold at each x. However, let each have the same mean, independent of x. Then, by elementary considerations [14], the average curve $I^{-1}\sum_i p(x|i)$ at point x is merely a "sample mean," and hence has a standard deviation $\sigma_i \propto 1/\sqrt{I}$. Finally, since $I \to \infty$, all the $\sigma_i \to 0$ so that the sample means at points x become the theoretical means, and these are by hypothesis all equal.

$$p(x_1, x_2, x_3|\bar{n}) = \frac{P(\bar{n}|x_1, x_2, x_3)}{P(\bar{n})} \times \text{constant}. \tag{A13}$$

This is proportional to (6) and hence will have *the same* maximum-likelihood solutions as before. A look at the quotient form of (26) shows, furthermore, the same posterior mean as before.

We conclude that even this higher state of ignorance does not lead to an ME solution. In fact, we went in the wrong direction. ME is maximum probable only when the die *is known* to be fair [5], i.e., when the state of ignorance is low, not high.

Acknowledgment

It is a pleasure to acknowledge many lively discussions of these questions with B. H. Soffer of Hughes Research Laboratories. R. Kikuchi of Hughes also provided useful comments. Finally, we want to thank one of the anonymous reviewers for his keen, critical, insights on the subject.

References

[1] J. P. Burg, "Maximum entropy spectral analysis," presented at the 37th Annu. Int. Meet. Soc. Explor. Geophys., Oklahoma City, OK, Oct. 1967.

[2] J. G. Ables, "Maximum entropy spectral analysis," *Astron. Astrophys.*, Suppl. 15, pp. 383–393, 1974.

[3] B. R. Frieden, "Restoring with maximum likelihood and maximum entropy," *J. Opt. Soc. Amer.*, vol. 62, pp. 511–518, Apr. 1972.

[4] R. Kikuchi and B. H. Soffer, "Maximum entropy image restoration. I. The entropy expression," *J. Opt. Soc. Amer.*, vol. 67, pp. 1656–1665, Dec. 1977.

[5] B. R. Frieden, "Unified theory for estimating frequency-of-occurrence laws and optical objects," *J. Opt. Soc. Amer.*, vol. 73, pp. 927–938, July 1983.

[6] G. Minerbo, "MENT: A maximum entropy algorithm for reconstructing a source from projection data," *Comp. Graphics Imag. Processing*, vol. 10, pp. 48–68, 1979.

[7] B. R. Frieden, "Image restoration using a norm of maximum information," *Opt. Eng.*, vol. 19, pp. 290–296, May/June, 1980.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[9] I. J. Good, "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables," *Ann. Math. Stat.*, vol. 34, pp. 911–934, 1963.

[10] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 227–241, 1968.

[11] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[12] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, 1982.

[13] J. MacQueen and J. Marschak, "Partial knowledge, entropy, and estimation," *Proc. Nat. Acad. Sci.* (USA), vol. 72, pp. 3819–3824, Oct. 1975.

[14] B. R. Frieden, *Probability, Statistical Optics and Data Testing*. New York: Springer-Verlag, 1983.

[15] M. H. de Groot, *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[16] See, e.g., H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[17] H. Theil and K. Laitinen, "Singular moment matrices in applied econometrics," in P. R. Krishnaiah, Ed., *Multivariate Analysis-V*. Amsterdam: North-Holland, 1980, pp. 629–649.

[18] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, "Consistent inference of probabilities for reproducible experiments," *Phys. Rev. Lett.*, vol. 52, pp. 1357–1360, 1984.