Bayesianism and the Fixity of the Theoretical Framework

by Donald Gillies, King's College London

Abstract

This paper compares the Bayesian with the classical approach to statistics. It is argued that the Bayesian approach works only if new evidence does not alter the framework of theoretical assumptions. This thesis is illustrated by two examples. The first is an investigation carried out by Neyman, in which the results of statistical tests led him to abandon one of his initial assumptions, and produce a better model based on a different assumption. It is argued that it would be hard to produce this pattern of reasoning within Bayesianism. The second example is De Finetti's use of exchangeability. It is argued that this gives reasonable answers if the process under consideration consists of independent events, but can go drastically wrong if the process is e.g. a Markov chain. This shows that, as regards many processes whose nature is not exactly known, statistical testing using the classical methodology is essential.

Contents

1. Introduction. Bayesianism versus Classical Statistics

Bayesianism is a powerful current of thought in quite a number of different areas, which include: artificial intelligence, decision theory, economics, philosophy of science, and statistics. In the present paper, I will deal only with Bayesianism in statistics. In fact since the beginning of this century, the principal controversy within statistics has been between

Bayesianism and the so-called classical statistics. I will begin therefore by attempting to characterise, in outline at least, these two approaches to statistics.

Let us start with Bayesianism. When this is applied in a particular problem situation, it is usually assumed that there is a given a set of possible statistical hypotheses H_θ where $\theta \varepsilon I$, for some set I, normally an interval of the real line. Some data or evidence e say is collected and the problem is to judge the hypotheses in the light of this evidence. To do this, the parameter θ is given a prior probability distribution $p(\theta)$ say. This represents the degree of belief of the statistician that θ has various values before the evidence e is considered. Given $p(\theta)$, the posterior probability distribution given e, i.e. $p(\theta \mid e)$ is then calculated using Bayes' Theorem. Our Bayesian statistician now adjusts his or her beliefs from $p(\theta)$ to $p(\theta \mid e)$, a process known as *Bayesian conditionalisation*. The merits of the various hypotheses H_θ are now judged using $p(\theta \mid e)$. Statistical inference on this account consists essentially of a change from a set of beliefs represented by $p(\theta)$ to another set represented by $p(\theta \mid e)$, or, to put it another way by a change of beliefs brought about by Bayesian conditionalisation.

While the concept of change of belief lies at the heart of Bayesianism, the corresponding concept for classical statistics is, in my view. that of hypothesis testing. I regard statistical tests as the core of classical statistics. This means that classical statistics, despite being allegedly 'classical', is in reality much more recent than Bayesianism. Bayesianism began with the publication of the famous paper of Bayes and Price in 1763. It received a powerful mathematical development from Laplace in his 1812. By contrast classical statistics can be dated from 1900 because, in a paper published that year, Karl Pearson introduced the χ^2 test – the first really important and widely used statistical test. Further statistical tests and a theory of statistical testing were subsequently developed by the founders of classical statistics 'Student' (W.S.Gosset), R.A.Fisher, E.S.Pearson, and J. Neyman. The methodology of classical statistics is essentially that of testing. Statistical hypothesis are put forward tentatively to explain observed data, and are then subjected to statistical tests. If they pass these tests, they continue to be held. If they fail the tests, they have to be abandoned or modified. The method here is that of conjectures and refutations as advocated by Popper in his 1963.

Having stated what I see as the difference between Bayesianism and classical statistics, I will now outline the criticism of Bayesianism which I wish to develop in this paper. It is not a criticism which attempts to show that Bayesianism is wrong in all circumstances. Indeed there are some situations where a Bayesian analysis seems to me quite correct – for just one example see my joint paper with Phil Dawid, published in 1989. What the argument seeks to do is to place a limit on the situations in which Bayesianism should be applied. Roughly the thesis is that Bayesianism can be validly applied only if we are in a situation in which there is a fixed and known theoretical framework which it is reasonable to suppose will not be altered in the course of the investigation.[1] I call this the condition of *the fixity of the theoretical framework*. For Bayesianism to be appropriate, the framework of general laws and theories assumed must not be altered during the procedure of belief change in the light of evidence. If this framework were altered at any stage, this could lead to changes in the probabilities which were made not in accordance with Bayes theorem and Bayesian conditionalisation. It follows that, if we are studying a process whose nature is not well known, statistical testing using the methodology of classical statistics is essential.

I will try both to elaborate this criticism and to render it plausible by considering two examples, one in each of the next two sections. The first of these examples is of an investigation carried out by an eminent classical statistician, and the second by an eminent Bayesian. Our eminent classical statistician is Jerzy Neyman, and his investigation was into the distribution of larvae in the plots of an experimental field. I will try to show in section 2 that this investigation was an admirable one, and that its success depended crucially on the use of the methodology of testing employed in classical statistics. Our eminent Bayesian is Bruno De Finetti, and in section 3 I will consider his use of exchangeability in his 1937. I will argue that this gives reasonable answers if the process under study consists objectively of independent events, but can go drastically wrong if the process is e.g. a Markov chain. This shows that we have to be very sure of the correctness of our theoretical framework (in this case that the process consists of independent events) before applying Bayesianism. After going through these two examples, I will conclude the paper in section 4, by considering some ways in which Bayesianism might be defended against the criticisms presented.

2. **An Investigation of Neyman's**

Neyman describes his investigation in his 1952, 33-7. His account begins as follows (1952, 33):

> 'Problems of pest control led to studies of the distribution of larvae in small plots. An experimental field planted with some crop is divided into a number of small plots, …. Then all the larvae found in each plot are counted. Naturally the number of larvae varies considerably from one plot to another.'

Neyman wanted to find a mathematical model which would account for this variation. The first such model which suggested itself to him was the Poisson distribution, according to which the probability of there being a number n of larvae in a small plot (p_n) is given by $p_n = \exp(-\lambda)\ \lambda^n/n!$ for some value of the parameter λ. In a loose sense this corresponds to the assumption that the larvae are distributed randomly throughout the field. It was thus a very plausible hypothesis, and indeed Neyman says explicitly that it was (1952, 33) '… one strongly suggested by intuition.' Neyman had moreover used the same hypothesis of a Poisson distribution for a very similar problem concerned with the distribution of bacteria on a Petri-plate, and there it had proved very successful. Despite these favourable a priori indications, Neyman followed the methodology of classical statistics by subjecting the hypothesis of a Poisson distribution to a series of tests, and, rather surprisingly, these showed that the hypothesis was false.

In his 1952, Neyman gives the results of 5 trials of the Poisson distribution hypothesis. In each case this hypothesis was subjected to a χ^2 test. In one case the test resulted in a confirmation with a value of χ^2 of 4.0 with 2 degrees of freedom, corresponding to 13.5%. The remaining four tests, however, were clear refutations with χ^2 values corresponding to around 0.1% or less, resulting in falsifications even at a 1% level of significance. There could be no doubt in the light of these results that the hypothesis of a Poisson distribution was incorrect. As Neyman says (1952, 34):

> 'In all cases, the first theoretical distribution tried was that of Poisson. It will be seen that the general character of the observed distribution is entirely different from that of Poisson. There seems to be no doubt but that a very serious divergence exists between the actual phenomenon of distribution of larvae and the machinery

> assumed in the mathematical model. When this circumstance was brought to my attention by Dr. Beall, we set out to discover the reasons for the divergence.'

As the last sentence of the quotation shows, Neyman did not consider any hypotheses other than that of the Poisson distribution until after the Poisson distribution hypothesis had been refuted by statistical tests. As so often in science, it was the *falsification* of a hypothesis which stimulated theoretical reasoning. This point will be important when we consider how this case might be analysed from the Bayesian point of view. Let us now see how Neyman continued with the investigation. He describes his next steps as follows (1952, 34-5):

> ' … if we attempt to treat the distribution of larvae from the point of view of Poisson, we would have to assume that each larva is placed on the field independently of the others. This basic assumption was flatly contradicted by the life of larvae as described by Dr. Beall. Larvae develop from eggs laid by moths. It is plausible to assume that, when a moth feels like laying eggs, it does not make any special choice between sections of a field planted with the same crop and reasonably uniform in other respects. Therefore, as far as the spots where a number of moths lay their eggs is concerned, it is plausible that the distribution of spots follows a Poisson Law of frequency, depending on just one parameter, say *m*, representing the average number of spots per unit area.
>
> However, it appears that the moths do not lay eggs one at a time. In fact, at each "sitting" a moth lays a whole batch of eggs and the number of eggs varies from one cluster to another. Moreover, by the time the counts are made the number of larvae is subject to another source of variation, due to mortality.
>
> After hatching in a particular spot, the larvae begin to look for food and crawl around. Since the speed of their movements is only moderate, it is obvious that for a larva to be found within a plot, the birthplace of this larva must be fairly close to this plot. If one larva is found, then it is likely that the plot will contain more than one from the same cluster.'

It is worth noting here that in his attempt to find a new better hypothesis to describe the distribution of the larvae, Neyman made use of background knowledge about the larvae obtained from the domain expert, Dr Beall. This led him to suppose that the larvae would be distributed in clusters round points where batches of eggs had been laid. The points where the eggs were

laid would follow a Poisson distribution, but not the larvae themselves. Neyman produced a mathematical model of this situation which led to the conclusion that the larvae would be distributed in what he called a 'Type A distribution' depending on two parameters. Using the same data as before, Neyman again applied the χ^2 test in the 5 cases, and this time all the tests confirmed the hypothesis. Neyman had clearly succeeded in explaining a surprising experimental finding, and his successful investigation shows the merits of classical statistics, or, what is the same thing, Popper's methodology of conjectures and refutations applied using statistical tests to obtain the refutations.

Neyman himself observes (1952, 37): 'In this example, in order to have agreement between the observed and predicted frequencies, it was imperative to adjust the mathematical model.' Moreover far from being dogmatic about his new Type A distribution, he is anxious to point out that it, like the Poisson distribution, has its limitations. Indeed he says (1952, 37):

> '… there are organisms (e.g., scales) whose distribution on units of area of their habitat does not conform with type A. An investigation revealed that the processes governing the distribution of these organisms were much more complex than that described and therefore, if a statistical treatment is desired, a fresh effort to construct an appropriate mathematical model is necessary.'

That concludes my account of Neyman's investigation of the distribution of larvae in a field, and I now turn to the question of whether a Bayesian statistician could have carried out this investigation as successfully as the classical statistician Neyman. I do not see how this could have been possible. A Bayesian would start in the same way by formulating a set of possible hypotheses H_λ where $0 < \lambda < \infty$. Here H_λ is just the Poisson distribution with parameter λ. The next step would have been to set up a prior probability distribution $p(\lambda)$ representing the Bayesian statistician's prior degree of belief in the various hypotheses. This would have been changed in the light of the evidence e to a posterior distribution $p(\lambda \mid e)$. Yet it is difficult to see how all these changes in degrees of belief by Bayesian conditionalisation could have produced the solution to the problem, namely a Type A distribution. The Bayesian mechanism seems capable to doing no more than change the statistician's degree of belief in particular values of λ. This illustrates very nicely my thesis that Bayesianism requires the fixity of

the theoretical framework. The theoretical framework at the beginning of the investigation was the assumption of a Poisson distribution. If this framework had been adequate, as it was in the example of bacteria in a Petri-plate, then Bayesianism would have dealt with the problem satisfactorily. However the theoretical framework was not adequate for the example of larvae in a field. It had to be changed from the assumption of a Poisson distribution to that of a Type A distribution, and the procedure of Bayesian conditionalisation cannot cope with such a change in belief.

To this it might be objected by a Bayesian that the initial set of possible hypotheses should have included both Poisson distributions and Type A distributions. If this had been done, then Bayesian conditionalisation would have dealt with the problem in a perfectly satisfactory manner. However, the difficulty with this proposal is that, as already pointed out, Neyman only thought of his Type A distribution after the assumption of a Poisson distribution had been refuted by a series of χ^2 tests. Neyman certainly did not consider Type A distributions as an a priori possibility at the beginning of the investigation. Indeed Type A distributions did not exist in the literature of probability and statistics at the beginning of Neyman's investigation. It was his analysis of the particular problem with the help of the domain expert Mr Beall, which caused Neyman to introduce Type A distributions for the first time. Moreover it was only the stimulus provided by the falsification of his initial hypothesis which led Neyman to carry out the subtle analysis which led him to formulate the Type A distribution.

A persistent defender of Bayesianism might still argue that a proper analysis of the problem at the beginning of the investigation could have led to the introduction of the Type A distribution at that stage. I rather doubt whether this is a serious possibility, but let us suppose for the moment that it is. The methodology corresponding to this approach would be for the Bayesian statistician to begin with a lengthy analysis of the problem, consulting domain experts, and introducing all the various distributions which might be relevant. While the views of Dr Beall suggested the Type A distributions, the views of other domain experts, since domain experts often disagree, might have suggested further possible distributions, say distributions of types B, C, and D. Moreover distributions other than Type A are sometimes necessary for problems of this kind, as Neyman's discussion of the distribution of scales, quoted earlier, shows. The Bayesian could then formulate his prior belief distribution over all these hypotheses, and proceed from there. Unfortunately such an approach could very often

prove a complete waste of time. Suppose a Bayesian statistician had tried such an approach on the example of bacteria on a Petri-plate. By the time he or she had formulated the first one or two of the hypothetical new distributions which might be possible, Neyman would already have confirmed by a series of χ^2 tests that the simple Poisson distribution was quite adequate in this case. This shows how easy and straightforward is the methodology of classical statistics. It allows us to start with a simple conjecture such as the Poisson distribution, provided only we obey the golden rule of testing our conjecture severely. If the conjecture passes our tests, then it can be accepted provisionally until some further investigations suggest the need for a modification. In the interim we have found a workable hypothesis without the need for elaborating a whole series of possible alternatives. Since Bayesianism depends on the fixity of the theoretical framework, Bayesian statisticians are faced with an awkward choice. Either they must, at the very beginning of the investigation, consider a whole series of arcane possible hypotheses, or they must risk never subsequently arriving at the hypothesis which constitutes the solution of the problem. Their difficulty here arises from the very essence of Baysianism, namely its limitation of changes of belief to those produced by Bayesian conditionalisation.

There are some further ways in which Bayesianism might be defended in the context of this particular example, but it will be convenient to postpone their consideration until section 4, and proceed in the next section to give my second example. In the first example, I have tried to show the merits of the methodology of classical statistics when applied by a leading classical statistician. In the second example I will move in the opposite direction by giving an analysis by a leading Bayesian, namely De Finetti's use of exchangeability, and then trying to show that this analysis only give satisfactory results if no changes are needed in the theoretical framework which is implicitly assumed.

3. **De Finetti on Exchangeability**

In Chapter III of his 1937, De Finetti poses the question (118): 'Why are we obliged in the majority of problems to evaluate a probability according to the observation of a frequency?', commenting that this question (119): 'includes in reality the problem of reasoning by induction.' He continues (119):

> 'In order to fix our ideas better, let us imagine a concrete example, or rather a concrete interpretation of the problem, which does not restrict its generality at all. Let us suppose that the game of heads or tails is played with a coin of irregular appearance.'

We will now explain how De Finetti analyses this example of the biassed coin from his subjective Bayesian point of view. It will emerge that this concrete example does, in a significant respect, fail to represent the full generality of the problem of reasoning by induction.

De Finetti's first step is to consider a sequence of tosses of the coin which we suppose gives results: $E_1, \ldots, E_n, \ldots$, where each E_i is either heads (H_i) or tails (T_i). So, in particular, H_{n+1} = Heads occurs on the n+1th toss. Further let e be a complete specification of the results of the first n tosses, that is a sequence n places long, at the ith place of which we have either H_i or T_i. Suppose that heads occurs r times on the first n tosses. The subjective Bayesian's method is to calculate $P(H_{n+1}| e)$, and to show that under some general conditions which will be specified later $P(H_{n+1}| e)$ tends to r/n for large n. This shows that whatever value is assigned to the prior probability $P(H_{n+1})$, the posterior probability $P(H_{n+1}| e)$ will tend to the observed frequency for large n. Thus different individuals who may hold widely differing opinions initially will, if they change their probabilities by Bayesian conditionalisation, come to agree on their posterior probabilities. Such is the argument. Let us now give, in our simple case, the mathematical proof which underpins it.

Suppose that $P(E_i) \neq 0$ for all i, so that also $P(e) \neq 0$. We then have by the definition of conditional probability

$$P(H_{n+1}| e) = \frac{P(H_{n+1} \& e)}{P(e)} \qquad (1)$$

To proceed further we introduce the condition of *exchangeability*. Suppose Mr B is making an a priori bet that a particular n-tuple of results ($E_{i1}\ E_{i2} \ldots E_{in}$ say) occurs. Suppose further that heads occurs r times in this n-tuple. Mr B's betting quotients are said to be *exchangeable* if he assigns the same betting quotient to any other particular n-tuple of results in which heads occurs r times, where both n and r can be chosen to have any finite integral non-negative values with $r \leq n$. Let us write his prior probability (or betting

quotient) that there will be r heads in n tosses as $\omega_r^{(n)}$. There are ${}^nC_r$ different ways in which r heads can occur in n tosses, where, as usual, ${}^nC_r = \frac{n!}{(n-r)!\ r!} = \frac{n(n-1)...(n-r+1)}{r(r-1)...1}$. Each of the corresponding n-tuples must, by exchangeability, be assigned the same probability, which is therefore $\omega_r^{(n)}/{}^nC_r$. Thus

$$P(E_{i1}\ E_{i2}\ ...\ E_{in}) = \frac{\omega_r^{(n)}}{{}^nC_r} \qquad (2)$$

Now e, by definition, is just a particular n-tuple of results in which heads occurs r times. Thus, by exchangeability,

$$P(e) = P(E_1\ E_2\ ...\ E_n) = \frac{\omega_r^{(n)}}{{}^nC_r} \qquad (3)$$

Now H_{n+1}& e is an (n+1)-tuple of results in which heads occurs r+1 times. Thus, by the same argument,

$$P(H_{n+1}\ \&\ e) = \frac{\omega_{r+1}^{(n+1)}}{{}^{n+1}C_{r+1}} \qquad (4)$$

And so, substituting in (1), we get

$$P(H_{n+1}|\ e) = \frac{{}^nC_r}{{}^{n+1}C_{r+1}}\ \frac{\omega_{r+1}^{(n+1)}}{\omega_r^{(n)}}$$

$$= \frac{n!}{(n-r)!\ r!}\ \frac{(r+1)!\ (n-r)!}{(n+1)!}\ \frac{\omega_{r+1}^{(n+1)}}{\omega_r^{(n)}}$$

$$P(H_{n+1}|\ e) = \frac{r+1}{n+1}\ \frac{\omega_{r+1}^{(n+1)}}{\omega_r^{(n)}} \qquad (5)$$

Formula (5) (which is De Finetti's formula (6), 1937, 122 with a slightly different notation) gives us the result we want. Provided only $\frac{\omega_{r+1}^{(n+1)}}{\omega_r^{(n)}}$ -> 1 as n -> ∞ (a very plausible requirement), we may choose our prior

probabilities $\omega_r^{(n)}$ in any way we please, and still get that, as n -> ∞, $P(H_{n+1}| e)$ -> r/n (the observed frequency), as required.

We can, however, obtain an even simpler result if we choose the prior probablities in a particular way. In n tosses, we can have either 0, 1, 2, ..., or n heads. So, by coherence,

$$\omega_o^{(n)} + \omega_1^{(n)} + \omega_2^{(n)} + ... + \omega_r^{(n)} + ... + \omega_n^{(n)} = 1 \qquad (6)$$

In the subjective theory, we can choose the $\omega_r^{(n)}$ (the prior probabilities) in any way we choose subject only to (6). However we can also, though this is not compulsory, make the 'principle of indifference' choice of making them all equal so that

$$\omega_o^{(n)} = \omega_1^{(n)} = \omega_2^{(n)} =... = \omega_r^{(n)} = ... = \omega_n^{(n)} = 1/(n+1) \qquad (7)$$

Substituting this in (5), we get

$$P(H_{n+1}| e) = \frac{r+1}{n+2} \qquad (8)$$

This is a classical result - Laplace's Rule of Succession, which De Finetti derives in the above way (1937, 144).

In the above calculations, De Finetti appears to show that subjective Bayesians will be led by the process of Bayesian conditionalisation to choose posterior probabilities which approximate to the observed frequency. He thus appears to have provided a foundation for reasoning by induction. I next want to argue that these calculations, despite their seeming generality, are only appropriate within a specific theoretical framework, and can lead us astray if used when that framework does not hold in reality. In order to identify this framework, I will now give some further results from De Finetti's 1937. These relate the concept of *exchangeability*, which De Finetti himself had introduced, to the older concept of *independence*. De Finetti's ideas on the relationship between exchangeability and independence are discussed in Galavotti (2001).

De Finetti proved a general theorem showing exchangeability and independence are linked, I will now state his result. Let us first define exchangeability for a sequence of random variables (or random quantities as

De Finetti prefers to call them) $X_1, \ldots, X_n, \ldots$. These are exchangeable if, for any fixed n, $X_{i1}, X_{i2}, \ldots, X_{in}$ have the same joint distribution no matter how $i_1, \ldots, i_n$ are chosen. Now let Y_n be the average of any n of the random quantities X_i i.e. $Y_n = (1/n)(X_{i1} + X_{i2} + \ldots + X_{in})$, since we are dealing with exchangeable random quantities it does not matter which $i_1, \ldots, i_n$ are chosen. De Finetti first shows (1937, 126) that the distribution $\Phi_n(\xi) = P(Y_n \leq \xi)$ tends to a limit $\Phi(\xi)$ as $n \to \infty$, except perhaps for points of discontinuity. He goes on to say (1937, 128-9):

> 'Indeed, let $P_\xi(E)$ be the probability attributed to the generic event E when the events $E_1, E_2, \ldots, E_n, \ldots$ are considered independent and equally probable with probability ξ; the probability P(E) of the same generic event, the E_i being exchangeable events with the limiting distribution $\Phi(\xi)$, is
>
> $$P(E) = \int_0^1 P_\xi(E)\, d\Phi(\xi).$$
>
> This fact can be expressed by saying that the probability distributions P corresponding to the case of exchangeable events are linear combinations of the distributions P_ξ corresponding to the case of independent equiprobable events, the weights in the linear combination being expressed by $\Phi(\xi)$.'

This general result can be illustrated by taking a couple of special cases. Suppose that we are dealing with a coin tossing example and the generic event E is that heads occurs r times in n tosses. Then

$$P_\xi(E) = {}^nC_r\, \xi^r\, (1 - \xi)^{n-r}$$

So

$$P(E) = \omega_r^{(n)} = {}^nC_r \int_0^1 \xi^r (1-\xi)^{n-r}\, d\Phi(\xi)$$

If, in particular, $\Phi(\xi)$ is the uniform distribution, we have

$$\omega_r^{(n)} = {}^nC_r \int_0^1 \xi^r\, (1 - \xi)^{n-r}\, d\xi$$

$= {}^{n}C_{r}\, B(r+1, n-r+1)$, where B is the Beta function

$= 1/(n + 1)$ (cf. formula 7 above)

Comparing these results with our earlier calculations involving exchangeability, we can see how exchangeability and independence are related.

Roughly speaking we can say that the situation which an objectivist would describe as one of independent events in which particular outcomes have fixed but unknown probabilities corresponds to what De Finetti would describe as one of exchangeable events. Of course De Finetti would not have liked this formulation, since he regarded the 'unknown probabilities' postulated by objectivists and classical statisticians as metaphysical and meaningless. Thus he says (1937, pp. 141-2):

> 'If … one plays heads or tails with a coin of irregular appearance, …, one does not have the right to consider as distinct hypotheses the suppositions that this imperfection has a more or less noticeable influence on the "unknown probability", for this "unknown probability" cannot be defined, and the hypotheses that one would like to introduce in this way have no objective meaning.'

De Finetti therefore concludes (1937, 142):

> ' … the nebulous and unsatisfactory definition of "independent events with fixed but unknown probability" should be replaced by that of "exchangeable events".'

Naturally I cannot agree with De Finetti's attempt to eliminate the concept of unknown probability. To postulate such probabilities, as is done in classical statistics, is neither meaningless nor metaphysical. Conjectures about such unknown probabilities can be tested statistical tests, and either confirmed or refuted, and this shows that such conjectures are scientific rather than metaphysical. It is thus both meaningful and scientific to postulate that a particular process consists of independent events with fixed but unknown probability. My thesis is that this postulate gives the theoretical framework within which De Finetti's calculations using exchangeability lead to sensible results. If we try to use these calculations in situations where this theoretical framework does not hold objectively, they are liable to give absurd and quite inappropriate conclusions. This can be

easily shown by seeing what happens when we apply the exchangeability calculations to a situation which is not one of independent events but of dependent events.

To illustrate my argument, it would be possible to use any one of a wide variety of sequences of events which are dependent rather than independent. To be concrete, I have first selected the simplest type of dependent sequence, namely a Markov chain, and then chosen one very simple and at the same time striking example of a Markov chain. This is the game of 'Red or Blue'.[2] At each go of the game there is a number s which is determined by the previous results. A fair coin is tossed. If the result is heads, we change s to s' = s+1, and if the result is tails, we change s to s' = s-1. If s' $\geq$ 0, the result of the go is said to be blue, while if s' < 0, the result of the go is said to be red. So, although the game is based on coin tossing, the results are a sequence of red and blue instead of a sequence of heads and tails. Moreover, while the sequence of heads and tails is independent, the sequence of red and blue is highly dependent. We would expect much longer runs which are all blue, than runs in coin tossing which are all heads. If we start the game with s = 0, then there is a slight bias in favour of blue which is the initial position. However, it is easy to eliminate this by deciding the initial value of s by a coin toss. If the toss gives heads we set the initial value of s at 0, and if the toss gives tails we set it at -1. This makes red and blue exactly symmetrical, so that the limiting frequency of blue must equal that of red and be 1/2. It is therefore surprising that over even an enormously large number of repetitions of the game, there is high probability of one of the colours appearing much more often than the other. Feller (1950, 82-3) gives a number of examples of these curious features of the game. Suppose for example that the game is played once a second for a year, i.e. repeated 31,536,000 times. There is a probability of 70% that the more frequent colour will appear for a total of 265.35 days, or about 73% of the time, while the less frequent colour will appear for only 99.65 days, or about 27% of the time.

Let us next suppose that a subjective Bayesian (Mr B) is asked to analyse a sequence of events, each member of which can have one of two values. Unknown to them this sequence is in fact generated by the game of red or blue. Possibly the sequence might be produced by a man-made device which flashes either 0 (corresponding to red) or 1 (corresponding to blue) onto a screen at regular intervals. However, it is not impossible that the sequence might be one occurring in the world of nature. Consider for example a sequence of days, each of which is classified as 'rainy' if some rain falls, or dry otherwise. In a study of rainfall at Tel Aviv during the

rainy season of December, January, and February, it was found that the sequence of days could be modelled successfully as a Markov chain. In fact the probabilities found empirically were: probability of a dry day given that the previous day was dry = 0.75, and probability of a rainy day given that the previous day was rainy = 0.66. (For further details see Cox & Miller,1965, 78-9.) It is clear that this kind of dependence will give longer runs of either rainy or dry days than would be expected on the assumption of independence. It is thus not impossible that the sequence of rainy and dry days at some place and season might be represented quite well by the game of red or blue.

Let us now return to our subjective Bayesian Mr B, who has been asked to deal with a process which is really governed, unknown to Mr B, by the game of 'Red or Blue'. Being an admirer of De Finetti's, Mr B will naturally make an assumption of exchangeability. Let us also assume that he gives a uniform distribution a priori to the $\omega_r^{(n)}$ (see formula 7 above) so that Laplace's rule of succession holds (formula 8). This is just for convenience of calculation. The counter-intuitive results would appear for any other coherent choice of the $\omega_r^{(n)}$. Suppose that we have a run of 700 blues, followed by 2 reds. Mr B would calculate the probability of getting blue on the next go using formula 8 with n = 702, and r = 700. This gives the probability of blue as 701/704 = 0.996 to 3 significant figures. Knowing the mechanism of the game, we can calculate the true probability of blue on the next go, which is very different. Go 700 gave blue, and go 701 gave red. This is only possible if s on go 700 was 0, the result of the toss was tails, and s became -1 on go 701. The next toss must also have yielded tails or there would have been blue again on go 702. Thus s at the start of go 703 must be -2, and this implies that the probability of blue on that go is zero. Then again let us consider one of Feller's massive sessions of 31,536,000 goes. Suppose the result is that the most frequently occurring colour appears 73% of the time (as pointed out above there is a probability of 70% of this result which is thus not an unlikely outcome). Mr B will naturally be estimating the probability of this colour at about 0.73 and so much higher than that of the other colour. Yet in the real underlying game, the two colours are exactly symmetrical.

We see that Mr B's calculations using exchangeability will give results at complete variance with the true situation. The reason for this is clear. By making the assumption of exchangeability, Mr B is implicitly assuming that the process he is considering consists of independent events with a fixed but unknown probability. As long as this theoretical framework

holds, his Bayesian calculations will give him reasonable results, but if the theoretical framework does not hold in a particular case, then the same Bayesian calculations will give him completely inappropriate results. My conclusion is, once again, that Bayesianism only works if the condition of the fixity of the theoretical framework is satisfied.

Our situation involving the game of 'Red or Blue' does not pose the same problems for a classical statistician. Suppose such a statistician (Ms C say) is confronted with a sequence of events which, unknown to her, is really governed by the game of 'Red or Blue'. It would be perfectly reasonable for Ms C to begin by making the simplest and most familiar conjecture, namely that the events are independent. Thus Ms C starts tackling the problem in much the same was as Mr B. However, being, unlike Mr B, a good Popperian, Ms C will test her conjecture rigorously with a series of statistical tests for independence. It will not be long before she has rejected her initial conjecture, and she will then start exploring other hypotheses involving various kinds of dependence among the events. If she is a talented scientist, she may soon hit on the red or blue mechanism, and be able to confirm that it is correct by another series of statistical tests. In this case the classical statistician seems better equipped to deal with the problem than the Bayesian. However there are some replies to this argument which could be made from the Bayesian point of view, and I will consider them in the final section of the paper (section 4).

4. **Possible Defences of Bayesianism**

De Finetti himself does say one or two things which are relevant to the problem. Having shown that exchangeable events are the subjective equivalent of the objectivist's independent and equiprobable events, he observes that one could introduce subjective equivalents of various forms of dependent events, and, in particular, of Markov chains. As he says (1937, Footnote 4, 146):

> 'One could in the first place consider the case of classes of events which can be grouped into Markov "chains" of order 1,2, ... , m, ... , in the same way in which classes of exchangeable events can be related to classes of equiprobable and independent events.'

We could call such classes of events *Markov-exchangeable*. De Finetti argues that they would constitute a complication and extension of his theory without causing any fundamental problem (1937, 145):

> ‘One cannot exclude completely *a priori* the influence of the order of events ... . There would then be a number of degrees of freedom and much more complication, but nothing would be changed in the setting up and the conception of the problem ... , before we restricted our demonstration to the case of exchangeable events ...’

Perhaps De Finetti has in mind something like the following. Instead of just assuming exchangeability, we consider not just exchangeability but various forms of Markov-exchangeability. To each of these possibilities we give a prior probability. No doubt exchangeability will have the highest prior probability. If the case is a standard one, like the biased coin, this high prior probability will be reinforced, and the result will come out more or less like that obtained by just assuming exchangeability. If, however, the case is an unusual one, then the posterior probability of exchangeability will gradually decline, and that of one of the other possibilities will increase until it becomes much more probable than exchangeability.

This approach to the problem is basically the same as that we attributed to the Bayesian in our discussion of Neyman’s investigation in section 2, and it is liable to the same difficulties which we noted there. If a Bayesian is to adopt this approach seriously, he or she must begin every investigation by considering all possible hypotheses which might be encountered in the course of the investigation. This is scarcely possible, and, even if it were possible, it would often be a waste of time. There are many situations in which the most obvious and straightforward hypothesis actually works so that a consideration of a large number of arcane alternatives would be useless toil. The classical statisticians do not need to indulge in such toil. They can begin with any assumption (or conjecture) they like, provided only they obey the golden rule of testing it severely. If the assumption passes such tests, it can be provisionally adopted. If it fails, some other better assumption must be sought. Thus the classical statistician proceeds, so to speak, one step at a time, and there is never any need to engage in the hopeless and time-wasting task of surveying all possible hypotheses which might apply to the problem in hand.

There are moreover, as Albert has shown in his contribution to the present volume, further difficulties in this defence of Bayesianism. To see what these are, let us go back to the formulation of Bayesianism given at the beginning of the paper. As I said there, it is usually assumed in a Bayesian statistical analysis that there is a given a set of possible statistical hypotheses

H_θ where $\theta \varepsilon I$, for some set I, normally an interval of the real line. The problem we are now considering is how the set H_θ where $\theta \varepsilon I$ should be chosen. If we select a rather narrow set H_θ we may leave out the hypothesis which would provide the required solution. If we try to make H_θ broad and inclusive, we set ourselves a very difficult task which may well prove a waste of time in a case in which the most simple and obvious solution actually works in practice. What Albert shows in his 2001 is that the second strategy of searching for a broad and inclusive set H_θ is liable to a further difficulty.

Albert considers the possibility of extending the set H_θ by including hypotheses involving chaos theory. Specifically he defines in section 4 of his paper what he calls a 'Chaotic Clock'. In a simple case in which we are considering a sequence of 0's and 1's generated by some unknown process, Albert formulates a set H_θ of hypotheses based on a mechanism involving a chaotic clock. He then gives in section 5.1 of his paper a remarkable result called the *Anything Goes Theorem*. Suppose Mr B adopts any learning strategy whatever, i.e. he chooses his conditional probabilities given evidence in any arbitrary way. There then exists a prior probability distribution p over the set H_θ of hypotheses based on the chaotic clock such that Mr B's probabilities are produced by Bayesian conditioning of p.

Albert's result is very striking indeed. His chaotic clock hypotheses are by no means absurd. After all chaos theory is used in both physics and in economics. Indeed hypotheses involving chaos are quite plausible as a means of explaining, for example, stock market fluctuations. If Mr B were really faced with a bizarre sequence of 0's and 1's, why should he not consider a hypothesis based on chaos theory? Yet if Mr B is allowed to consider the chaotic clock set of hypotheses, then any learning strategy he adopts becomes a Bayesian strategy for a suitable choice of priors. In effect Bayesianism has become empty.

It follows that a Bayesian (Mr B say) is caught on the horns of a dilemma. Mr B may adopt a rather limited set of hypotheses to perform his Bayesian conditionalisation, but then, as the example of the game of Red or Blue shows, if his set excludes the true hypothesis, his Bayesian learning strategy may never bring him close to grasping what the real situation is. This is the first, or 'Red or Blue', horn of the dilemma. If Mr B responds by saying he is prepared to consider a wide and comprehensive set of

hypotheses, these will surely include hypotheses from chaos theory and thus anything he does will become Bayesian, making the whole approach empty. This is the second, or 'Chaotic Clock', horn of the dilemma.

The Bayesian is faced with quite severe difficulties here, but there is one further way out which is sometimes suggested, and I will conclude the paper by giving it a brief consideration. The suggestion is that in we should start with a reasonably specific set of initial hypotheses H_θ but add to this set a 'catch all' hypothesis K, which simply says that some hypothesis other than the H_θ is correct. We then give our prior distribution over the H_θ and K. If it is a standard case, then one of the H_θ will emerge as the most probable hypothesis given the evidence. If, however, we are dealing with a non-standard case, then K will gain in probability while the probability of each of the H_θ becomes very small. In such a situation, we will divide up K into some specific set J_θ say, and a new catch all K', and repeat the process. In this way we should, even in a problematic situation, be led to the correct hypothesis.

While such a procedure sounds very reasonable when stated in outline, any attempt actually to implement it in detail brings to light a whole host of difficulties and complexities, and it is not surprising that there is no instance to my knowledge of such a plan being actually carried out in detail by a Bayesian. Let us begin by considering how the prior probabilities should be divided between the set H_θ and the catch all K. Surely K should have a very large prior probability since our background knowledge concerning the development of science would suggest that most hypotheses considered at a particular time are eventually shown to be inadequate to some degree or in some respects. Yet if K is given a large prior probability, this may prevent any of the H_θ ever acquiring a large probability, even in a straightforward case.

Suppose this initial difficulty is overcome, we are the faced with another. Let us take one of the problematic cases in which we assume to begin with one set of hypotheses H_θ say, and another set J_θ are in fact correct. H_θ could be Poisson distributions and J_θ could be Type A distributions, or H_θ could be the hypothesis of independent events with fixed probability θ and J_θ could be hypotheses of a Markov chain of some type. In this case we have got to show how the probability of the catch all K

changes from its prior value p(K) say to a posterior value p(K | e) in the light of evidence. How is such a calculation to be carried out? It is no easy matter, and it must be done in such a way that p(K | e) increases to such a value that we decide to abandon the H_θ and subdivide K into J_θ and the new catch all K'. I really think such a calculation is scarcely possible. Of course a Bayesian could show that I am wrong by carrying out such a calculation in one of the cases dealt with in this paper, but the result would undoubtedly be very complicated. At this point one can reasonably ask why the Bayesian wants to get involved in such complexities rather than to adopt the methods of classical statistics which, as I have shown, deal with the problem in an extremely simple and straightforward way, using the method of conjectures and refutations.

My conclusion is that Bayesianism should only be applied if we are in a situation in which there is a fixed and known theoretical framework which it is reasonable to suppose will not be altered in the course of the investigation, that is to say if the condition of the fixity of the theoretical framework is satisfied. As regards many processes whose nature is not exactly known, statistical testing using the methodology of classical statistics is essential.

Notes

1. The phrase 'a fixed theoretical framework' comes from Lakatos (1968, 161), although he uses it in a somewhat different sense. Lakatos is criticising Carnap's inductive logic, and points out that Carnap's confirmation function (*c*-function) depends on the language employed so that it cannot cope with changes in the language. Lakatos puts the argument like this (1968, 161):

> 'Although growth of the evidence *within* a fixed theoretical framework (the language L) leaves the chosen *c*-function unaltered, growth of the theoretical framework (introduction of a *new* language L*) may change it radically.'

Lakatos here identifies a theoretical framework with a language. By contrast I am using 'theoretical framework' to refer to the set of theories under consideration. Thus a theoretical framework in my sense changes when a new theory is introduced even though the language does not change. Despite this difference, the general structure of Lakatos' argument is quite similar to that of the argument developed in this paper.

2. The game of 'Red or Blue' is described in Feller,1950, 67-95 which contains an interesting mathematical analysis of its curious properties. Popper read of the game in Feller, and had the idea of using it to argue against various theories of induction. In his 1957, 358-60 (reprinted in his 1983, 301-5) he uses the game to criticize what he calls 'the simple inductive rule'. I have adapted this argument of Popper's to produce the critique of De Finetti's use of exchangeability given here.

References

Albert, M. (2001) 'Bayesian Learning and Expectations Formation: Anything Goes' in present volume, 000-000.

Bayes, T. and Price, R. (1763) 'An Essay towards Solving a Problem in the Doctrine of Chances', reprinted in E.S.Pearson and M.G.Kendall (eds.) *Studies in the History of Statistics and Probability*, Griffin, 1970, 134-53.

Cox, D.R. and Miller, H.D. (1965) *The Theory of Stochastic Processes*. Methuen.

Dawid, P. and Gillies, D.A. (1989) 'A Bayesian Analysis of Hume's Argument concerning Miracles', *The Philosophical Quarterly*, **39**, 57-65.

De Finetti, B. (1937) 'Foresight: Its Logical Laws, Its Subjective Sources'. English translation in H.E,Kyburg and H.E.Smokler (eds.), *Studies in Subjective Probability*, Wiley, 1964, 93-158.

Feller, W. (1950) *Introduction to Probability Theory and Its Applications*. Third edition, 1971, Wiley.

Galavotti, M.C. (2001) 'Subjectivism and Objectivity in Bruno De Finetti's Bayesianism', in present volume, 000-000.

Lakatos, I. (1968) 'Changes in the Problem of Inductive Logic'. Reprinted in John Worrall and Gregory Currie (eds.), *Imre Lakatos, Philosophical Papers, Volume 2*, Cambridge University Press, 128-200.

Laplace, P.S. (1812) *Théorie analytique des probabilités*. Reprinted as vol. 5 of *Oeuvres complètes de Laplace*, 14 vols, Gauthier-Villars, 1878-1912.

Neyman, J. (1952) *Lectures and Conferences on Mathematical Statistics and Probability*. 2nd edition, revised and enlarged, Washington: Graduate School of U.S. Department of Agriculture.

Pearson, K. (1900) 'On the Criterion that a given System of Deviations from the probable in the case of a Correlated System of Variables is such that it can be reasonably be supposed to have arisen from Random Sampling', reprinted in *Karl Pearson's Early Statistical Papers*, Cambridge University Press, 1956, 339-57.

Popper, K.R. (1957) 'Probability Magic or Knowledge out of Ignorance', *Dialectica*, **11**(3/4), 354-74.

Popper, K.R. (1963) *Conjectures and Refutations*. Routledge and Kegan Paul.

Popper, K.R. (1983) *Realism and the Aim of Science*. Hutchinson.