PRIOR PROBABILITIES AND TRANSFORMATION GROUPS

E. T. Jaynes[†]

Washington University, St. Louis, Mo.

Summary: The problem of translating prior information uniquely into a prior probability assignment has heretofore seemed insoluble due to lack of invariance under parameter changes. We show that this ambiguity can be removed by finding the group of operations which transform the problem into an equivalent one, and applying a basic desideratum of consistency. The method is illustrated for the case of location and scale parameters, leading to a procedure for determining prior distributions which is completely "objective" in the sense that it is independent of the parameterization and allows no arbitrary choice on the part of the user.

[†]Supported by the National Science Foundation, Grant No. NSF G23778

1. INTRODUCTION

Since the time of Laplace, applications of probability theory have been hampered by difficulties in the treatment of prior information. In realistic problems of inference, we often have prior information which is highly relevant to the question being asked; to fail to take it into account is to commit the most obvious inconsistency of reasoning, and may lead to absurd or dangerously misleading results. As an extreme example, we might know in advance that a certain parameter $\theta \leq 6$. If we fail to incorporate that fact into our equations, then a conventional statistical analysis might easily lead to the conclusion that the "best" estimate of θ is $\theta^* = 8$, and a shortest 90 percent confidence interval is $(7 \leq \theta \leq 9)$. Yet we do not seem to have the principles needed to translate prior information into a definite prior probability assignment.

The "orthodox" school of thought, represented by most statisticians, seeks to avoid the problem by rejecting the use of prior probabilities altogether, except in the case where the prior information consists of frequency data. However, as the above example shows, this places a great restriction on the class of problems which can be fully treated. Usually, the prior information does <u>not</u> consist of frequency data; but it is nonetheless cogent. As Kendall and Stuart [1] point out, this is a major weakness of the principle of confidence intervals.

The "personalistic" school of thought [2], [3] recognizes this deficiency, but proceeds to overcompensate it by offering us too many priors. Surely, the most elementary requirement of consistency demands that two persons with the same relevant prior information should assign the same prior probabilities. One stands aghast at the fact that personalistic doctrine makes no attempt to meet this desideratum, but instead attacks it as representing a naive "necessary" view of probability; and proclaims as one of its fundamental tenets (Ref. [2], p. 3) that we are free to violate it without being unreasonable!

2. THE BASIC DESIDERATUM

Let us belabor the point just made. A prior probability assignment not based on frequencies is necessarily "subjective" in the sense that it describes a state of knowledge, rather than anything which could be measured directly in an experiment. But if our methods are to have any relevance to science, the prior distribution must be completely "objective" in the sense that it is independent of the personality of the user; i.e., it should describe the prior information, and not anybody's personal feelings. To introduce prior probabilities which everyone is free to choose arbitrarily according to his own fancy, is hardly an advance over orthodox practice.

Evidently, we need to find a middle ground between

the orthodox and personalistic approaches, which will give us just one prior distribution for a given state of prior knowledge. Historically, orthodox rejection of Bayesian methods was not based at first on any ideological dogma about the "meaning of probability," and certainly not on any failure to recognize the importance of prior information; this has been noted by Kendall and Stuart [1], Lehmann [4], and many other orthodox writers. The really fundamental objection (stressed particularly in the remarks of E. S. Pearson in Ref. [3]) was the lack of any principle by which the prior probabilities could be made "objective" in the aforementioned sense. We Bayesians must concede that this is a very sound objection and that Bayesian methods, for all their advantages, will not be entirely satisfactory until we face the problem squarely and show how this requirement may be met.

For later purposes it will be convenient to state this basic desideratum as follows: in two problems where we have the same prior information, we should assign the same prior probabilities. This is stated in such a way that it seems psychologically impossible to quarrel with it; indeed, it may appear so trivial as to be without useful content. The main purpose of the present paper is to show that, in spite of first appearances, this desideratum may be formulated mathematically in a way which has nontrivial consequences.

We are not entirely without clues as to how this uniqueness problem might be solved, at least in some cases.

The principle of maximum entropy (i.e., the prior probability assignment should be the one with the maximum entropy consistent with our prior information) gives us a definite rule for setting up priors, which is impersonal and has an evident intuitive appeal [5], [6], [7], which indicates that it does accomplish the purpose of assigning a prior. In practice, we find that it is easy to apply, and leads to useful results [8], [9] that could be obtained otherwise only by the most awkward and artificial devices.

The application of this principle to the case of continuous parameters is, however, ambiguous because the results depend on our choice of parameters. More generally, all prior probability assignments to continuous parameters, whether based on maximum entropy or not, suffer from this same ambiguity. We have not, heretofore, had any "objective" criterion telling us which parameterization to use.

Since this same difficulty confronts us already in the problem of expressing "complete ignorance," we may hope to make progress by considering this simpler, but still unsolved, problem. Bayes suggested, in one particular case, that we assign a uniform prior probability density; and the domain of useful application of this rule is certainly not zero, for Laplace was led to some of the most important discoveries in celestial mechanics by using it in analysis of astronomical

data. However, Bayes' rule has the obvious difficulty that it is not invariant under a change of parameters.

Jeffreys [10], [11] suggested that we assign a prior $d\sigma/\sigma$ to a continuous parameter σ known to be positive, on the grounds that we are then saying the same thing whether we use the parameter σ or σ^m. Such a desideratum is surely a step in the right direction; however, it cannot be extended to more general parameter changes. We do not want (and obviously cannot have) invariance of the form of the prior under all parameter changes. What we want is invariance of content; but the rules of probability theory already determine how the prior must transform, under any parameter changes, so as to achieve this.

The real problem, therefore, must be stated rather differently; we suggest that the proper question to ask is: For which choice of parameters does a given form of prior distribution such as that of Bayes or Jeffreys; or a given principle such as maximum entropy, apply? Our parameter spaces seem to have a mollusk-like quality which prevents us from answering this, unless we can find some new principle which gives them a property of "rigidity".

Stated in this way, we recognize that problems of just this type have already appeared, and have been solved, in other branches of mathematics. In Riemannian geometry and General Relativity theory, we allow arbitrary continuous coordinate transformations; yet the property of "rigidity" is

maintained by the concept of the invariant line element, which enables us to make statements of definite geometrical and physical meaning, independently of our choice of coordinates. In the theory of continuous groups, the group parameter space had just this mollusk-like quality until the introduction of the concept of invariant group measure, by Hurwitz [12] and Haar [13], [14]. We seek to do something very similar to this for the parameter spaces of statistics.

In the following section we give a very elementary argument which shows that the concept of "complete ignorance" may be defined precisely by specifying the transformation group of the problem. The above basic desideratum of consistency may then be stated mathematically in the form of functional equations which must be satisfied by the prior distributions and which, at least in some cases, uniquely determine the form of the prior.

3. TRANSFORMATION GROUPS

We sample from a continuous two-parameter distribution

$$p(dx|\mu,\sigma) = \phi(x,\mu,\sigma)\ dx \qquad (1)$$

and consider:

Problem A: Given a sample $\{x_1 \ldots x_n\}$, estimate μ and σ.

The problem is indeterminate, both mathematically and conceptually, until we introduce a definite prior distribution

$$f(\mu,\sigma)\, d\mu d\sigma, \tag{2}$$

but if we merely specify "complete initial ignorance", this does not tell us which function $f(\mu,\sigma)$ to use.

Suppose we carry out a change of variables to the new quantities $\{x',\mu',\sigma'\}$ according to

$$\begin{aligned} \mu' &= \mu+b \\ \sigma' &= a\sigma \\ x'-\mu' &= a(x-\mu) \end{aligned} \tag{3}$$

where $0 < a < \infty$, $-\infty < b < \infty$. The distribution (1) expressed in the new variables is

$$p(dx'|\mu',\sigma') = \Psi(x',\mu',\sigma')\, dx' = \phi(x,\mu,\sigma)\, dx$$

or from (3),

$$\psi(x',\mu',\sigma') = a^{-1}\cdot\phi(x,\mu,\sigma) \tag{4}$$

Likewise, the prior distribution is changed into a new one $g(\mu',\sigma')$, where from the jacobian of the transformation (3),

$$g(\mu',\sigma') = a^{-1} f(\mu,\sigma). \tag{5}$$

The above relations will hold whatever the distributions $\phi(x,\mu,\sigma)$, $f(\mu,\sigma)$.

Now suppose the distribution (1) is invariant under the group of transformations (3), so that ψ and ϕ are the same function:

$$\psi(x,\mu,\sigma) = \phi(x,\mu,\sigma) \tag{6}$$

whatever the values of a, b. The condition for this invariance is that $\phi(x,\mu,\sigma)$ must satisfy the functional equation

$$\phi(x,\mu,\sigma)=a\phi(ax-a\mu+\mu+b,\mu+b,a\sigma) \tag{7}$$

Differentiating with respect to a, b and solving the resulting differential equation, we find that the general solution of (7) is

$$\phi(x,\mu,\sigma) = \frac{1}{\sigma} \; h\left(\frac{x-\mu}{\sigma}\right) \tag{8}$$

where h(q) is an arbitrary function. Thus the usual definition of a location parameter μ and a scale parameter σ is equivalent to specifying that the distribution shall be invariant under the group of transformations (3).

What do we mean by the statement that we are "completely ignorant" of μ and σ except for the knowledge that μ is a location parameter and σ is a scale parameter? To answer this, we might reason as follows. If a change of scale can make the problem appear in any way different to us, then we were <u>not</u> completely ignorant; we must have had some kind of information about the absolute scale of the problem. Likewise, if a shift of location can make the problem appear in any way different, then we must have had some prior information about location. In other words, "complete ignorance" of a location and scale parameter is a state of knowledge such that <u>a change of scale and shift of location does not change that</u>

state of knowledge. We shall presently have to state this more carefully, but first let us see its consequences. Consider, therefore,

Problem B: Given a sample $\{x_1' \ldots x_n'\}$, estimate μ' and σ'

If we are "completely ignorant" in the above sense, then we must consider A and B as entirely equivalent problems; they have identical sampling distributions, and our state of prior knowledge about μ' and σ' in problem B is exactly the same as for μ and σ in problem A.

Our basic desideratum now acquires a nontrivial content; for we have formulated two problems in which we have the same prior information. Consistency demands, therefore, that we assign the same prior probability distribution in them. Thus, f and g must be the same function:

$$f(\mu,\sigma) = g(\mu,\sigma) \tag{10}$$

whatever the values of (a,b). But the form of the prior distribution is now uniquely determined; for combining Equations (3), (5), and (10), we see that $f(\mu,\sigma)$ must satisfy the functional equation

$$f(\mu,\sigma) = a\, f(\mu + b,\, a\sigma) \tag{11}$$

whose general solution is

$$f(\mu,\sigma) = \frac{(\text{const.})}{\sigma} \tag{12}$$

which is the Jeffreys rule!

We must not jump to the conclusion that the prior (12) has been determined by the form (8) of the population. Indeed, it would be very disconcerting if the form of the prior were determined merely by the form of the population from which we are sampling; any principle which led to such a result would be suspect. Examination of the above reasoning shows, however, that the result (12) was uniquely determined by the transformation group (3); and not by the form of the distribution (8).

To illustrate this, note that there is more than one transformation group under which (8) is invariant. In the transformations (3) we carry out a change of scale by a factor "a" and a translation b. Denoting this operation by the symbol (a,b), we can carry out the transformation (a_1,b_1), then (a_2,b_2); and from (3) obtain the composition law of group elements:

$$(a_2,b_2)(a_1,b_1) = (a_2a_1, b_2+b_1) \tag{13}$$

Thus the group (3) is Abelian, the direct product of two one-parameter groups. It has a faithful representation in terms of the matrices

$$\begin{pmatrix} a & o \\ o & e^b \end{pmatrix} \tag{14}$$

Now consider the group of transformations in which we first carry out a change of scale "a" on all quantities;

and follow this by a translation b. This group is given by

$$\begin{aligned} \mu' &= a\mu+b \\ \sigma' &= a\sigma \\ x' &= ax+b \end{aligned} \tag{3'}$$

These transformations have the composition law

$$(a_2,b_2)(a_1,b_1) = (a_2a_1, a_2b_1+b_2) \tag{13'}$$

and so the group (3') is non-Abelian; it has a faithful representation in terms of the matrices

$$\begin{pmatrix} a & b \\ o & 1 \end{pmatrix} \tag{14'}$$

which cannot be reduced to diagonal form. Therefore, (3) and (3') are entirely different groups.

If we specify the transformation group (3') instead of (3), equations (5) and (7) are modified to

$$g(\mu',\sigma') = a^{-2} f(\mu,\sigma) \tag{5'}$$

$$\phi(x,\mu,\sigma) = a\phi(ax+b, a\mu+b, a\sigma) \tag{7'}$$

But we find that the general solution of (7') is also (8); and so both groups define location and scale parameters equally well. However, their consequences for the prior are different; for the functional equation (11) is modified to

$$f(\mu,\sigma) = a^2 f(a\mu+b, a\sigma) \tag{11'}$$

whose general solution is

$$f(\mu,\sigma) = \frac{(\text{const.})}{\sigma^2} \qquad (12')$$

Thus, the state of knowledge which is invariant under the group (3) is not the same as that which is invariant under (3'); and we see a new subtlety in the concept of "complete ignorance". In order to define it unambiguously, it is not enough to say merely, "A change of scale and shift of location does not change that state of knowledge". We must specify the precise manner in which these operations are to be carried out; i.e. we must specify a definite group of transformations.

We thus face the question: which group, (3), or (3'), or perhaps some other, defines the state of prior knowledge which we have, in realistic problems, about location and scale parameters? In spite of several attempts, I have not been able to invent any problem in which I feel that (3') really describes the prior information. The difficulty with (3') lies in the equations $x' = ax + b$, $\mu' = a\mu + b$; thus the change of scale operation is to be carried out about two points denoted by $x = o, \mu = o$. But, if we are "completely ignorant" about location, then the condition $x = o$ has no particular meaning; what determines this fixed point about which the change of scale is to be carried out?

In every problem which I have been able to imagine, it is the group (3); and therefore the Jeffreys prior probability rule, which seems appropriate. Here the change of

scale involves only the difference $(x - \mu)$; thus it is carried out about a point which is itself arbitrary; and so no "fixed point" is defined by the group (3). However, it will be interesting to see whether others can produce examples in which the point $x = 0$ always has a special meaning, justifying the stronger prior (12).

To summarize: if we merely specify "complete initial ignorance," we cannot hope to obtain any definite prior distribution, because such a statement is too vague to define any mathematically well-posed problem. We are defining this state of knowledge far more precisely if we can specify a set of operations which we recognize as transforming the problem into an equivalent one. Having found such a set of operations, the basic desideratum of consistency then places nontrivial restrictions on the form of the prior.

Further analysis shows that, if the number of independent parameters in the transformation group is equal to the number of parameters in the statistical problem, the "fundamental domain" of the group [12] reduces to a point, and the form of the prior is uniquely determined; thus specification of such a tranformation group is an exhaustive description of a state of knowledge.

If the number of parameters in the transformation group is less than the number of statistical parameters, the fundamental domain is of higher dimensionality, and the prior

will be only partially determined; for example, if in the group (3')we had specified only the change of scale operation, and not the shift of location, repetition of the argument would lead to the prior

$$f(\mu,\sigma) = \sigma^{-2}\, k(\mu)$$

where $k(\mu)$ is an arbitrary function.

It is also readily verified that the transformation group analysis is consistent with the desideratum of invariance under parameter changes mentioned above; i.e. that while the form of the prior distribution cannot be invariant under all parameter changes, its content should be. If the transformation group (3) or (3') had been defined in terms of some other choice of parameters (α,β), the form of the transformation equations and functional equations would, of course, be different; but the prior to which they would lead in the (α,β)-space would be just the one that we obtain by solving the problem in the (μ,σ)-space and transforming the result to the parameters (α,β) by the usual jacobian rule.

4. DISCUSSION

The above analysis enables us to see the Jeffreys prior probability rule in a new light. It has, perhaps, always been obvious that the real justification of Jeffreys' rule cannot lie merely in the fact that the parameter is positive. As a simple example, suppose that μ is known to be a location parameter; then both intuition and the above analysis agree

that a uniform prior density is the proper way to express "complete ignorance" of μ. The relation $\mu = \theta - \theta^{-1}$ defines a 1:1 mapping of the region $(-\infty < \mu < \infty)$ onto the region $(0 < \theta < \infty)$; but the Jeffreys rule does not apply to the parameter θ, consistency demanding that its prior density be taken proportional to $d\mu = (1 + \theta^{-2})\, d\theta$. It appears that the fundamental justification of the Jeffreys rule is, not merely that a parameter is positive, but that it is a scale parameter.

This also suggests a simple interpretation of some relations between Bayesian and orthodox procedures. Many of us have been surprised to discover how many orthodox procedures involving location and scale parameters lead to results mathematically identical with the Bayesian ones based on the Jeffreys prior. For example, the shortest confidence intervals for the mean or variance of a normal distribution, and the width of a rectangular distribution, are identical with the shortest Bayesian posterior probability intervals at the same level. Likewise, the orthodox F-test and t-test against one-sided alternatives turn out to be identical with the corresponding Bayesian tests, in the following sense: the critical confidence level, at which the null hypothesis is just rejected, is equal to the Bayesian posterior probability that the alternative is true. Thus, in spite of their diametrically opposed philosophies, the two procedures lead us to exactly the same final conclusions. It is curious that so

many textbook authors, after warning the reader against use of the thoroughly discredited Bayesian methods, proceed to choose just these problems to demonstrate the superiority of orthodox methods!

From "empirical" evidence of this sort, it appears to be the rule rather than the exception that, if the orthodox procedure is based on a sufficient statistic, it is mathematically equivalent to the Bayesian procedure based on the Jeffreys prior. In other words, refusal to use prior probabilities at all amounts, mathematically, to the same thing as assigning prior probabilities describing "complete ignorance". Undoubtedly, exceptions to this statement can be found, since the orthodox and Bayesian criteria of performance are so different; nevertheless, this interpretation does seem to make good intuitive sense.

The fact that the prior distributions found above cannot be normalized may be interpreted in two different ways. One can say that it arises simply from the fact that our formulation of the problem of "complete ignorance" was an idealization -- a useful idealization, but one which does not strictly apply in any real problem. A shift of location from a point in St. Louis to a point in the Andromeda nebula; or a change of scale from the size of an atom to the size of our galaxy, does not transform any problem of earthly concern into a completely equivalent one. In practice we will always have some

kind of prior knowledge about location and scale; and in consequence the group parameters (a,b) cannot vary over a truly infinite range. Therefore the transformations (3) do not, strictly speaking, form a group. However, over the range which does express our prior ignorance, the above arguments still apply. Within this range the functional equations, and the resulting form of the prior, must still hold.

In most problems, as is well known, use of non-normalizable priors causes no difficulty; for if the random experiment is providing us with any useful information at all, the likelihood function is such that the posterior distribution vanishes strongly at both ends. In fact, normalization of priors is always unnecessary, because in the application of Bayes' theorem the prior appears both in numerator and denominator, and any normalization constant cancels out.

However, there is a more constructive way of looking at this. Finding the distribution representing "complete ignorance" may be regarded as only the first step in finding the prior for any realistic problem. The "pre-prior" distribution representing complete ignorance does not strictly represent any realistic state of knowledge; but it does define the basic "invariant measure" for our parameter space, without which the problem of finding a realistic prior is mathematically indeterminate. In other words, we cannot answer the question, "What prior distribution represents this

specific prior information?" unless we first learn how to answer, "What prior distribution represents complete ignorance?" Having answered this, the "invariant measure" is known, and application of the principle of maximum entropy to incorporate specific prior information then becomes independent of our choice of parameters. Demonstration of the consistency of this interpretation, and examples of use of the resulting procedure in practical statistical problems, will be deferred to a later article.

5. CONCLUSION

It might be objected that we have not, by these considerations, solved the problem of "complete ignorance". We have merely shifted the problem back to that of choosing some transformation group, and we still lack a completely "impersonal" principle that tells us which one to choose. Indeed, this analysis gives us no reason to think that specifying a transformation group is the only way in which "complete ignorance" may be precisely defined. Furthermore, the procedure suggested here is not necessarily applicable in all problems; and so it remains an open question whether other approaches may be as good or better.

However, before we would be in a position to make any comparative judgments, it would be necessary that some definite alternative procedure be suggested, and that there exist specific problems in which both methods are applicable.

Since such comparisons are not available at present, one can only point out some properties of the method here suggested. The class of problems in which it can be applied is that in which (1) the statement of the problem suggests some definite transformation group, which establishes the invariant measure, representing "complete ignorance", in our parameter space, and (2) in order to apply the principle of maximum entropy, the specific prior information must be of such a nature that, given any proposed prior probability assignment, we can determine unambiguously whether it does or does not agree with that information.

It is difficult to see how any procedure could incorporate prior information which does not satisfy condition (2). We note that satisfying these conditions is, to a large extent, simply a matter of formulating the problem more completely than is usually done.

If these conditions are met, then we have the means for incorporating prior information into the problem, which is independent of our choice of parameters and is completely "impersonal," allowing no arbitrary choice on the part of the user. Few orthodox procedures, and to the best of my knowledge no other Bayesian procedures, enjoy this complete "objectivity." Thus, while the above criticisms are undoubtedly valid, I think it will be granted that this analysis does constitute an advance in the precision with which we are able

to formulate statistical problems, as well as an extension of the class of problems in which statistical methods are useful. The fact that this has proved possible gives hope that further work along these lines -- directed in particular toward learning how to formulate statistical problems so that condition (1) is satisfied -- may yet lead to the final solution of this ancient but vital puzzle; and thus achieve full "objectivity" for Bayesian methods.