The constraint rule of the maximum entropy principle

Jos Uffink
Department of History and Foundations of Mathematics and Science
University of Utrecht, P.O. Box, 80.000, 3508 TA Utrecht, the Netherlands
e-mail: uffink@fys.ruu.nl

18 October 1995

Contents

Abstract

The principle of maximum entropy is a method for assigning values to probability distributions on the basis of partial information. In usual formulations of this and related methods of inference one assumes that this partial information takes the form of a constraint on allowed probability distributions. In practical applications, however, the information consists of empirical data. A constraint rule is then employed to construct constraints on probability distributions out of these data. Usually one adopts the rule to equate the expectation values of certain functions with their empirical averages. There are, however, various other ways in which one can construct constraints from empirical data, which makes the maximum entropy principle lead to very different probability assignments. This paper shows that an argument by Jaynes to justify the usual constraint rule is unsatisfactory and investigates several alternative choices. The choice of

a constraint rule is also shown to be of crucial importance to the debate on the question whether there is a conflict between the methods of inference based on maximum entropy and Bayesian conditionalization.

1 Introduction

It is an idea, common to all interpretations of probability that whatever the value of the probability of an event or a combination of events may be, the event may still occur, or fail to occur. As a consequence, one cannot determine the values of the probabilities in question by simple inspection of whether the events do, in fact, occur or not. Probabilities are in this sense unobservable quantities. This point creates the problem of how to assess the numerical values to be attached to probabilities and how empirical data can play a role in guiding one's choice. This is the problem of statistical inference, and forms a field full of controversy and debate.

Part of this controversy is due, of course, to the prevailing disagreement on the meaning of probability. But beside this, methods of statistical inference differ widely in how they perceive the goal of the enterprise. Thus the problem of 'assessing' probability values can be construed as a problem of estimation, or one of hypothesis testing or one of finding a best fit, etc. Many different approaches are known, but it is not our purpose to review them. Instead, we focus on a particular method of statistical inference known as the principle of maximum entropy.

The principle of maximum entropy (MEP) is a method of statistical inference which aims at assigning numerical values to probabilities when certain partial or incomplete information is given. To be precise, suppose that probabilities are to be assigned to the set of n mutually exclusive possible outcomes of an experiment. The principle says that these probability values are to be chosen such that the (Shannon) entropy of the distribution $p = (p_1, \ldots, p_n)$, i.e. the expression

$$H(p) = -\sum p_i \log p_i \tag{1}$$

attains its maximum value under the condition that p agrees with the given information.

A closely related method of inference is the principle of maximum *relative* entropy (MREP). This principle has been proposed as a method for updating the values of a previous or prior probability assignment in the light of new partial information. It says that if our prior probability assignment is p and new information is obtained which induces us to revise these values, one should replace the probabilities p_i by new (posterior) values q_i, which agree with the partial information and maximize the relative entropy[1]

$$H(q, p) := -\sum q_i \log \frac{q_i}{p_i}.$$

[1]The relative entropy is also known as (minus) the relative information, and the principle of maximum relative entropy is sometimes called the principle of minimum information.

Both versions of the maximum entropy principle have been subject to considerable controversy. In a previous paper (Uffink, 1995) the merits of these principles have been compared to a class of alternative rules for statistical inference employing a more general entropy expression: the maximum Rényi entropy principles, which can likewise be given an absolute and a relative formulation. In that study it has been assumed without question that the partial information, upon which these methods of statistical inference operate, is represented by constraints on probability distributions, i.e. as unequivocal restrictions on the allowed values of the probabilities with which a distribution p can only agree or disagree. This kind of information is sometimes called 'testable' (Jaynes, 1968).

On the one hand, this testable information is thus supposed to single out a particular subset in the set of all probability distributions. On the other hand, the information is often described as 'data' or 'evidence' i.e. as being of empirical origin. There is an obvious conceptual tension between these two ideas concerning the nature of the information, and this is the source of confusion and controversy quite apart from the question of which exact inference method to apply to this information. The present paper is devoted to a study of this issue.

The tension just mentioned arises because in the approach of the MEP a probability distribution is intended to characterize a judgment, i.e. a state of mind, rather than a state of affairs in the external world. As a consequence, empirical data, conceived of as the results of some experiment, or as an account of a state of affairs, will typically be formulated without recourse to this notion of probability. Such data may, to be sure, consist of average values obtained in repeated measurements, or of observed fluctuations etc.; but they usually do not involve personal beliefs or expectations. The question is then how data, not referring to probability, can nevertheless be modeled as a constraint on probability distributions. The point at stake was formulated in a particularly pungent manner by Skyrms (1985):

> "What do we learn when we learn the constraints? [...] The constraint is a partial specification of what our posterior degrees of belief should be, [...] nothing more or less. But what we learn when we learn something which requires updating our degrees of belief is typically something more or something less, or at any rate something different."

Indeed, taking the subjective view on probability seriously, one would naturally assume that in order to learn something about the value of a probability, one should inspect our mind instead of the external world. In order to find out, say, whether or not there exists life on Mars, it is natural to take account of data collected by space craft. But in order to obtain a (partial) specification of our state of knowledge about the existence of Martians, more obvious methods of enquiry are introspection, or observing one's willingness to lay bets, or, perhaps, some very advanced technique in brain surgery. In short, how and why should data about the state of the world constrain probability distributions, if these probability distributions are not themselves characteristic of the state of the world?

It would be a mistake, however, to think that this gap between empirical data and constraints on probability distributions is peculiar to the subjective view on probability alone. Also within the frequentist view an essential distinction is to be made between the actually observed sample and the population or ensemble from which the sample is drawn. Here too it is not allowed, and often undesirable, to identify or interchange statements about the finite sample and the infinite ensemble.

As one might already expect, the problem we are facing is connected with the thorny problem of induction. Let us see how the maximum entropy principle copes with it.

2 The constraint rule

A celebrated simple application of the MEP, called the Brandeis dice problem, has been discussed repeatedly by Jaynes and his commentators. In 1978, it was formulated as follows (Jaynes, 1983, p. 244):

> "Suppose a die has been tossed N times and we are told only that the average number of spots up was not 3.5 as one might expect for an "honest" die but 4.5. Given this information, *and nothing else*, what probability should we assign to i spots in the next toss? Let us see what solution the Principle of Maximum Entropy gives for this problem, if we interpret the data as imposing the mean value constraints
>
> $$\sum_{i=1}^{6} ip_i = 4.5" \tag{2}$$

Note that Jaynes is careful to say that the data are *interpreted* as imposing a constraint; not that they constitute a constraint by themselves. The interpretational step here is that a statement about the sample average, which concerns the actually observed relative frequencies $n_i = N_i/N$ of throws showing i spots up in our N tosses, namely

$$\sum_{i=1}^{6} in_i = 4.5, \tag{3}$$

is transformed into a statement about an expectation value characterizing probability distributions. Apparently, a rule of the form

$$\sum_i in_i = \sum_i ip_i, \tag{4}$$

or more generally

$$\bar{f} = \langle f \rangle, \tag{5}$$

is employed to bridge the gap between empirical data and probability distributions. (Here, $\bar{f}$ denotes the observed average of a function of the possible outcomes, and $\langle f \rangle$ is its expectation value.) We shall call this the *constraint rule*.

After this step, the solution of the problem according to the MEP is straightforward. One can solve the problem of maximizing (1) under the constraint (2) by the Lagrange multiplier technique. This yields a probability distribution of the form:

$$p_i = \frac{e^{-\beta i}}{Z(\beta)} \tag{6}$$

with

$$Z(\beta) = \sum_i e^{-\beta i} \quad \text{and} \quad -\frac{d}{d\beta} \log Z(\beta) = 4.5.$$

The numerical values are calculated by Jaynes up to five decimal places:

$$(p_1, \ldots, p_6) = (0.05435,\ 0.07877,\ 0.11416,\ 0.16545,\ 0.23977,\ 0.34749). \tag{7}$$

Our problem is the question how to understand and motivate the constraint rule. At first sight, the rules (4) or (5) may seem plausible by an innocent identification of the empirical relative frequencies n_i and the probabilities p_i. Indeed, this may be why in an earlier formulation of the problem (Jaynes, 1983, p. 42) the adoption of the constraint (2) in response to the data (3) was simply considered to be "evident". However, a simplistic identification of frequencies and probabilities is not tenable on any usual view of the meaning of probability whatsoever, at least not for finite N. Thus, the constraint rule (5) is not obvious, and it is important to note, as in the careful formulation Jaynes uses above, that the rule involves an interpretational step which is still open for discussion. Before we go into the explanation Jaynes offers for this rule, let us explore some of its properties.

Consider the case where the number 4.5 in the above example is replaced by 6. This means that the die showed an even more extraordinary behaviour: all our N throws have shown the face with six spots up. Application of the constraint rule (4) now leads to the constraint $\sum_i i p_i = 6$ which allows only a single probability assignment:

$$(p_1, \ldots, p_6) = (0, 0, 0, 0, 0, 1) \tag{8}$$

i.e. we should regard any face other than 6 at the next toss as practically impossible. This seems a rather radical judgment, especially if the number of throws N is small. It violates Jaynes' own judgment that "It is unreasonable to assign zero probability to *any* situation unless our data really rules out that case" (Jaynes, 1983, p. 43). One might feel, therefore, that the probability assignment should in general depend on the value of N too.[2]

[2]It can be argued that the value of N is not available for the probability assignment, because of the "nothing else" clause emphasized in the quotation above. In a later paper Jaynes elucidated this clause by adding: "nothing else (i.e. not making use of any additional information that you or I might get from inspection of the die or from past experience with dice in general)" (Jaynes 1983, p. 323). To me, this suggests that the use of N is allowed, since this number is in the 'public domain', i.e. it is mentioned in the statement of the problem, and not disclosed to you or me alone. But another passage in his work (Jaynes 1983, pp. 271-2) might suggest the contrary reading. In my reading of this last passage, the issue of whether the probability assignment is allowed to depend on N is conflated with the question whether the sample average is known precisely. However this may be, it is clear that also when the value of N is not given it is not thereby safe to assume that it is large.

	$p_1 = p_6$	$p_2 = p_5$	$p_3 = p_4$
N=2	0.1667	0.1667	0.1667
N=4	0.1500	0.1667	0.1833
N=20	0.1440	0.1658	0.1901
N=30	0.1432	0.1658	0.1909
N=60	0.1423	0.1658	0.1919

Table 1: *The probability of i spots on the $N+1^{th}$ throw given that the average numbers of spot in the previous N throws is 3.5, according to the method of inverse probability. (Calculated from equation (40) in the appendix).*

This desire is satisfied in an alternative approach to the problem, by the classical method of inverse probability of Bayes and Laplace. In this approach one starts with the assumption that the die is characterized by an initially unknown 'true' probability distribution $p = (p_1, \ldots, p_6)$, and that the possible values of this distribution are all *a priori* equally likely, i.e. one adopts a uniform prior probability distribution over p. Under these assumptions one can calculate the posterior distribution for p conditional on the given data $\sum_i i n_i = 6$. A standard calculation (given in the appendix) leads to the rule of succession which gives

$$p_1 = \ldots = p_5 = \frac{1}{N+6}, \quad p_6 = \frac{N+1}{N+6}. \tag{9}$$

This assignment remains somewhat more conservative than the MEP assignment (8). Still, it is pleasing that the two rules agree in the limit $N \to \infty$.

However, the result obtained by the method of inverse probability does not always tend to agree with the maximum entropy assignment for large N. Suppose the original value of 4.5 is replaced by the value which one would expect for an unloaded die:

$$\sum i\, n_i = 3.5, \tag{10}$$

Under the constraint rule (4) the MEP now obviously leads to the uniform assignment $p_i = \frac{1}{6}$, which is the distribution with largest entropy attainable for six-faced dice. The assignment resulting from the method of inverse probability, conditionalized on the average (10) in N throws can be calculated as before by the rule of succession. The result (equation (40) in the appendix) is not easy to determine analytically, but some numerical values are collected in Table 1. With increasing N, the assignment disagrees more and more with the MEP assignment!

At first sight this result may seem rather absurd: the occurrence of an event which ought to be expected if the die were honest, apparently induces us to believe it is biased. An (admittedly crude) explanation is that the data (10) are to be expected not only of an honest die; they are in fact even more likely for a die biased towards 3 and 4. It is

not so odd, therefore, that a method of inference should take the data (10) as evidence supporting the hypothesis of a loaded die.

A more serious surprise is that the behaviour shown in Table 1 contradicts the oft-stated belief that in the limit of large N the same results are obtained, whether one uses the method of maximum entropy with a constraint on the expectation value $\langle i \rangle = \alpha$ or conditionalizes on the sample average $\overline{i} = \alpha$. Sometimes this is said to be a consequence of a theorem of van Campenhout and Cover (1981). One should note, however, that the problem addressed by the van Campenhout-Cover theorem is different from the one we are dealing with in the Brandeis dice problem in the sense that in their theorem one considers the probability that one of the *previously* performed throws shows i spots, instead of the *next* throw. Also, it is assumed in this theorem that the throws are independent, instead of the mere conditional independence in the Bayes-Laplace method.[3] (cf. (37) and ((38) in the appendix.)

Of course, the method of inverse probability of Bayes and Laplace, is no less controversial than the maximum entropy approach, and the above comparison is not intended as an objection to the MEP. The literature abounds with a copious amount of objections against this method and the ensuing rule of succession. (See e.g. Venn, 1866, pp. 146–166; Peirce, 1878; Bertrand, 1889, pp 171–174; Fisher, 1973, pp. 24-38 or Jeffreys, 1961, p. 127–132.) In particular, one may well doubt the cogency of the underlying assumptions of this method, i.e. the construction of a prior probability distribution over unknown 'true' probability distributions and the choice that this prior distribution should be uniform, or ask whether all this complies with the clause that 'nothing else' is known, which was emphasized in the statement of the problem. All I can say is that these assumptions would be deemed appropriate *precisely* under this clause by most classical probability theorists. Also, it has been shown by De Finetti that this construction can be replaced by the more appealing and parsimonious assumption of 'exchangeability'.

The main point of displaying the alternative Bayesian method of inverse probability here is to show that doubt on the constraint rule (5) need not lead to sceptical despair. There is a multitude of alternative methods of statistical inference, which can be applied to problems of the Brandeis dice type. Indeed the problem how to connect empirical data to probability assignments is just the very heart of the debate on the foundations of statistics. This naturally leads to the question what arguments can be adduced in favour of the particular constraint rule (5) that is proposed by Jaynes.

Before we turn to this question, it is interesting to ask what solution to the Brandeis dice problem is obtained when the more general principles of maximum Rényi entropy, mentioned in the introduction, are applied. The Rényi entropy of order r of a probability

[3]Note also that in a later paper of 1979 (Jaynes, 1983, p. 323), Jaynes interprets the maximum entropy values (7) of the Brandeis dice problem not as probabilities p_i for i spots at the next toss but instead as estimates of the previously obtained frequencies n_i. This is conceptually *very* different from the problem discussed here. In fact in the original Brandeis lecture of 1962, (ibid., p. 41) this interpretation was expressly rejected. Yet another reading of the problem to which the maximum entropy principle is intended is proposed by Skyrms (1987).

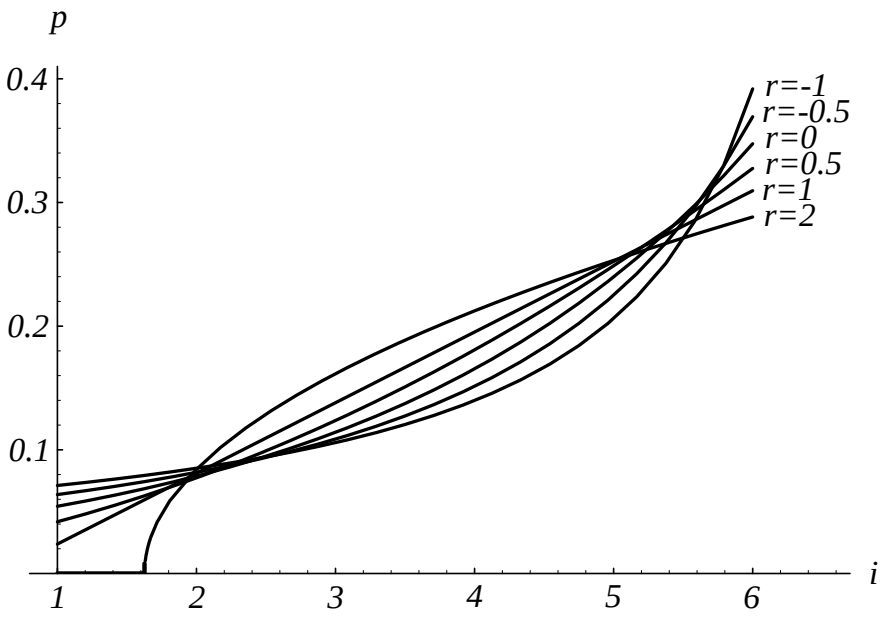


Figure 1: *The probabilities p_i in the Brandeis dice problem according to the maximum Rényi entropy principles for various values of r. The maximum (Shannon) entropy values are labeled by $r = 0$. Note that only the discrete values p_i, with $i = 1, \ldots, 6$ have meaning, the continuous curves are drawn only for convenience. For $r \geq 1$ one has $p_1 = 0$, and in the limit $r \to \infty$ the value p_2 vanishes as well.*

distribution p is defined as

$$\log U_r(p_1, \ldots, p_n) = -\frac{1}{r} \log \sum p_i^{1+r}$$

By the same multiplier technique as before, maximization of U_r leads to the assignment

$$p_i = \sqrt[r]{\lambda(\mu) + \mu i} \tag{11}$$

where λ and μ are again determined by the normalization and the constraint conditions. Fig. 1 displays the form of these solutions for various values of r under the constraint condition (2). Note that, despite the rather different analytical appearance of the expression (6) and (11), the values of this former equation are approached continuously by those of the latter when $r \to 0$.

3 Justification of the constraint rule

In the paper quoted earlier (Jaynes 1978) it is noted that the step equating expectation values (or probability averages) and sample averages is by no means obvious:

> "There is a [...] point of logic about the use of maximum entropy that has troubled some who are able to see the distinction between probability and frequency. In imposing the mean-value constraint [(2)] we are simply appropriating a *sample* average obtained from N measurements [...] and equating it to a *probability* average. [...] Is there not an element of arbitrariness about this? A cynic might say that after all these exhortations about probability and frequency, we proceed to confuse them after all, by using the word 'average' in two quite different senses." (Jaynes, 1983, p. 268)

He goes on to offer a justification for the general rule (5), which relies on the theory of parameter estimation:

> "If we decide to use maximum entropy based on expectations of certain specified functions $f_1(x), \ldots, f_m(x)$, then we know in advance that our final distribution will have the mathematical form
>
> $$p_{\beta_1 \ldots \beta_m}(x) = \frac{1}{Z(\beta_1, \ldots, \beta_m)} \exp\left(-\beta_1 f_1(x) \cdots - \beta_m f_m(x)\right) \tag{12}$$
>
> and nothing prevents us from thinking of this as defining a class of sampling distributions parametrized by the Lagrange multipliers β_k [...] Choosing a specific distribution from this class is then equivalent to making an estimate of the parameters β_k. But parameter estimation is a standard problem of statistical inference." (ibid. p. 269.)[4]

In the theory of parameter estimation, one assumes that an experiment is governed by some probability distribution (the 'sampling' distribution) from an (general) family of distributions $p_{\beta_1 \ldots \beta_m}$. The goal of the theory is to estimate the unknown values of β_k by means of suitably chosen estimators, i.e. functions $\hat{\beta}_k(\vec{x})$ of the outcomes $\vec{x} = (x_1, \ldots, x_N)$ obtained in N independent repetitions of the experiment.

In particular, the well-known method of maximum likelihood is an estimation procedure in which one adopts as estimators $\hat{\beta}_k(\vec{x})$ those values of β_k for which the likelihood function

$$L_{\vec{x}}(\beta_1, \ldots, \beta_m) := p_{\beta_1 \ldots \beta_m}(\vec{x}) \tag{13}$$

attains its maximum for given $\vec{x}$.

Now, consider the estimation of the parameters $\beta_1, \ldots \beta_m$ in the family of distributions (12), assuming that the data $\vec{x}$ are obtained from independent repetitions. We can write the likelihood function as

$$L_{\vec{x}}(\beta_1, \ldots, \beta_m) = \prod_{j=1}^{N} p_{\beta_1 \ldots \beta_m}(x_j) = \frac{1}{Z(\beta_1, \ldots, \beta_m)^N} \exp\left(-\sum_{j=1}^{N} \sum_{k=1}^{m} \beta_k f_k(x_j)\right)$$

[4] In this and following quotations I have adapted the notation to achieve some uniformity throughout the present article.

This function (or, equivalently, its logarithm) is stationary at its maximum, so that one obtains the conditions

$$\frac{\partial}{\partial \beta_k} \log L_{\vec{x}}(\beta_1, \ldots, \beta_m) = 0, \quad k = 1, \ldots, m \tag{14}$$

if the derivatives are evaluated at the maximum likelihood estimates $\beta_k = \hat{\beta}_k(\vec{x})$. In fact, as Jaynes showed, a stationary point must in this case be a maximum if the functions f_k are linearly independent (as we shall assume), so that the conditions (14) are in fact necessary and sufficient for the likelihood function to be maximal.

The equations (14) are, as shown by straightforward calculation, equivalent to

$$\bar{f}_k := \frac{1}{N} \sum_{j=1}^{N} f_k(x_j) = -\frac{\partial}{\partial \beta_k} \log Z(\beta_1, \ldots, \beta_m). \tag{15}$$

The normalization of (12) gives

$$Z(\beta_1, \ldots, \beta_m) = \sum_{x \in S} \exp\left(-\beta_1 f_1(x) \cdots - \beta_m f_m(x)\right)$$

so that

$$-\frac{\partial}{\partial \beta_k} \log Z(\beta_1, \ldots, \beta_m) = \sum_{x \in S} f_k(x) p_{\beta_1, \ldots \beta_m}(x) = \langle f_k \rangle_{\beta_1 \ldots \beta_m}. \tag{16}$$

Combining (15) and (16) one recovers the constraint rule (5). Jaynes concluded from this result:

> "This appears to the writer as a rather complete answer to some objections that have been raised against the constraint rule. We are not, after all, confusing two averages; it is a derivable consequence of probability theory that we *should* set them equal." (ibid. p. 271)

The argument impressed some critics (see Lavis and Milligan, 1985). However, I believe there are three objections to it.

First of all, the argument does not quite offer a justification for the constraint rule (5), because the argument proceeds from the assumption that we adopt the very rule to be justified. Thus, for comparison, suppose we decided to use maximum entropy based on harmonic means. That is, suppose we equate $\overline{f_k^{-1}}^{-1}$ with $\langle f_k^{-1} \rangle^{-1}$. Then the MEP yields a different family of distributions of exponential form (12), with the functions f_k^{-1} replacing f_k. But then too, the maximum likelihood estimator for the Lagrange parameters will pick a value which sets $\langle f_k^{-1} \rangle$ equal to the averages $\overline{f_k^{-1}}$. And so this alternative constraint rule would be justified by the same argument.

For another example, take $m = 1$, $f(x) = x \in I\!N$ and suppose that we decide to equate the maximum in the observed sample with the (essential) supreme value, i.e. with the least value that, with probability one, will not be exceeded. Thus the constraint rule becomes:

$$\max_{j=1,\ldots,N} x_j = \inf\{c \ : \ P(x \leq c) = 1\} \tag{17}$$

Maximizing entropy under this constraint rule leads to a family of block-shaped probability distributions:

$$p_\beta(x) = \begin{cases} \beta^{-1} & \text{when } 1 \leq x \leq \beta \\ 0 & \text{elsewhere.} \end{cases}$$

These distributions are now not of the exponential form, as in (12). Also, condition (14) is not applicable since the likelihood function is not stationary at its maximum. Nevertheless, the maximum likelihood estimator for this case is well-known:

$$\hat{\beta}(x_1, \ldots, x_N) = \max x_j.$$

And so again the maximum likelihood argument justifies the constraint rule (17) from which we started. But of course an argument which will justify all rules really justifies none of them.[5]

As a second objection, it appears awkward to justify maximum entropy inference by taking recourse to the method of maximum likelihood. The latter method is certainly not a derivable consequence of probability theory, as the above quotation suggests. It forms the core of a general (and debatable) theory of statistical inference in its own right. Moreover, the goal of this method, namely to select a unique probability distribution in the light of experimental data, is very similar to that of the MEP. Maximum likelihood should therefore rather be seen a competitor preying in the same arena than as a support for maximum entropy, even if the results of these two methods happen to agree in some simple cases.

As a final objection, one may question whether Jaynes' proposal of treating the Lagrange multipliers as parameters to be estimated is actually consistent with other aspects of his approach. We shall see in more detail in section 5 that on other occasions Jaynes rather forcefully denied the cogency of this idea.

We conclude that the argument to justify the constraint rule (5) by recourse to the maximum likelihood method of parameter estimation is not successful. This raises the question whether there are viable alternatives to the constraint rule or, more generally, other rules to connect empirical data with a choice of probability distribution. The maximum likelihood method, as already mentioned, provides an often-used and well-known alternative with a very similar goal.

But a more extensive debate in the literature has developed over the comparison with another general rival method for statistical inference, which is also claimed by many to be the unique 'consistent' or 'coherent' method. This is the method of Bayesian conditionalization.

[5]The argument does not, literally speaking, justify all constraint rules. The rule to set the sample median of a continous bounded variable $0 \leq x \leq 1$ equal to the 'probability median', i.e. to the least value that is exceeded with probability $1/2$, gives a family of maximum entropy probability densities of the form $p_\beta(x) = (2\beta)^{-1}$ for $0 \leq x < \beta$; $p_\beta(x) = (2 - 2\beta)^{-1}$ for $\beta \leq x \leq 1$. Numerical examples show that the maximum likelihood estimator for β here does not always equal the sample median. Thus, maximum likelihood and maximum entropy do not always agree.

4 Maximum entropy in peaceful coexistence with Bayesian conditionalization

There are at least four different views of the relationship between Bayesian conditionalization and maximum entropy inference. In Jaynes' view there is no conflict between these methods: he emphasizes that each operates on a quite distinct problem area. According to Friedman, Shimony and Dias (Friedman and Shimony, 1973; Dias and Shimony, 1981; Shimony, 1985) the two methods do have an area of application in common on which they lead to contradictory results. Williams (1980) also claims that the two methods have common ground. But he claims that they agree on this ground, and that Bayesian conditionalization is a mere special case of the maximum entropy inference. Skyrms (1985), by contrast, argues that maximum entropy inference is a special case of Bayesian conditionalization.

Thus all conceivable views on the mutual relationship have been advocated in the literature. The ensuing debate (see Hobson, 1972; Shimony, 1973; Gage and Hestenes, 1973; Tribus and Motroni, 1973; Friedman, 1973; Cyranski, 1978) has not brought much clarification. It would be vain to attempt to present a review of all these positions and developments. Instead I shall try to sketch and evaluate three main positions. I will argue that the variety in the verdicts of the above authors is due, not so much on different understandings of Bayesian inference or maximum entropy, but rather to their relying on different construals of the constraint rule.

Let us briefly recall the method of Bayesian conditionalization. Let A, B stand for events, i.e. for subsets of a set S consisting of a total of n possible mutually exclusive elementary events, over which the probability distributions to be considered are defined. The conditional probability of A given B is defined as

$$P(A|B) = P(A \cap B)/P(B). \tag{18}$$

The rule of Bayesian conditionalization then consists of replacing the probability $P(A)$ for any event A by the probability $P(A|B)$ once the event B is observed. I.e. the old (prior) probability assignment $P(A)$ is replaced by a new, (posterior) probability $Q_B(A)$ after the observation of B, which equals the prior conditional probability:

$$Q_B(A) = P(A|B). \tag{19}$$

Note that this is a rule of updating (or probability kinematics), sometimes called Bayes' rule or Bayes' theorem, which involves two different instances in time. It should be firmly distinguished from the relation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{20}$$

also often called Bayes' theorem, but relating only 'synchronous' probabilities: this is an immediate consequence of the definition (18). Both rules, in turn, should be

distinguished from the Bayesian method of inference which operates from the principle that probabilities can be attached to unknown probabilities. This latter method is also known as the method of inverse probability. Note also that one can conditionalize on events only, i.e. on measurable subsets of S, or, when repeated experiments are considered, on subsets in S^N.

The principle of maximum entropy was originally not conceived of as a rule for updating but as a method of assigning values to probabilities. However, as discussed more fully in (Uffink, 1995, section 5), the principle can be generalized into the principle of relative maximum entropy (MREP). This generalized principle is, according to many authors, to be seen as a rule for updating a probability assignment under new information.[6] In particular, the prescription is to choose as a posterior distribution the probability distribution Q that satisfies the constraint set by the new information and maximizes the relative entropy $H(Q, P)$ with respect to the prior distribution P. Here, the relative entropy $H(Q, P)$ for two probability measures is defined as

$$H(Q, P) = -\sup \sum_i Q(A_i) \log \frac{Q(A_i)}{P(A_i)}$$

where the supremum is to be taken over all partitions $(A_1, \ldots, A_m)$ of S. In the case where S contains a finite number of elements, the supremum is attained by the finest partition possible, i.e. when the A_i are singletons and the relative entropy is

$$H(Q, P) = -\sum_{x \in S} q(x) \log \frac{q(x)}{p(x)}.$$

This principle of maximum relative entropy is thus comparable in intent to Bayesian conditionalization. But a fundamental distinction between the two is that whereas Bayesian conditionalization updates on an event, the MREP updates on a constraint, i.e. on a set of probability distributions.[7] These two forms of input are conceptually still very different, and Jaynes rightfully emphasized this important distinction:

> "If a statement d referring to a probability distribution in a space S is testable (for example if it specifies a mean value $\langle f \rangle$ for some function f defined on S), then it can be used as a constraint in the PME; but it cannot be used as a conditioning statement in Bayes' theorem because it is not a statement about any event in S or any other space.
>
> Conversely, a statement D about an event in the space S^N (for example, an observed frequency) can be used as a conditioning statement in applying Bayes' theorem, [...] but it cannot be used as a constraint in applying MEP in space S, because it is not a statement about [...] any probability distribution over S, i.e., it is not testable information in S." (Jaynes 1983, p. 250)

[6] However, Skyrms (1987) proposes the interesting view to regard the principle as a rule for supposing rather than for updating.

[7] In the most general formulation, proposed by Shore and Johnson, the constraint set is allowed to be an arbitrary closed and convex set of probability distributions over S. (Cf. Uffink, 1995, section 6.)

In this view, then, the two methods are designed for distinct types of problems, and they cannot possibly come into conflict.

However, even if we adopt a strict separation between statements about events and those about probability, some sort of bridge between the realms of observable events and of probability distributions will be needed if we are to put the MEP to practical use. Indeed this is exactly what the constraint rule (5) provides by transforming a statement about an event ("$\overline{f} = 4.5$") into a statement about probability distributions ("$\langle f \rangle = 4.5$"). But then the division of labour between maximum entropy and Bayesian conditionalization becomes less clear-cut, and it is here that space for conflict is created. It is into this space that Friedman, Shimony and Dias enter.

5 Maximum entropy in conflict with Bayesian conditionalization

To discuss the objections of Friedman, Shimony and Dias, assume that an N-fold repetition of an experiment is arranged for, such that each performance will lead to a result x in a set S containing n possible outcomes. Assume that, when performed, this procedure yields evidence E_α that incites us to adopt a constraint specifying the expectation value of some real-valued function f defined on S to be equal to α. Thus, the constraint singles out the set of probability distributions

$$\mathcal{I}_\alpha = \{p \ : \ \sum_{x \in S} p(x) f(x) = \alpha\} \tag{21}$$

for some value of α. This evidence may, as usual, consist of the event that the sample average of f in the repeated trial has the value α, thus:

$$E_\alpha = \{(x_1, \ldots, x_N) \ : \frac{1}{N} \sum_j^N f(x_j) = \alpha\} \subset S^N \tag{22}$$

in which case the incitation is by means of the constraint rule (5).

Application of the MEP under the constraint (21) gives a distribution of the form

$$p_\beta(x) = \frac{e^{-\beta f(x)}}{Z(\beta)} \tag{23}$$

where

$$Z(\beta) = \sum_x e^{-\beta f(x)} \tag{24}$$

and β is determined by

$$-\frac{d}{d\beta} \log Z(\beta) = \alpha. \tag{25}$$

Different values of α in (21) lead to different values of β. Thus, Friedman and Shimony argue, one can regard β as a function of α. It is easy to show that this function must be monotonic and hence invertible, because, using (24) and (25), one finds

$$-\frac{d\alpha}{d\beta} = \frac{d^2}{d\beta^2} \log Z(\beta) = \langle (f - \langle f \rangle_\beta)^2 \rangle_\beta$$

which is strictly positive for all β. Hence, the constraint (21) can equivalently be expressed in terms of β. Friedman and Shimony exploit this one-to-one correspondence to relabel the evidence (22):

$$\hat{E}_\beta := E_{\alpha(\beta)}.$$

with

$$\alpha(\beta) = \langle f \rangle_\beta. \tag{26}$$

We now imagine two stages in our experimentation. First consider the instant at which we have not yet obtained the evidence.[8] One may assume that at this stage, our state of knowledge about x is represented by the uniform distribution

$$p(x) = 1/n, \quad \forall x \in S \tag{27}$$

At the same time we are also uncertain about the precise value of α. It is a fundamental tenet of the Bayesian approach that whenever one is uncertain about an event, this uncertainty can be represented by a probability distribution. This means that our state of knowledge is more fully represented as a joint probability distribution about x and α. Since α is continuous, it is convenient to write this joint distribution as $p(x|E_\alpha)\rho(\alpha)$ where $\rho(\alpha)$ is some probability density. The theorem of total probability then gives

$$p(x) = \int p(x|E_\alpha)\rho(\alpha)\, d\alpha \tag{28}$$

Using the relabeling procedure above, we can also write this as

$$p(x) = \int p(x|\hat{E}_\beta)\hat{\rho}(\beta)\, d\beta \tag{29}$$

where

$$\hat{\rho}(\beta) = \rho(\alpha(\beta))|\frac{d\alpha}{d\beta}| \tag{30}$$

Now consider the second stage, at which the evidence E_α (and nothing else) has been obtained. This observation lifts our previous ignorance about α, or β, and updating by Bayesian conditionalization now makes us replace the prior $p(x)$ by the posterior

$$q(x) = p(x|E_\alpha(\beta)) = p(x|\hat{E}_\beta)$$

[8] Friedman and Shimony describe this stage as one in which nothing but the structure of the system is known, i.e. just the fact that there are only n possible values for x. However, it is essential to the argument to assume that evidence enforcing (21) will be forthcoming. Thus, a change of plan, say to collect data constraining another expectation $\langle g \rangle$ instead, must already be excluded at this stage.

Using the MEP (or equivalently, the MREP with the values (27) as the prior distribution), on the other hand, the distribution after the reception of E_α is of the form (23). Consistency between the two methods demands that the two results agree, i.e.

$$p_\beta(x) = p(x|\hat{E}_\beta). \tag{31}$$

Friedman and Shimony showed that this agreement is impossible, under some mild conditions, unless the probability density $\hat{\rho}(\beta)$ reduces to a Dirac delta function:

$$\hat{\rho}(\beta) = \delta(\beta). \tag{32}$$

Following (Skyrms, 1987) one can give the theorem an elegant geometrical formulation, by representing probability distributions as points in an n-dimensional linear space. The set of all possible probability distributions on S forms a $n-1$-dimensional convex subset called an n-simplex. The distributions p_β in this simplex are represented by a curve, parametrized by β. In general, when $n > 3$, such a curve need not be plane. Let us call a curve *strictly convex* if no point on the curve coincides with a proper convex combination of other points from this curve; i.e. if all its points are extreme points of its own convex hull. The (slightly generalized[9]) result of Friedman and Shimony then reads:

Theorem 1 *If a function $f: S \to \mathbb{R}$ on a finite set S takes at least three distinct values and the probabilities*

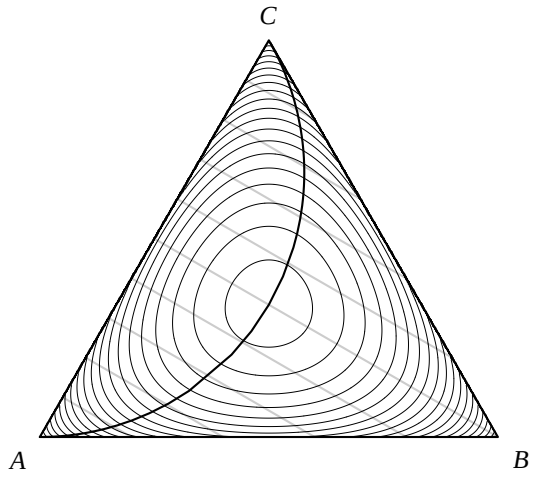
$$p_\beta(x) = \frac{e^{-\beta f(x)}}{Z(\beta)} \quad \textit{for } \beta \in \mathbb{R}$$

are the barycentric coordinates of a point p_β in a n-simplex, then the curve connecting the points p_β is strictly convex. This is not the case if f takes on less than three distinct values.

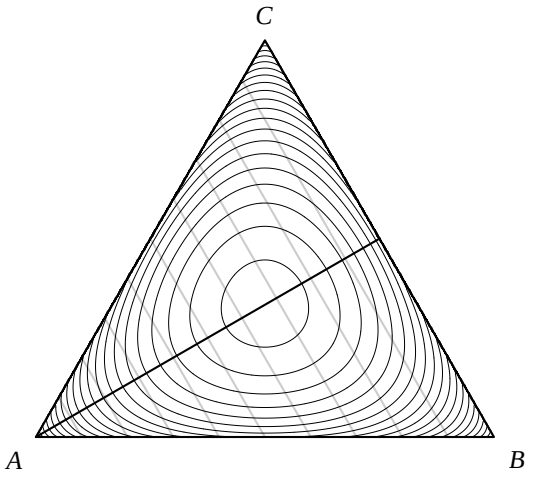
The theorem is illustrated in Fig. 2 and 3 for the case when $n = 3$, so that the simplex becomes an equilateral triangle. A proof of the theorem is given in the Appendix. To see that this theorem entails the result (32), note that under the identification (31) the integral in (29) can be seen as a weighed mixture of points from the curve p_β. For a strictly convex curve such a mixture cannot coincide with any point on the curve, in particular not with the point $p_{\beta=0}$ as demanded by (27), unless the weighing is by a Dirac delta function.

The original article of Friedman and Shimony (1971) concluded from this argument that the MEP is "inconsistent with the principles of probability theory". In later publications this verdict was weakened into "a difficulty verging on an inconsistency" (Friedman, 1973) or an "anomaly that is [...] *almost* a demonstration that the MEP is

[9] Friedman and Shimony prove the therorem in the special case of convexity at the point $\beta = 0$ and under the further assumption that there exists an $x_0 \in S$, such that $f(x_0) = \frac{1}{n}\sum_{x\in S} f(x)$. Seidenfeld (1986) already showed that in a slightly different problem (see section 6) this last restriction can be relaxed into the assumption that f takes at least three distinct values on S.



(Fig. 2)



(Fig. 3)

Figure 2: *The simplex representation of probability distributions over three possible events. Every point in the perfect triangle represents a distribution, the vertices A, B and C representing the extreme distributions* $(1,0,0)$, $(0,1,0)$, *and* $(0,0,1)$ *respectively, and the center point the uniform distribution* $(1/3,1/3,1/3)$. *For every point in the triangle the values* $p(x_1), p(x_2)$ *and* $p(x_3)$ *correspond to the distances to the sides BC, AC and AB respectively. The thin contours connect distributions of equal Shannon entropy; the grey lines represent the constraint sets* $\mathcal{I}_\alpha = \{p \ : \ \langle f \rangle = \alpha\}$, *for various choices of* α, *with* $f(x_i) = i$. *The fat convex curve* p_β *connects the distributions with maximum entropy under these constraints for different values of* α.

Figure 3: *The situation of Fig. 2 in the case where the constraint function f takes on only two distinct values. (Here:* $f(x_2) = f(x_3) \neq f(x_1)$.*) The curve* p_β *is now straight, and all points on this curve except the two endpoints are representable as a proper mixture of other points from the curve.*

inconsistent with Bayesian probability theory"(Shimony, 1985). Indeed, taken by itself, there is nothing inconsistent with probability theory in the idea that our judgment over β should be represented by a Dirac delta. The anomaly is only that this judgment of complete certainty, before the experiment is performed, seems to conflict with the background motivation for the MEP as a code of honesty, of not assuming information which we actually do not possess.

But a more direct conflict with Bayesianism is not very far away from the conclusions of Friedman and Shimony. It is sufficient to note that in Bayesian conditionalization a judgment of complete certainty, as represented by a Dirac delta, is irreversible; i.e. no future evidence can induce the holder of that judgment to a change of mind about β. So if the distribution (32) truly represents our conviction that $\beta = 0$ prior to the experiment's performance, this ought to remain so after the experiment. But this

cannot be reconciled with the idea that in general one should put $\beta = \beta(\alpha) \neq 0$ after the receipt of α.

In his responses to the objection Jaynes argued that the argument rests on a failure "to see the distinction between the Lagrange multiplier β of a maximum entropy problem and an estimated parameter [...]" (Jaynes, 1985, p. 135). More specifically:

> "the quantity β had no previous existence; it is a Lagrange multiplier that is 'created' only in the process of entropy maximization. [...] β is not "estimated" but *defined* by the MEP formalism. [...] It does not make sense therefore, to speak of having prior knowledge of β, much less of honestly representing that knowledge. A Lagrange multiplier does not have a probability distribution." [ibid. p.136]

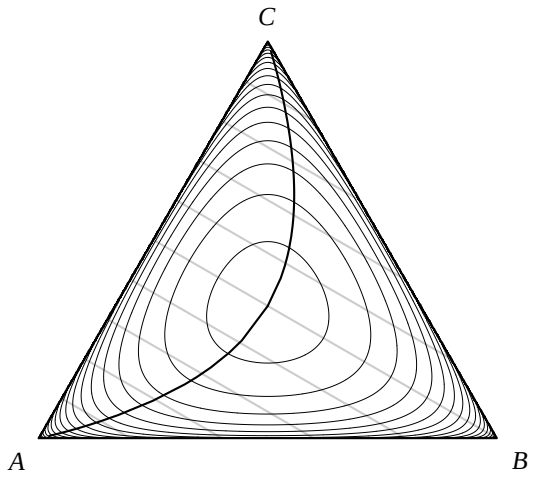
The distinction between parameters which are created and those having previous existence seems somewhat elusive to me. I take it that Jaynes means, in line with his earlier statements, that one cannot define a probability distribution over β, because the value of β does not specify an event, but only labels a probability distribution.

In defence of Shimony and his coworkers one may note, however, that in a previous passage, devoted to the justification of his constraint rule and quoted on page 9, Jaynes seemed to see no problem ("nothing prevents us ...") in treating Lagrange multipliers as unknown parameters to be estimated. It appears to me that the conceptual distinction between probability distributions and events, emphasized by Jaynes to answer the present objections, contradicts his own strategy to justify the constraint rule. Hence, the Friedman-Shimony-Dias objections point to a real dilemma in the maximum entropy approach: it seems one is at least forced to give up one of the two arguments.

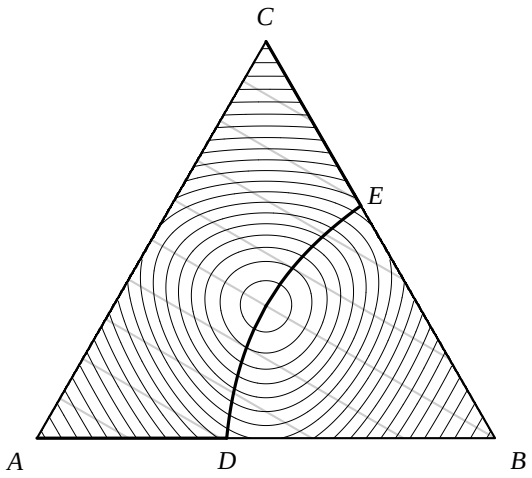
Personally, I think that the constraint rule is not justified by the maximum likelihood argument anyway, and would opt to follow Jaynes in his advice that we strictly distinguish between the parameters that label probability distributions and those that label events. From that perspective, the manoeuvre of Friedman and Shimony to argue for the equivalence of the parametrization in terms of α and β seems suspect. Their way of speaking about these parameters as quantities existing by themselves tends to obscure what they label. If the distinction is maintained strictly, the probability of an event E_α to occur need not be equated with that of the distribution $p_{\beta(\alpha)}$ to be adopted, and their argument fails. However, one cannot really criticize these authors for resting their argument on a one-to-one link between events E_α and corresponding distributions p_β: for this link results from the constraint rule that is used by maximum entropy advocates as a matter of routine.

It is of interest to ask whether the maximum Rényi entropy principles fare differently under the Friedman-Shimony-Dias objection. Although I have not found a proof, it seems highly likely that the Friedman-Shimony theorem holds for the curves obtained from these principles also, in the case of Rényi entropies of order $r \leq 0$. The theorem fails, however, for $r > 0$. Figure 4 displays some typical examples.

It is dubious, however, whether one can derive substantial comfort from this failure. It seems, to me at least, that the most interesting aspect of the Friedman-Shimony-Dias



(Fig. 4a)



(Fig. 4b)

Figure 4: *The same situation of Fig. 2, but with Rényi entropies of order $r = -0.5$ (in Fig. a) and $r = 3$ (Fig. b) replacing the Shannon entropy. The contours are lines of equal Rényi entropy, and the fat curves connect the points of maximum entropy. Note that in Fig. a the entropy surface is flatter at the center and steeper at the edges than in Fig. 2; for Fig. b it is more peaked. In Fig. b the maximum entropy curve (ADEC) moves along the sides of the simplex for a while before crossing, and all its points except four are representable as proper convex mixtures of other points on the curve. This is due to the fact that with Rényi entropies of order $r > 0$ the isentropic contours intersect the boundaries of the simplex at a non-zero angle.*

analysis is the point that if one decides to maximize entropy on constraints of the form (21), before the experimental data are available, one already places strong restrictions on what the final distribution can look like. For example, in the cases depicted in Fig. 2 and 4a, no evidence will make the probability of x_2 increase above its initial value of $1/3$. The paradoxical aspect of this was pointed out by Friedman (1973). He argued that if it is certain beforehand that a probability value will be revised downward, this value must have been too high to start with, and could not have been a faithful representation of our opinion. But similar excluded regions exist in the examples of Fig. 3, 4a and 4b too. The curve in Fig. 4b is not strictly convex, but here one finds that $p(x_2)$ can only decrease from its initial value of $1/3$ unless $p(x_1)$ or $p(x_3)$ is zero. Also, for all Rényi entropies and for all constraints of the form $\langle f \rangle = \alpha$ for a real three-valued constraint function f on $S = \{x_1, x_2, x_3\}$ with $f(x_1) < f(x_2) < f(x_3)$ one has $p_\beta(x_2) \leq 1/2$. That the set of available final judgments is so highly structured is rather unexpected. The remarkable point is not just that the above restrictions are not founded on an empirical basis; but rather that they will not be removed, whatever empirical data may

be obtained in the experiment, or, indeed, how often it may be repeated.[10] It seems that only the suspension of the constraint rule can avoid this conclusion.

6 Maximum entropy as a generalization of Bayesian conditionalization

In section 4 we have discussed the view that the MEP is fundamentally different from Bayesian inference in the sense that whereas Bayesian conditionalization uses events (empirical data) as input, the MEP operates on constraints on probability distributions. However, the MEP routinely employs an additional constraint rule (5), which connects the constraints with empirical data. It is through the adoption of this constraint rule that the two methods may come into conflict, as we have seen in the previous section.

In this section we turn the approach around. Instead of adding a constraint rule to the MEP, that brings this method of inference into contact with the Bayesian procedure, can't we also enrich the procedure of Bayesian conditionalization with some rule to concoct a constraint from the empirical data, and so bring Bayes into the sphere of competence claimed for the MREP? This is indeed the approach taken by Williams (1980) and by Van Fraassen (1981,1989).

In order to find such a rule, note that the posterior probability distribution Q_E obtained by Bayesian conditionalization (19), gives probability one to the observed event: $Q_E(E) = 1$. This suggests the constraint rule that the actually observed data be given probability one. Or, in other words, with the observed event E we associate the set of probability distributions

$$\mathcal{I}_E = \{Q \ : \ Q(E) = 1\} \tag{33}$$

or in the case of probability densities:

$$\mathcal{I}_E = \{q(x) \ : \ \int_E q(x)\, dx = 1\}$$

In this way the data determine a constraint which is exactly of the general form required by the MREP, i.e. $\mathcal{I}_E$ is a closed and convex set of probability distributions.

Thus the two methods are brought once again in each other's reach. But now there is no conflict. Instead, under the constraint rule (33), the MREP leads immediately

[10] Note that if the constraint $\langle f \rangle = \alpha$ is based on a given observed sample average $\bar{f} = \alpha$ and nothing else, as Jaynes proposes, one cannot in general deduce the actually observed frequencies n_1, n_2, n_3. Thus it need not be strange that there should be a non-trivial upper bound on $p(x_2)$: any contribution of a number of x_2's in the observed sample to the sample average can, when $f(x_) < f(x_2) < f(x_3)$, often also be obtained from an appropriate number of x_1's and x_3's. Then even if $\bar{f} = f(x_2)$, one can still not be sure that all performances of the experiment gave the result x_2, But in exceptional cases, say, $f(x_1) = -1, f(x_2) = 0$ and $f(x_3)$ positive and irrational, and the sample average $\bar{f} = 0$, one can deduce the relative frequencies: all repetitions of the experiment must in this case have yielded outcome x_2. To assign a probability of $p(x_2) \leq 1/2$ in that case seems counterintuitive, especially if N is large.

to the Bayesian posterior! This follows by noting that for given P and E the relative entropy $H(Q,P)$ becomes maximal for the distribution Q which is proportional relative to P on the domain where it does not vanish. Hence, it is identical to the Bayesian posterior and the two methods of choosing a posterior give the same result. In the notation of Shore and Johnson (cf. Uffink, 1995, eq. (14)) we can write:

$$Q_E = I_E \circ P$$

where $\circ$ means updating by maximum relative entropy and I_E denotes the constraint specifying that the posterior distribution belongs to the set $\mathcal{I}_E$. Since the MREP is equipped to handle other types of constraints as well, it becomes natural to regard Bayesian conditionalization as just a special case of the Maximum entropy updating! Indeed, Williams showed that by generalizing the condition $Q(E) = 1$ to $Q(E) = \alpha$, for $0 \leq \alpha \leq 1$ in the constraint rule (33), the application of the MREP reproduces the procedure of Jeffrey conditionalization (Jeffrey, 1965), a method which is therefore encompassed by this principle as well. Several remarks should be made with respect to these startling results.

First, a natural question is how the present approach succeeds in escaping the inconsistency found by Friedman and Shimony. To see this, note that the constraint rule (33) can be put in the same form as the rule (5) in terms of the characteristic function of an event E, namely

$$\mathcal{I}_E = \{Q \ : \ \langle \chi_E \rangle = 1\}$$

where

$$\chi_E(x) = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{otherwise} \end{cases}$$

This evades the conditions of theorem 1 because χ_E only takes on two values. Thus a quite radical change in the relationship between the two inference methods is due simply to the restriction of the constraint rule (5) to bivalued functions. The theorem still stands as a limitation to the variety of constraints the MREP can handle without running into a contradiction with Bayesian conditionalization. Seidenfeld (1986) showed that this variety is in fact already exhausted by the two cases mentioned by Williams.

Secondly, it should be emphasized that the present constraint rule (33) is actually radically different from the rule (5) proposed by Jaynes. To see this more clearly, let us return to the problem of $N+1$ tosses with a die, discussed in the Brandeis dice problem of section 3. Let $\vec{x} = (i_1, \ldots, i_{N+1})$ denote a sequence of outcomes and suppose we have observed an average of 4.5 spots up in the first N throws. In the previous, Jaynesian, approach to the MREP one starts with prior probabilities $p_1, \ldots, p_6$ for the outcome of the $N + 1^{th}$ throw and then imposes the constraint

$$\mathcal{I}_{\langle i \rangle} = \{q_i \ : \ \sum_{i=1}^{6} i q_i = 4.5\} \tag{34}$$

to obtain the posterior probabilities: $q_i = I_{\langle i \rangle} \circ p_i$. Where $I_{\langle i \rangle}$ now denotes the constraint that the posterior probabilities belong to (34). This is to be contrasted to the

present, Williamsian approach, in which one first has to extend the domain of the probability distributions to the set S^{N+1} of all possible sequences $\vec{x}$. That is, an extended prior probability distribution $P(\vec{x})$ is assumed of which the previously mentioned values $p_1, \ldots, p_6$ are marginal probabilities:

$$\sum_{i_1=1}^{6} \cdots \sum_{i_N=1}^{6} P(i_1, \ldots, i_{N+1}) = p_i \text{ where } i = i_{N+1}.$$

Only on this domain can one formulate the constraint regarding the observed average of 4.5:

$$\mathcal{I}_E = \{Q(\vec{x}) \ : \ Q(\{\vec{x} \ : \ \frac{1}{N}\sum_{j=1}^{N} i_j = 4.5\}) = 1\} \tag{35}$$

Updating under this constraint one derives the posterior $Q = I_E \circ P$. Finally, the desired probabilities for i spots at the $N+1^{th}$ toss are found from Q by marginalization (i.e. by summing over $i_1, \ldots, i_N$). A special case of the last approach is provided by the method of inverse probability, discussed in section 2, and we have seen already that the results of these approaches need not agree.

It is also clear, however, that in the present approach the prior distribution over the outcomes of the $N+1^{th}$ toss and the constraint (33) by themselves do not settle the values of the posterior probabilities for the $N+1^{th}$ toss. A crucial role is, of course, played by the extension of the prior to a much wider domain. It is by means of this extension that we are provided (among other things) with the (prior) conditional probabilities $P(i_{N+1}|E)$ with which the posterior can be equated once the event E occurs. The present procedure remains silent on how this crucial ingredient is to be chosen.

Further, although the constraint rule (33), as an alternative to (5), is obviously more plausible, it is likewise still open to further questioning. Instead of equating our expectations for the future with an average observed in the past the rule tells us to respond to observed events simply by regarding them as completely certain. But why? It seems that in the present constraint rule one simply picks out a salient property of Bayesian posterior probability distributions and declares that posteriors should always be chosen from the class I_E of all distributions sharing that property. But it is not clear why the non-Bayesian posteriors within the constraint set $\mathcal{I}_E$ are more acceptable than the non-Bayesians outside of this class. The set $\mathcal{I}_E$ might still turn out to be too small or too large.

I know of no attempt to motivate the rule (33) except for those that already uniquely lead to Bayesian conditionalization. An interesting discussion however is given by Van Fraassen (1989, p. 320), who conceives of this rule as a voluntary decision of a person to accept whatever is revealed by observation. He argues that deliverances of experience should be *identified* with commands that constrain one's future opinion. In this view experience literally speaks to us, sometimes with "the voice of an angel" (i.e. fully reliable), but always in an imperative mood.

I find this view intriguing and original, but also somewhat puzzling. It is utterly different from the usual way of modeling data as events (i.e. as sets or as propositions). However, I believe it is more readily suited to the case of immediate responses to raw sense data (as the term 'deliverances of experience' suggests) than to the situation for which maximum entropy or statistical inference in general is intended, namely the case where data are scientific reports of experimental observations, of which the die throw is a mere prototype. The interesting question in such situations is how to update one's judgment, not about the data themselves but about something else: the outcome of a next experiment, the validity of a theory, etc. Experimental data are surely not identical to imperatives on our beliefs concerning such issues.

Further, we note that as a generalization of Bayesian conditionalization, maximum entropy inference is not unique. All expressions of the form

$$\tilde{H} = \sum_i \phi(\frac{q_i}{p_i})p_i$$

with concave ϕ have the same property of being maximal when the probabilities q_i and p_i are proportional (see Uffink, 1995). And so, a maximum $\tilde{H}$ principle would yield a similar generalization of Bayesian conditionalization under the constraint rule (33).

Finally, we briefly discuss a more general type of constraint (33), introduced by Van Fraassen, in what he called the 'Judy Benjamin problem'. Suppose $S = \{x_1, x_2, x_3\}$ and assume that a constraint is given on the value of a posterior conditional probability[11]

$$\mathcal{I}_{\rm cond} = \{Q \ : \ q(x_1|E) = \alpha\}, \tag{36}$$

where $E = \{x_1, x_3\}$. This type of constraint, again, determines a convex closed set as demanded by the most general formulation of the MREP. In fact the constraint sets can also be characterized by a fixed expectation value. If we write:

$$\begin{aligned} f(x_1) &= -(1-\alpha) \\ f(x_2) &= 0 \\ f(x_3) &= \alpha \end{aligned}$$

the constraint (36) is equivalent to the constraint to put $\langle f \rangle = 0$.

Noting that f takes three values, it would seem that we immediately run into the limitations imposed by the Friedman-Shimony theorem. Note, however, that the present rule differs from that assumed in the previous section because here the function f itself depends on α. As a result, the set of constraints is very different from that of Fig. 2 and the Friedman-Shimony theorem is not applicable. (See Fig. 5.) Nevertheless, under this type of constraints the MREP is again inconsistent with Bayesian conditionalization:

[11]The question what empirical data induces one to adopt this type of constraint is circumvented in van Fraassen's formulation of this problem in a manner characteristic of the above view on the nature of experience. The heroine of the Judy Benjamin story is simply informed what her state of belief should be by a helpful expert voice on her radio.

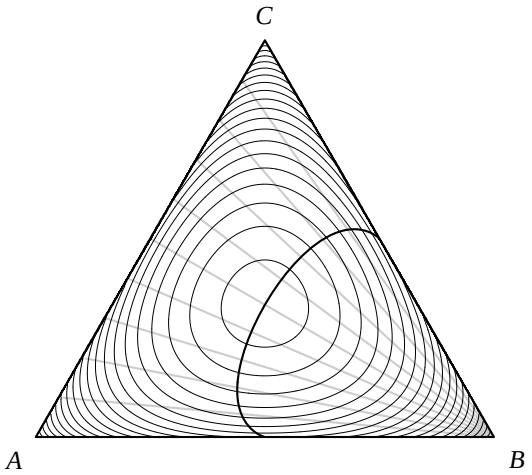


Figure 5: *The maximum entropy distributions for the constraint type proposed by van Fraassen and Seidenfeld.*

Seidenfeld (1986) showed that, starting from an arbitrary prior, the probability of x_2 can only increase when we update : $q(x_2) = (I_{\text{cond}} \circ p)(x_2) > p(x_2)$, whatever the value of α, unless $q = p$. Thus, again, the prior distribution cannot be a proper convex mixture of posteriors, as required by conditionalization. Seidenfeld's result can easily be generalized from Shannon to Rényi entropies of arbitrary order $r \geq -1$. Note however the contrast between this case and the constraint type considered by Friedman and Shimony: in Fig. 2 the probability of x_2 could only *decrease* from its initial value!

7 Summary and discussion

All methods of statistical inference must let empirical data play a role in the statements they wish to make about probability distributions. In the method of maximum entropy the role of the data is to supply a constraint on the set of allowed probability distributions. The question studied here is how, and in what form, data can deliver such a constraint.

It has been my point of departure that the kind of empirical data for which this method is intended do not deliver constraints on probability distributions automatically. Probabilities are not directly observable. If they were, i.e. if we could simply 'read off' the specification of probabilities from observational data, the whole problem of statistical inference would exist no longer as a special field. I take it that a constraint has to be constructed from the data, by a constructive rule.

The constructive rule adopted by Jaynes is to equate observed averages and expectation values of certain specified functions. This is what we called the constraint rule. We have shown that if the MEP is equipped with this constraint rule its prob-

ability assignments can be very different from those obtained by the classical method of inverse probability (or the rule of succession). It has been argued further that the argument Jaynes gave to justify his rule, relying on maximum likelihood estimation, is inconclusive. The argument brings back an echo of old objections against the principle of insufficient reason (cf. Keynes 1973, Fisher, 1973): the probability assignments are sensitive to the choices of scale in the presentation of the empirical data, Thus it appears that even though the Bertrand paradox (Bertrand, 1889) has been exorcized in the maximum entropy method, as discussed in a previous paper (Uffink, 1995), its ghost re-enters through the back door when the constraint rule is adopted.

The problem of choosing a constraint rule, is also of crucial importance in judging the relationship between ME inference and Bayesian conditionalization. If one applies the rule proposed by Jaynes, an inconsistency between these two methods is obtained, by the argument of Friedman and Shimony. However, we have also argued that this mutual inconsistency can be escaped by maintaining a strict distinction between probabilities and events.

A different constraint rule has been proposed by Williams: to regard the observed events as certain, i.e. to put the probability of the observed event equal to one. When this rule is adopted, no contradiction between Bayesian conditionalization and maximum entropy arises. However, in this case much more detailed information needs to be assumed, in the sense that the prior probability distribution must be given on a much wider domain.

It is clear that the choice of a constraint rule is by no means less important in maximum entropy inference than the choice of a mathematical expression to measure entropy. It remains an interesting problem how to characterize viable constraint rules in maximum entropy inference, whether they be of the Shannon or of the Rényi variety. With regard to this problem a further question looms. All the inference methods studied here comply with a format in which some mathematical functional on probability distributions is maximized. The role of the data is restricted *only* to the provision of a constraint. One may well ask on what grounds this format is adopted. This question becomes more pointed when one considers a contrasting example: the estimation method of maximum likelihood. This provides an inference rule that is comparable to maximum entropy in the sense that it also selects a probability distribution in the light of observed empirical data, but in a very different format. Here, one also maximizes a functional of probability distributions: the likelihood function (13). However, this likelihood function explicitly depends on the data, while the set of probability distributions over which the maximization is to be performed is determined by a statistical model, and assumed to be independent of the data. In view of the widespread application and intuitive appeal of this technique, I suspect that a good argument for the idea that empirical data merely act as constraints upon the maximization of a data-independent expression will not be easy to come by.

Appendix

1. Inverse probability applied to the Brandeis dice problem In the approach of inverse probability, one models the Brandeis dice problem by assuming that the die is characterized by an unknown 'true' probability distribution $p = (p_1, \ldots, p_6)$. It is further assumed that our initial ignorance about the values of p_i is represented by a normalized uniform prior probability density over the range of their possible values:

$$\phi(p_1, \ldots, p_6) = 5!\, \delta(\Sigma_i p_i - 1) \quad \text{for} \quad p_i \geq 0, i = 1, \ldots, 6$$

where δ is a Dirac delta. Finally, it is assumed that when the true probabilities are given, the probability of obtaining a sequence of outcomes $x_1, \ldots, x_N$ showing the frequencies $(N_1, \ldots, N_6)$ (with $\sum_i N_i = N$) is:

$$P(N_1, \ldots, N_6 \,|\, p_1 \ldots, p_6) = \frac{N!}{N_1! \cdots N_6!}\, p_1^{N_1} \cdots p_6^{N_6} \tag{37}$$

(This last assumption is a version of Lewis' 'Principal Principle'. See Howson and Urbach, 1989, p. 229.) Now, the theorem of total probability gives

$$\begin{aligned} P(N_1, \ldots, N_6) &= \int\cdots\int P(N_1, \ldots, N_6 \,|\, p_1, \ldots, p_6)\, \phi(p_1, \ldots, p_6)\, dp_1 \cdots dp_6 \\ &= \frac{5!N!}{N_1! \cdots N_6!} \int\cdots\int p_1^{N_1} \cdots p_6^{N_6}\, \delta(\Sigma_i p_i - 1)\, dp_1 \cdots dp_6 \\ &= \frac{5!N!}{(N+5)!} \end{aligned} \tag{38}$$

Where we have used the fact that

$$\int\cdots\int p_1^{N_1} \cdots p_k^{N_k}\, \delta(\Sigma_i p_i - 1)\, dp_1 \cdots dp_k = \frac{N_1! \cdots N_k!}{(N+k-1)!}$$

when the integral is restricted to positive values of p_i. Note that according to (38), for a given N, all possible frequencies $(N_1, \ldots, N_6)$ are equally probable.

Next, one calculates the posterior probability density of the values of $p_1, \ldots, p_6$ given that the frequencies $N_1, \ldots, N_6$ have occurred, using Bayes' theorem:

$$\begin{aligned} \phi(p_1, \ldots, p_6 \,|\, N_1, \ldots, N_6) &= \frac{P(N_1, \ldots, N_6 | p_1, \ldots, p_6)\phi(p_1, \ldots, p_6)}{P(N_1, \ldots, N_6)} \\ &= \frac{(N+5)!}{N_1! \cdots N_6!}\, p_1^{N_1} \cdots p_6^{N_6}\, \delta(\Sigma_i p_i - 1) \end{aligned}$$

From this one easily obtains the conditional probability for the next throw to show i spots under the condition that the previous N throws have shown the frequencies

$(N_1, \ldots, N_6)$:

$$\begin{aligned} P(i_{N+1} = i | N_1, \ldots, N_6) &= \int\cdots\int p_i \, \phi(p_1, \ldots, p_6 | N_1, \ldots, N_6) \, dp_1 \cdots dp_6 \\ &= \frac{(N+5)!}{N_1! \cdots N_6!} \, \frac{(N_i+1)! \prod_{j \neq i} N_j!}{(N+6)!} \\ &= \frac{N_i + 1}{N+6}. \end{aligned} \tag{39}$$

This is the famous rule of succession.

This result is not yet applicable to the Brandeis dice problem, however, where only the average $\frac{1}{N}\sum_i iN_i$ instead of the entire set of frequencies $(N_1, \ldots, N_6)$ is given. For that purpose, one has to calculate:

$$\begin{aligned} P(x_{N+1} = i \,|\, \frac{1}{N}\sum iN_i = \alpha) &= \frac{\sum' P(x_{N+1} = i | N_1, \ldots, N_6) P(N_1, \ldots, N_6)}{P(\frac{1}{N}\sum_i N_i = \alpha)} \\ &= \frac{\sum'(N_i + 1)}{\sum_{k=1}^{6} \sum'(N_k + 1)} \end{aligned} \tag{40}$$

where the prime indicates that the summation is to be performed over all sets of frequencies $(N_1, \ldots, N_6)$, obeying the conditions $N_k \in I\!N$ and $\sum iN_i = \alpha N$. Thus, for example, if $N = 2$ and $\alpha = 4.5$, only two sets of frequencies, (0,0,0,1,1,0) and (0,0,1,0,0,1), obey the conditions; so that $\sum'(N_i + 1) = (2, 2, 3, 3, 3, 3)$ and the probabilities (40) become $(\frac{1}{8}, \frac{1}{8}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}\frac{3}{16})$. In the case considered in section 2, $\alpha = 6$, the result is easy: the primed summation is over one set of frequencies only, and this leads to the result (9). In the case $\alpha = 3.5$ some values of the distribution (40) are tabulated in Table 1.

2. Proof of theorem 1. Here we show that if f takes more than two values the curve joining the points p_β is strictly convex, or in other words, all these points are extreme elements of their own convex hull. Thus, let

$$\mathcal{D} = \{p \,:\, \exists(\gamma_1, \ldots, \gamma_m), \gamma_j \geq 0, \sum_{j=1}^{m} \gamma_j = 1, \text{ such that } p = \sum_{j=1}^{m} \gamma_j p_{\beta_j}\}$$

be the convex hull of the maximum entropy curve. Note that, by virtue of Carathéodory's theorem (cf. Roberts and Varberg, 1973), we may choose $m \leq n$. Now to prove that the points p_β are extreme points of $\mathcal{D}$, we argue as follows. Let

$$p = \sum_{i=1}^{m} \gamma_i p_{\beta_i}$$

be a proper mixture of points from the maximum entropy curve (i.e., all β_i are mutually different and all $\gamma_i > 0$) and consider the point $p_{\bar\beta}$ on the maximum entropy curve where

$$\bar\beta = \alpha^{-1}\left(\sum_i \gamma_i \alpha(\beta_i)\right)$$

where $\alpha(\beta)$ is defined by (26). Obviously, $p_{\bar{\beta}}$ belongs to the same constraint set $\mathcal{I}_\alpha$ (defined by (21)) as p, because

$$\langle f \rangle_p = \sum_i \gamma_i \langle f \rangle_{\beta_i} = \sum_i \gamma_i \alpha(\beta_i) = \alpha(\bar{\beta}) = \langle f \rangle_{\bar{\beta}}. \tag{41}$$

Now each of the mutually disjoint constraint sets contains a unique point from the maximum entropy entropy curve. (A proof of this can be found e.g. in the appendix of (Uffink, 1995). Thus it suffices to show that

$$p \neq p_{\bar{\beta}} \tag{42}$$

in order to conclude that p cannot coincide with any point of this curve.

In order to show (42), consider the function Φ on $\mathcal{D}$ which maps p to the variance of f in the distribution p:

$$\Phi \ : \ p \to \Phi(p) = \langle f^2 \rangle_p - \langle f \rangle_p^2$$

One easily finds, with the help of (41),

$$\Phi(p_{\bar{\beta}}) - \Phi(p) = \langle f^2 \rangle_{\bar{\beta}} - \sum_i \gamma_i \langle f^2 \rangle_{\beta_i}.$$

This will be strictly negative, and *a forteriori* we will have $p_{\bar{\beta}} \neq p$, if one can show that $\langle f^2 \rangle_{\beta(\alpha)}$ is a strictly convex function of α, i.e. if, putting $\alpha_i = \alpha(\beta_i)$,

$$\langle f^2 \rangle_{\beta(\sum_i \gamma_i \alpha_i)} < \sum_i \gamma_i \langle f^2 \rangle_{\beta(\alpha_i)}.$$

To prove this, we must show that the second derivative of $\langle f^2 \rangle$ as a function of α is strictly positive. To simplify the calculation, we note that by subtracting an irrelevant constant from $f(x)$ we can achieve $\alpha(\beta) \equiv \langle f \rangle_\beta = 0$ without loss of generality. Then, using $dp_\beta(x)/d\beta = -f(x)p_\beta(x)$, we find

$$\begin{aligned} \frac{d^2}{d\alpha^2}\langle f \rangle_{\beta(\alpha)} &= \frac{d\beta}{d\alpha}\frac{d}{d\alpha}\frac{d\beta}{d\alpha}\frac{d}{d\beta}\sum_x f(x)\frac{e^{-\beta f(x)}}{Z\beta)} \\ &= \frac{1}{\langle f^2 \rangle^6}(\langle f^2 \rangle \langle f^4 \rangle - \langle f^3 \rangle^2) \end{aligned}$$

Applying the Cauchy inequality $\langle A^2 \rangle \langle B^2 \rangle \geq \langle AB \rangle^2$ to the case $A = f, B = f^2$, one sees that this is non-negative, and in fact strictly positive unless

$$\forall x \in S \ : \ f(x)^2 p_\beta(x) = c f(x) p_\beta(x)$$

for some constant c. Since $p_\beta(x) > 0$ for all $x \in S$ and $\beta \in \mathbb{R}$, this can only occur if $f(x) = c$ whenever $f(x) \neq 0$. i.e. when f takes at most two values.

Acknowledgments

Special thanks to Tim Budden for sharpening my thoughts in many parts of this paper and correcting many mistakes. I also owe much to discussions with Jasper Boessenkool, Dennis Dieks, Fred Muller and Pieter Vermaas. I thank Henk Bos for help in the proof of theorem 1. It is a pleasure to thank Jeremy Butterfield and all other members of the Department for History and Philosophy of Science in Cambridge and Harvey Brown at the Department for Philosophy at the University of Oxford for hospitality and encouragement. This work was supported by a grant from the British Council and the Netherlands Organization for Scientific Research.

References

Bertrand, J. (1889), *Calcul des Probabilités*, (Paris: Gauthier-Villars).

Campenhout, J.M. van and Cover, T.M. (1981), 'Maximum Entropy and Conditional Probability', *IEEE Transactions on Information Theory* **IT-27**, 483–489.

Cyranski, J.F. (1978), 'Analysis of the Maximum Entropy Principle "debate" ', *Foundations of Physics* **8**, 493–506.

Dias, P.M. and Shimony, A. (1981), 'A Critique of Jaynes' Maximum Entropy Principle', *Advances in Applied Mathematics* **2**, 172–211.

Fisher, R.A. (1973), *Statistical Methods and Scientific Inference*, (New York: Hafner Press, 3rd edition).

Fraassen, B.C. van (1981), 'A Problem for Relative Information Minimizers in Probability Kinematics', *British Journal for the Philosophy of Science* **32**, 375–379.

Fraassen, B.C. van (1989), *Laws and Symmetry*, (Oxford: Clarendon Press).

Friedman, K. and Shimony, A. (1971), 'Jaynes's Maximum Entropy Prescription and Probability Theory', *Journal of Statistical Physics* **3**, 381–384.

Friedman, K. (1973), 'Replies to Tribus and Motroni and to Gage and Hestenes', *Journal of Statistical Physics* **9**, 265–269.

Gage D.W. and Hestenes, D. (1973), 'Comment on the Paper "Jaynes's Maximum Entropy Prescription and Probability Theory"' *Journal of Statistical Physics* **7**, 89-90.

Hobson, A. (1972), 'The Interpretation of Inductive Probabilities', *Journal of Statistical Physics* **6**, 189–193.

Howson, C. and Urbach, P. (1989), *Scientific Reasoning*, (La Salle, Illinois: Open Court).

Jaynes, E.T. (1968), 'Prior Probabilities', *IEEE Transactions on Systems Science and Cybernetics* **SSC-4**, (1968), 227–241. Reprinted in (Jaynes 1983), pp. 116–130.

Jaynes, E.T. (1978), 'Where do we stand on maximum entropy?', in R.D. Levine and M. Tribus (eds), *the Maximum Entropy Formalism*, (Cambridge Massachusetts: MIT Press) pp. 15–118. Reprinted in (Jaynes, 1983) pp. 211–314.

Jaynes, E.T (1983), *Probability, Statistics and Statistical Physics*, R. Rosenkrantz (ed) (Dordrecht: Reidel).

Jaynes, E.T. (1985), 'Some random observations', *Synthese* **63**, 115–138.

Jeffrey, R.C (1965), *The Logic of Decision*, (Chicago: Chicago University Press).

Jeffreys, H. (1961), *Theory of Probability*, (Oxford: Clarendon Press 3rd edition).

Keynes, J.M. (1921), *Treatise on Probability, The Collected Works of J.M. Keynes*, Vol. 8, (London: Macmillan).

Lavis, D.A. and Milligan, P.J. (1985), 'The Work of E.T. Jaynes on Probability, Statistics and Statistical Physics', *British Journal for the Philosophy of Science* **36**, 193–210.

Peirce, C.S. (1878), 'The probability of induction' *Popular Science Monthly* **12**, 705–718; also in C. Hartshorne and P. Weiss (eds), *Collected Papers of Charles Sanders Peirce*, Vol.2, (Cambridge Massachusetts: Belknap Press of Harvard College) 1933, §§ 669–693.

Roberts, A.W. and Varberg D.E. (1973), *Convex Functions*, (New York: Academic Press).

Seidenfeld, T. (1979), 'Why I am not an Objective Bayesian', *Theory and Decision* **11**, 413–440.

Seidenfeld, T. (1986), 'Entropy and Uncertainty' *Philosophy of Science* **53**, 467–491.

Shimony, A. (1973), 'Comment on the Interpretation of Inductive Probabilities', *Journal of Statistical Physics* **9**, 187–191.

Shimony, A. (1985), 'The Status of the Principle of Maximum Entropy', *Synthese* **63**, 35–53.

Skyrms, B. (1985), 'Maximum Entropy Inference as a Special Case of Conditionalization', *Synthese* **63**, 55- -74.

Skyrms, B. (1987), 'Updating, Supposing and Maxent' *Theory and Decision* **22**, 225–246.

Tribus M. and Motroni, H. (1972), 'Comments on the Paper "Jaynes's Maximum Entropy Prescription and Probability Theory"' *Journal of Statistical Physics* **4**, 227–228.

Uffink, J. (1990), *Measures of Uncertainty and the Uncertainty Principle*, (Utrecht: Utrecht University).

Uffink, J. (1995), 'Can the maximum entropy principle be explained as a consistency requirement?', (to appear in *Studies in the History and Philosophy of Modern Physics*).

Venn, J. (1866), *The Logic of Chance*, (London: Macmillan).

Williams, P.M. (1980), 'Bayesian Conditionalisation and the Principle of Minimum Information' *British Journal for the Philosophy of Science* **31**, 131–144.