REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 7, 1989	Final Report, 1 Apr 87 to 31 Mar 89

4. TITLE AND SUBTITLE
BAYESIAN NONPARAMETRIC PREDICTION AND STATISTICAL INFERENCE

5. FUNDING NUMBERS
AFOSR-87-0192
61102/ 2304/A5

6. AUTHOR(S)
Burce M. Hill

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Department of Statistics
Ann Arbor, MI 48109-1092

8. PERFORMING ORGANIZATION REPORT NUMBER
AFOSR-87-0192

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
Building 410
Bolling AFB, DC 20332-6448

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFOSR-TR- 90-0211

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The problem of Bayesian nonparametric prediction and statistical inference is formulated and discussed. A solution is proposed based upon A_n and H_k as in Hill (1968). The meaning of parameters in the subjective Bayesian theory of Bruno de Finetti is discussed in connection both with A_n and with conventional parametric models. It is argued that the usual sharp distinction between prediction and parametric inference is largely illusory. The finite version of de Finetti's theorem is emphasized for the practice of statistics, with the infinite case used only to obtain approximations and insight.

14. SUBJECT TERMS

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

BAYESIAN NONPARAMETRIC PREDICTION AND STATISTICAL INFERENCE

Bruce M. Hill*

September 7, 1989

Abstract

The problem of Bayesian nonparametric prediction and statistical inference is formulated and discussed. A solution is proposed based upon A_n and H_n as in Hill (1968). The meaning of parameters in the subjective Bayesian theory of Bruno de Finetti is discussed in connection both with A_n and with conventional parametric models. It is argued that the usual sharp distinction between prediction and parametric inference is largely illusory. The finite version of de Finetti's theorem is emphasized for the practice of statistics, with the infinite case used only to obtain approximations and insight.

1 Introduction

Bayesian nonparametric statistics consists of methods for statistical inference and prediction based upon weak apriori knowledge as to the form of the underlying population. In real world problems one typically does not have the type of sharp apriori knowledge usually assumed about models. Indeed, it is well known that in the practice of statistics the most difficult and important phase consists of the specification of such models. In this article we wish to discuss the case in which it is difficult, or impossible, to model the data in terms of conventional parametric models such as the Gaussian, exponential, or even exponential family, at least without resorting to complex mixtures of such distributions. Our inference will instead be based upon A_n, the nonparametric Bayesian approach of Hill (1968, 1980a, 1988b, 1987b). A version of this approach was originally suggested from a fiducial point of view by R. A. Fisher (1939, 1948). See also Dempster (1963).

In his celebrated article La Prévision (1937), Bruno de Finetti proposed a subjective Bayesian solution to the problem of scientific induction, as formulated, for example, by the Scottish philosopher, David Hume (1748). De Finetti

*This work was supported by the U. S Air Force under grant AFOSR-87-0192, and by the National Science Foundation under grant DMS-8901234.

did so in terms of the concept of exchangeability, which is a special form of dependence that he introduced and studied extensively (1937). Other key references are Hewitt and Savage (1955), Savage (1972), Heath and Sudderth (1976), and Diaconis and Freedman (1980, 1981). In this article I shall first give a somewhat personal review of the history and substance of the connection between induction and subjectivistic perceptions of symmetry, with particular attention to A_n and H_n, which I developed for the case of vague or diffuse prior knowledge as to the shape of the underlying distribution of the observables.

The problem of induction is the problem of drawing inference about the future based upon the past. This problem has long plagued philosophers and others, partly because there is no way to prove that induction works (apart from induction itself), and also because in the real world it can be extremely difficult to formulate inferential or decision procedures, i.e., inductive techniques, that are appropriate in a given situation. The problem is best thought of in terms of the probabilistic prediction of potentially observable random quantities (not necessarily exchangeable), say $X_1, \ldots, X_n$. Given the values of the first n observations, $X_1 = x_1, \ldots, X_n = x_n$, what can we say about X_{n+1} or any other future observations? In the Bayesian approach this is done in terms of the evaluation of a probability distribution for the future observables, given the data $X_1 = x_1, \ldots, X_n = x_n$. Conventional Bayesian methods, using a prior distribution for a 'parameter,' such as the parameter of a Bernoulli sequence, yield such a predictive posterior distribution, in addition to the more customary posterior distribution for the parameter. In such situations, once a statistical model and prior distribution have been formulated and specified, the posterior distribution of the future observations, given the past, is thereby completely determined. Such a scheme may be called inductive, since it prescribes a (coherent) mode of inference and behavior with respect to the future observables, given any set of data.

This scheme, as usually interpreted, requires that there exist 'true' known probabilities that represent the conditional distribution of the data, given the parameter, i. e., a conventional statistical model. However, at the deepest level, where 'true' probabilities either do not exist, or even if in some as yet unknown sense they do exist, they are at least unknown, the conventional model-based Bayesian theory is incomplete, since it is difficult even to give operational meaning to the assertion that a particular model is 'true,' much less to find such a model. The problem that de Finetti clearly formulated and largely solved was the problem of giving meaning to Bayesian inferential procedures without relying upon the usual crutch of an assumed statistical model. Before the fundamental work of de Finetti, the assumption of such a true model was simply an unjustified act of faith. One could, of course, refer to some underlying physical theory, or to the central limit theorem, or some previous analogous data, to support belief in such a model. But deep down this remained at best a matter of delicate subjective judgment, and it was not even clear how to express what such subjective judgments concerned. For example, consider the use of

the normal or Gaussian distribution. Poincaré (1912, p. 171) states in connection with this distribution, "Tout le monde y croit cependant, me disait un jour M. Lippmann, car les expérimentateurs s'imaginent que c'est un théoréme de mathématiques, et les mathématiciens que c'est un fait expérimental," or "everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, and the mathematicians because they think it is an experimental fact." In the real world it is justified, in fact, by neither. In Hill (1969) it is shown that the use of the normal distribution can instead be based simply upon a subjective judgment of spherical symmetry for the 'actual' errors in the observations. See also Borel (1914, p. 66, 90-93) and Borel (1906). Hill (1969, p. 95) gives the exact density for the marginal distribution of n coordinates based upon spherical symmetry, or conditional uniformity, on the N-dimensional sphere. By Scheffé's lemma that convergence of densities to a proper density implies convergence in distribution, it immediately follows that each fixed r-dimensional marginal distribution of the joint distribution of the n coordinates converges to the Gaussian, even as n goes to infinity as well as N. In this sense spherical symmetry implies approximate normality. It should be noted that my statement of the result, which is for the case of spherical symmetry without a constraint on the average of all N coordinates, agrees with that of Borel, who discovered and stated the result for the case of one coordinate, and appears to have understood the general case. When there is also a constraint on the average of the N coordinates, then my exponent N - n - 2 should be changed to N - n - 3.

In the theory that I proposed spherical symmetry, or more generally, conditional uniformity on surfaces, is itself only an approximation based upon the available knowledge, and does not purport to be more than this, or to have any other objective meaning. For example, in the case of errors of measurement, one may view the usual orthogonal axes of a coordinate system as arbitrary, and therefore introduce rotational symmetry. Ultimately, it is simply a matter of judging that spherical symmetry represents a sufficiently good approximation to one's opinions in order to be useful for inference, prediction, and decision-making. In my opinion there is no hope to demonstrate that such a judgment is either 'correct' or 'incorrect,' other than empirically, for example, by seeing how well it works predictively.

At an even more basic level, as de Finetti first realized, induction can be based upon a direct subjective judgment of exchangeability for the sequence of observables. Once this subjective judgment is made, it is a mathematical fact that one will be acting (nearly) as though some statistical model were true. Conditional upon the parameters of such a model, the data will be regarded as independent and identically distributed, if the exchangeable sequence is infinite, and approximately so if it is a sufficiently long finite sequence. See Diaconis and Freedman (1980). (It should be noted that the conventional assumption of independent, identically distributed observations, with an unknown distribution, corresponds to the subjective Bayesian assumption of exchangeability.) This is

de Finetti's theorem, and its significance is that it provides a subjective justification for the use of statistical models and for conventional model-based Bayesian inference, provided that care is taken in the interpretation of such techniques. See Savage (1972) for a treatment of exchangeability from this point of view. De Finetti stressed the fact that the judgment of exchangeability is itself only a subjective judgment, and perhaps only an approximation to one's actual opinions. To the extent that one judges the sequence as exchangeable, then one is led to conventional Bayesian inferential techniques. Often, in fact, approximate. or even partial exchangeability, will suffice to justify such techniques. Furthermore, as will be discussed in Section 3, I believe that it is necessary to integrate the conventional Bayesian theory with data-analytic methods for the selection of models, parameters, and hypotheses. See Hill (1987a) for discussion of the deeper structures that may underlie conventional statistical models, Hill (1985) for Bayesian selection of models, and Hill (1988) for a theory of Bayesian data analysis.

In a practical sense, based upon the subjective judgment of exchangeability, de Finetti had completely solved the problem of inductive inference for the case of Bernoulli data, or, more generally, for multinomial data with a known finite number of categories. Combined with the beautiful result of W. E. Johnson (1932), as discussed in Zabell (1982), there was little more to be said at the foundational or even practical level for these cases, other than to elaborate on the choice of prior distribution for the Bernoulli parameter p or for the parameter θ of a multinomial distribution. Thus the precise measurement (or stable estimation) argument of L. J. Savage (1961, p. Ch. 4; 1962, p.20), or as presented in Degroot (1970, p. 199), deals with the case in which the prior distribution is diffuse relative to the likelihood function, so that it is of little consequence, and the posterior density for the parameter can be approximated by the likelihood function. On the other hand, H. Jeffreys's theory of hypothesis testing covers the most important situations in which the prior is not diffuse. See Edwards, Lindman and Savage (1963), and Hill (1974a, 1982) for discussions. The problem of so-called 'uninformative' priors has also been dealt with very effectively by a number of people for the case of multinomial data. See Good (1965, Ch. 4) for a review and discussion. Furthermore, it has long been recognized that many real world problems can be adequately modelled by such finite partitions, Fisher (1959, p. 111), Savage (1961, p. 4.23), so when this can be done effectively there is available a more or less complete system (apart from details and various complications that arise in practice) for inductive inference and decision-making, including the prediction of future observations. In de Finetti's own words (1937, p. 147): "It is thus that when the subjectivistic point of view is adopted , the problem of induction receives an answer which is naturally subjective but in itself perfectly logical, while on the other hand, when one pretends to *eliminate* the subjective factors one succeeds only in *hiding* them (that is, at least, in my opinion), more or less skillfully, but never in avoiding a gap in logic. It is true that in many cases–as for example on the hypothesis

of exchangeability–these subjective factors never have too pronounced an influence, provided that the experience be rich enough; this circumstance is very important, for it explains how in certain conditions more or less close agreement between the predictions of different individuals is produced, but it also shows that discordant opinions are always legitimate. This does not make any change in the purely subjective character of the whole theory of probability." See also de Finetti (1974, Ch. 11).

Thus in the exchangeable case the only type of situation that had not been essentially resolved was that in which no finite partition model was appropriate, or more generally, when the number of parameters requisite realistically to model the data is large relative to the number of observations. One can speak of this either in terms of multinomial data with an infinite number of categories, or alternatively, in terms of an unknown and possibly quite large or even infinite number of categories. Still again, to suggest the general type of problem, one can speak of Bayesian nonparametric statistics, or of Bayesian inference about an 'unknown' distribution function. Whatever words we may use, what we are trying to describe is the situation in which no conventional parametric statistical model is thought to be appropriate for the exchangeable sequence of observations. This situation seems often to arise in the p.actice of statistics. Indeed, from my own point of view, which will be explained further below, they in fact represent the great majority of statistical situations, with Gaussian and other conventional parametric models being appropriate only in very limited contexts.

What then can be said about the nonparametric case from a subjective Bayesian point of view? The first thing to observe is that de Finetti's theorem still holds, so that in the case of an infinite exchangeable sequence of observables, one will be mixing over a dummy variable that represents the 'unknown' distribution, say F, in the population. De Finetti (1937, Ch. 4), had already given an insightful development of the mathematics of this situation for exchangeable random quantities. Diaconis and Freedman (1980, 1981) have presented easily accessible proofs for even more general cases. Just as in the case of exchangeable events, what must be specified in order to implement the Bayesian approach, is the mixing function, or apriori distribution for F. In the case of exchangeable events, F is concentrated at only two known values, 0 and 1, while in the present case the distribution F can in principle be any distribution function on the real line. Special mixing distributions will correspond to the subset of F appropriate for a finite multinomial situation, in which the categories are coded numerically, or for sure knowledge that F is Gaussian, etc. The fully nonparametric case is that in which F cannot be restricted to such special subsets of the space of all distribution functions. In principle, what the Bayesian must do is to specify a prior distribution, π, on the space of all distribution functions, F; and then, given the data, update this prior distribution to become a posterior distribution, π^*, in accord with Bayes's theorem.

How then can a subjective Bayesian specify a prior distribution that ex-

presses a realistic degree of vagueness about F? Since we are dealing with possibly infinitely many parameters, it is clear that the problem is formidable even from the point of view of the mathematics involved, and of course even much more so conceptually. Furthermore, after observing a sample from the population, and obtaining π^*, in order to obtain the posterior predictive distribution of future observations, one would have to integrate the conditional distribution of the future observations, given F, with respect to this posterior distribution of F. Here the 'unknown' F plays the same role as the 'unknown' θ of a conventional multinomial model, but the mathematics is again enormously more complicated. In addition, a basic difficulty arises here that did not appear in the case of finite multinomial models. It is no longer the case that one can rely on some form of Savage's precise measurement argument. Thus the distribution function F may have infinitely many parameters, and no matter how large a finite sample is taken from the population, the prior distribution for F may still play a crucial role. Even if, more realistically, we regard F as having a large finite number of parameters, in a practical sense the same phenomenon occurs, since realistic sample sizes will be small relative to the total number of parameters. See Hill (1975b) for a discussion of this phenomenon. Typically there is no such thing as global robustness (for all possible distributions of F), or in other words, the posterior distribution may be extremely sensitive to the prior distribution for F.

The problem is not, however, so hopeless of solution as may first appear. The first modern day hint or suggestion as to the nature of a possible solution occurs in the work of R. A. Fisher (1939, 1948), who proposed a fiducial interpretation for what I later called A_n. and who gives credit to 'Student' for the underlying idea.

Consider a conventional formulation of statistical inference, in which the observations are conditionally independent with cumulative distribution function $F(x;\phi)$, where ϕ is a conventional unknown parameter. Assume that the distribution function is continuous in x for each ϕ. Let $X_{(i)}$ denote the ascending order statistics of the data, for i = $1,\ldots,n$. Then let $\theta_i = F(X_{(i)}\ ;\ \phi)$ - $F(X_{(i-1)};\phi)$, for i = $1,\ldots,n+1$, where by definition $X_{(0)}$ = - ∞ and $X_{(n+1)}$ = ∞. *Before* the data are drawn, clearly the distribution of the θ_i is a uniform distribution on the n-dimensional simplex, i . e., a special Dirichlet distribution in which all the parameters are equal to unity. This is the fundamental frequentistic intuition with regard to A_n, and which Fisher presumably used to put forth his proposed fiducial solution. Thus Fisher suggested (or implied) that even when the random variables $X_{(i)}$ are replaced by their observed values $x_{(i)}$, that the uniform distribution for the θ_i would still be appropriate. It should be mentioned that the articles by Fisher (1939, 1948) only briefly and cryptically discuss the nonparametric fiducial case that we are concerned with. The first clear formal statement of something like A_n is by Dempster (1963), who stated the Fisherian argument precisely, changed the name from fiducial to 'direct' probability, and applied the argument for prediction of future observables as

well. Dempster also asserted that what I later called A_n "does not appear to have a Bayesian interpretation."

In my 1968 article I showed that in fact A_n does have a Bayesian interpretation. Before discussing this, however, let me note that Fisher's proposed fiducial distribution is an example of a posterior predictive distribution, since whatever the rationale, it is posterior to the data, and does partially specify a probability distribution for the future data. This predictive distribution is not completely specified, since what it does is to attach a probability of $\frac{1}{n+1}$ to each of the n + 1 open intervals formed by the consecutive order statistics of the given sample, assuming that there are no ties, and goes no further. The fiducial argument that Fisher gave for this evaluation depends upon one's willingness to persist with the pre-data evaluation of the distribution of, say, $F(X_{(i)};\phi) - F(X_{(i-1)};\phi)$, *after* the $X_{(i)}$ are replaced by their observed numerical values. Such a fiducial argument, although intriguing, was logically suspect. See Edwards (1972, p. 207) for a totally devastating example against the logic of the fiducial argument. Also, Lindley (1958) had already shown, ***under certain special conditions,*** that the fiducial argument leads to a genuine posterior distribution only in simple cases reducible to that of a location parameter. However, Fisher, in a footnote (1959, p.51), asserted that "Probability statements derived by arguments of the fiducial type have often been called statements of 'fiducial probability'. This usage is a convenient one so long as it is recognized that the concept of probability involved is entirely identical with the classical probability of the early writers, such as Bayes. It is only the mode of derivation which was unknown to them."

In short, the situation with regard to A_n was anything but clear, and at the time it was not even known whether A_n was a coherent evaluation in the sense of de Finetti. If it was, then presumably it could have been derived by means of a prior distribution for F, Bayes theorem, and an integration with respect to the posterior distribution of F, as discussed above. The problem that I addressed in my 1968 article was that of giving such a derivation of A_n. The first step in my formulation was to consider the case of arbitrary finite populations, in which the size of the population may be unknown, rather than infinite populations, or in other words, to deal with finite exchangeable sequences.

This step is not only more realistic, since in the real world we are not ordinarily called upon to deal with more than finite populations or sequences of observables, but it also greatly simplifies the mathematics. Indeed, the proofs of de Finetti's theorem for the infinite case by Heath and Sudderth (1976) and by Diaconis and Freedman (1980, 1981), proceed by taking limits for the finite case. Thus for the case of events, one can condition on the sum of the indicators after N observations, and note that because of exchangeability, all paths to a given sum are equally likely. The exchangeable distribution of any N indicators (for example, for red and white balls in an urn) is then identical with one that arises from sampling without replacement from a 'randomly selected' urn with N balls, i. e., the draws are made from an urn with composition (R,W), R + W = N, with a probability equal to the original subjective probability that the

sum of the N indicators is R. More generally, one can consider the empirical distribution function that arises from 'sampling' the first N coordinates from an exchangeable sequence. Conditional on this empirical distribution function, the individual coordinates are distributed uniformly over the collection of N-tuples having this empirical distribution. Because sampling without replacement is, for large N, close to sampling with replacement from this empirical distribution, and because of the convergence of this empirical distribution to some limiting distribution (with probability one, under exchangeability), one obtains de Finetti's theorem for the infinite sequence in this way. See Diaconis and Freedman (1980, p. 749; 1981, p. 209) for details. These authors have emphasized the importance of the finite case even for the underlying mathematics, and as I shall argue later, it is also the appropriate formulation for inferential purposes. The model for A_n in Hill (1968, p. 679) is actually equivalent to the specification of a diffuse prior distribution for the empirical distribution of a finite population. In the context of Heath and Sudderth, or of Diaconis and Freedman, it is equivalent to specifying the subjective distribution for the sufficient statistic based upon N trials within their models. Thus instead of specifying a diffuse prior distribution on the space of all distribution functions F, what I have done is to specify such a prior distribution for the empirical distribution function of the entire finite population from which a simple random sample has been drawn. Here the number of units in the finite populations can be unknown, as well as the number of jump-points of the empirical distribution of the population, and the points at which the jumps occur and the sizes of the jumps. The details concerning this diffuse prior distribution will be given in Section 2. Here what I want to discuss is the underlying sampling model.

The statistical model for this problem can be thought of in terms of sampling with or without replacement from a finite population of units. Imagine that each unit carries an attached tag or label, for example giving the color of the unit, or the name of the species to which the unit belongs, or a numerical value such as the mass of the unit, or the future time of death of the unit. It does no harm to visualize the population of units, with their attached labels, as sitting in an urn. A simple random sample, with or without replacement, is then drawn, and we observe the value of the label for each unit in the sample. It is assumed for simplicity here that the label or numerical value is observed without error, although the theory can easily be extended to deal with errors of measurement.

The population of labels or numerical values can be described in terms of the empirical distribution of such labels or values. Indeed, because we are dealing with only finite populations, the case of colors or names can be viewed as a special case of the case of numerical values. Thus we can imagine, without loss of generality, that the finite collection of colors or names in the entire population have been encoded numerically, thus yielding numerical values. For simplicity, we shall describe the situation for the case of such numerical values. When we return to the case of 'colors,' as in the species sampling problem, we shall point out the special features that arise in this case. For the time being, visualize the

urn population as consisting of the numerical values attached to the units, such as their masses. This population can then be described in terms of the number of units in the population, say N, and the empirical distribution of the values in the population. Note, for example, that if sampling is with replacement from this population, and if the number of distinct values in the population is known to be say, M, then this model is a special case of a conventional multinomial model with exactly M non-empty categories. In general, of course, M need not be known, except that $M \leq N$, and sampling can be without replacement. In any case, the number of units in the population, N, and the empirical distribution of population values, completely characterize the finite population of values. In fact here the empirical distribution for the entire finite population of values plays a similar role to that of the 'unknown' probabilities in a conventional statistical model. Following the spirit of de Finetti (1937), I regard the fundamental problem of induction to be reducible to that which arises in sampling without replacement from an urn consisting of units that are labelled with numerical values.

The solution that I proposed for this problem, which consists of a model for a generalized version of A_n in which ties can occur, will be discussed in Section 2. Historically, the sequence of events concerning A_n after Dempster (1963) was as follows. I proved in my 1968 article that A_n cannot hold exactly for *countably additive proper prior distributions*, in the case of exchangeable sequences in which ties have probability 0. At the same time I recommended it as an approximation for a variety of situations, that can be roughly described as situations in which the data is measured on a "rubbery scale," and gave several models in which it would be appropriate. I also proved in Hill (1968, p. 686) that A_n for all n implies that the posterior distribution of the θ_i defined earlier is the uniform Dirichlet distribution on the (n − 1) dimensional simplex, thus giving support to Fisher's fiducial argument. Also, Hill (1967) derived the posterior expectation of a future observation, and of the mean of the population, using A_n. The next historically significant development regarding A_n was the proof by Lane and Sudderth (1978), using finite additivity, that A_n for all n is coherent in the sense of de Finetti, i. e., it is impossible to be made a sure loser, and the further result by the same authors (1984) that it is predictively coherent. The robustness and invariance properties of A_n were investigated by myself in Hill (1980a) with the general result that it is robust in the modern Bayesian sense of Berger (1985), Hill (1980b). Then my doctoral student Peter Lenk (1984) showed, along with many other things, that A_n can arise as a limit of proper priors, using a log gaussian model for the prior distribution of the unknown density function of the population. Next, Berliner and Hill (1988) used A_n to obtain the predictive distribution for future observations in the case of censored data, as for example in survival analysis. Finally, in Hill (1987b) I have constructed a class of simple parametric models, called splitting processes, such that A_n holds for all n. A modification of this construction also yields H_n for all n. The Dirichlet process of Ferguson (1973) turns out to be a very

special type of splitting process. It is also shown how A_n arises from sampling of complex mixtures of distributions, and the relationship with the oneway random effects analysis of variance is explained.

In the next Section I will restate A_n, give my model for inference and prediction, and suggest a new and compelling (for me) subjectivistic argument for A_n.

2 A_n and H_n

In Hill (1968) a direct specification, denoted A_n, for the posterior predictive distribution of future observations was proposed. A_n was meant to express extremely vague subjective prior knowledge as to the form of the underlying population distribution. For the case of n = 1 and 2, A_n follows from conventional parametric models (Gaussian, for example) with a uniform prior distribution on the location parameter, or on the location parameter and logarithm of the scale parameter, respectively, Jeffreys (1961, p. 171), Hill (1968, p. 688). For example, when n = 1, suppose that the parameter θ is the mean of a normal population with known standard deviation of unity. Given an observation $X_1 = x_1$ from this population, the posterior distribution of θ is N(x_1, 1). The predictive distribution of the next observation, X_2, is then easily seen to be N(x_1, 2). Hence the posterior probability that $X_2 \leq x_1$ is .5. Note that for any prior distribution which is diffuse relative to the likelihood function, A_1 will hold to a good approximation, since the posterior distribution of θ will still be approximately N(x_1, 1). A similar analysis applies in the case n = 2. At the time of Hill (1968), it was not known whether A_n could be obtained for conventional parametric models when $n \geq 3$. However, Hill (1987b) shows that this is the case for both A_n and H_n.

A_n for untied data, or H_n for the case of ties, are exactly appropriate for data measured on a merely ordinal scale, or with a trivial modification, for data that consists of labels (such as the names of species, as in the species sampling problem), and can yield an extremely good approximation for data on a ratio or interval scale, such as the weights in a population of penguins, as will be discussed at the end of this section. The cases where it is exactly appropriate can be described as data measured on a "rubbery" scale. Just as with other nonparametric models, it is hardly necessary for the assumptions to hold literally, in order that the conclusions be appropriate to a very good approximation.

The condition A_n is defined as follows. A_n asserts that conditional upon $X_1, \ldots, X_n$, the next observation X_{n+1} is equally likely to fall in any of the open intervals between successive order statistics of the given sample (Hill, 1968, p. 677). Note that in our *definition* of A_n we do not assume that the sequence is necessarily exchangeable or that ties have probability 0. Thus, we can also include cases where there is a positive probability that the next observation ties

one of the previous observations, and also partially exchangeable situations that satisfy A_n. At the present time I wish to slightly modify this notation, use H_n to denote the situation in which ties can occur, and reserve A_n for the special case of H_n in which their are no ties (or ties have probability 0). In this article I will also assume that the observations are exchangeable, although this will not be included in the definition of A_n and H_n.

A_n specifies a predictive distribution for one future observation. If also A_{n+1} holds, then by conditioning upon which interval the first new observation falls in, we can obtain a predictive distribution for two new observations, and by extension for an arbitrary number of new observations. See Hill (1968, p.684) for such predictive schemes. [1] Furthermore, we can use this same idea to deal with censored data, again by conditioning upon which intervals the censored observations will fall in. Berliner and Hill (1988) carry through such an analysis for the case of survival data, present upper and lower bounds for the survival function, and simple algorithms with which to make the analysis. In the survival problem, for example, we assess the predictive probability distribution for the time of death of new patients given a treatment, using as data the death times, and the intervals in which censoring occurred, i. e., the partial censoring information, for a previous group of patients who were given the treatment, and with whom the new patients are regarded as exchangeable. Chang (1989) provides additional computational algorithms, and extends the results to the two sample case.

In addition to de Finetti (1937, 1974), other key references on exchangeability are Hewitt and Savage (1955), Savage (1972), Heath and Sudderth (1976), and Diaconis and Freedman (1980, 1981). (The article by Heath and Sudderth gives an extremely simple and yet rigorous proof of de Finetti's theorem for the case of events. The articles by Diaconis and Freedman do so for the general case.) The definition of exchangeability that we shall use is that motivated by the subjective Bayesian viewpoint, namely, in terms of a subjective judgment that the order is irrelevant. (Mathematically, this is the same as all other definitions of exchangeability but psychologically it is different, in that we do not assume the sequence is 'truly' exchangeable, but merely that one regards it as exchangeable, perhaps only as an approximation to the truth.)

To be precise, let $X_1, \ldots, X_{k-1}$, be k − 1 random variables that are (finitely) exchangeable in the subjective Bayesian sense; that is, the joint distribution of any r distinct variables is the same as that for any other such r variables, r = $1, \ldots, k$. An infinite sequence of such variables is said to be exchangeable if the above condition is true for each k. Such models arise from the following Bayesian formulation: Assume that, given some distribution, say F, $X_1, \ldots, X_n$, are independent and identically distributed according to F. For F unknown it is natural for the Bayesian to model F itself as 'random' with some apriori prob-

[1] Note that the equation on the top of page 684 is only valid if $i \neq j$. When $i = j$, it is necessary to add another term which corresponds to the possibility that the second new observation ties the first. A similar correction is necessary in the formula for $E(\theta_i \times \theta_j)$.

ability specification. This can be done either parametrically or nonparametrically. In either case, 'integrating out' F leads to an exchangeable unconditional joint distribution for the X's. Conversely, de Finetti's theorem implies that if the exchangeable sequence is infinite, then their exists a distribution on F, called the prior distribution of F, for which the joint distribution of the observations obtained by 'integrating out' F is the original exchangeable distribution. See Hewitt and Savage (1955). Heath and Sudderth (1976), and Diaconis and Freedman (1980, 1981) for proofs. The 1980 article, which emphasizes the finite exchangeable case, is particularly appropriate for my purposes. Thus the authors show that the most general exchangeable sequences arise by taking limits of the finite exchangeable sequences that arise in sampling without replacement from urns.

I will now present my model for A_n, or more precisely, for my generalization of A_n, called H_n, which allows for ties, and of which A_n is a special case.

We assume that there exists a finite population of units, with each unit having an attached value or label. For example the value might be the mass of the unit, or the label might be the name of the species to which the unit belongs. We assume that the set of values is simply ordered, or at least can be simply ordered. By a simple ordering we mean a relationship, say $\leq$, which for any two elements x,y, of the set of values, is such that either $x \leq y$ or $y \leq x$, and which is transitive. (See Jeffreys (1957, Chs. 5-6), Luce and Narens(1987), and Whitrow (1980, Sec. 4.7) for discussions of the concept of measurement.)

Thus masses would be on a ratio scale, and are certainly simply ordered; while labels can be simply ordered for a finite population simply by designating an ordering. (This can be done for infinite populations as well, using the well-ordering theorem, but there is no need to go into such things here.) Suppose there are N units in the population, and that the set of attached values or labels is $\{Z_i, i = 1, \ldots, N\}$. We shall now refer only to values, with it being understood that we include labels as a special case after the finite population has been simply ordered. Some of the values Z_i may be equal to one another. Suppose that in fact there are only M distinct values amongst the Z_i, and denote these in ascending order of magnitude as $X_{(1)} < X_{(2)} < \ldots < X_{(M)}$, where of course $M \leq N$. Finally, suppose that the value $X_{(i)}$ occurs in L_i units, where $L_i \geq 1$, since by assumption the value $X_{(i)}$ does in fact occur, and $\sum_{i=1}^{M} L_i =$ N.

The above model constitutes our description of the finite population of values Z_i. Note that this determines the empirical distribution of values in the finite population, i. e., the empirical distribution has jumps occurring at $X_{(1)}$, ..., $X_{(M)}$, and the jump that occurs at $X_{(i)}$ has height L_i/N . Of course in general all of these quantities are unknown, i. e., N, M, the $X_{(i)}$, and the L_i. From the subjective Bayesian point of view one must then specify a probability distribution for all of these quantities. It should be noted that this point of view corresponds exactly to the recent probabilistic treatments of exchangeability for

finite sequences, as in Diaconis and Freedman (1980), where the finite exchangeable sequence of length N is the vector $Y_1, \ldots, Y_N$, which would be generated by sampling without replacement all N elements of the finite population, so that these Y_i are some permutation of the Z_j.

The case of an infinite exchangeable sequence may be viewed as an idealization of this scheme, and gives rise to de Finetti's theorem. But the model in terms of sampling from a finite population is simpler, avoids difficulties and paradoxes of infinity, is more realistic, and in view of the results of Diaconis and Freedman, loses no generality in any case. For example, in my model we require only a prior distribution for the composition of the finite population, i. e., for (M, N, $\underline{X}$, $\underline{L}$), rather than a prior distribution on F, the theoretical distribution for an infinite exchangeable population. It is far simpler to specify such a prior distribution on the finite number of parameters (at most 2N +2) needed to describe this finite population, than to do so on the infinite dimensional space of distribution functions F. Furthermore, we shall argue that there is a natural way to represent vagueness for the finite population, which would be much more difficult to achieve for an infinite population (for example, one would have to confront some basic issues concerning the difference between countable and finite additivity).

Now let us consider the data that we shall be analyzing. It is assumed that a simple random sample is drawn without replacement from the finite population that we have described above. Let the sample size be n. The data will consist of the numerical values attached to the n units that are thus selected from the finite population. Let $x_{(1)} < x_{(2)} < \ldots < x_{(m)}$, be the ascending order statistics of the sample, with m distinct values, $1 \leq m \leq n$, and with n_i sample units having the value $x_{(i)}$. Thus $n_i \geq 1$, and $\sum_{i=1}^{m} n_i = n$. It is assumed here that the values are measured without error, so that each $x_{(i)}$ is necessarily some $X_{(j)}$ in the population, but of course we do not know with certainty which. By data we mean the set of m distinct $x_{(i)}$ values, and the n_i. Thus the data determines the empirical distribution of the sample, but is more informative because n and the n_i are known as well. We now require only one further bit of notation. Given the data, define J_i to be the rank, in the population, of the the value $x_{(i)}$ in the sample, for $i = 1, \ldots, m$. The vector $\underline{J} = (J_1, \ldots, J_m)$ then gives the true ranks, in the population, corresponding to the sample values $x_{(i)}$. Thus $1 \leq J_1 < J_2 < J_3 \ldots < J_m \leq M$. because of the fact that the $x_{(i)}$ and $X_{(j)}$ are strictly ordered. We now are ready for the basic equations of the Hill model for H_n.

In the first equation, we condition on the true composition of the finite population, by which we mean the unknown quantities $\underline{X}$, $\underline{L}$, M, and N. This equation gives the probability for observing the data together with $\underline{J} = \underline{j}$, for each possible vector of ranks $\underline{j}$:

$$Pr\{\ data,\ \underline{J} = \underline{j} \mid \underline{X},\ \underline{L},\ M,\ N\ \} = \binom{N}{n}^{-1} \times \prod_{i=1}^{m} \binom{l_{j\,i}}{n_i}, \qquad (1)$$

if $X_{(j\,i)} = x_{(i)}, i = 1, \ldots, m$, and is otherwise 0.

Note that this would be the likelihood function for the population quantities, except that we have included $\underline{J} = \underline{j}$ together with the data, because this is the key to making an effective evaluation; the ordinary likelihood function would involve a mixture of (1) with respect to $\underline{j}$. For sampling with replacement, it is only necessary to replace the factor $\binom{l_{j\,i}}{n_i}$ by $(l_{j\,i})^{n_i}$, etc. We shall not further deal with the case of sampling with replacement, since sampling without replacement is the more common, more difficult to analyze, and more important form of sampling.

The next step is to integrate out over the unknown $\underline{X}$ values in the population. In general such an integration requires the assumption of countable additivity, or conglomerability in the finitely additive theory, as in Hill and Lane (1985). However, in the present case with only two values for the probability in question, i. e., that given by (1), or else the value 0, it follows without any additional assumptions, that

$$Pr\{\ data,\ \underline{J} = \underline{j} \mid \underline{L} = \underline{l},\ M,\ N\ \}$$

$$= \binom{N}{n}^{-1} \times \prod_{i=1}^{m} \binom{l_{j\,i}}{n_i} \times Pr\{\ X_{(j\,1)} = x_{(1)}, \ldots X_{(j\,m)} = x_{(m)} \mid \underline{L} = \underline{l},\ M,\ N\ \}. \qquad (2)$$

We thus obtain the basic result,

$$Pr\{\ \underline{J} = \underline{j},\ \underline{L} = \underline{l},\ data \mid M, N\ \} = Pr\{\ data,\ \underline{J} = \underline{j} \mid \underline{L} = \underline{l},\ M,\ N\ \} \times Pr\{\ \underline{L} = \underline{l} \mid M, N\ \}$$

$$= \binom{N}{n}^{-1} \times \prod_{i=1}^{m} \binom{l_{j\,i}}{n_i} \times Pr\{\ X_{(j\,1)} = x_{(1)}, \ldots X_{(j\,m)} = x_{(m)} \mid \underline{L} = \underline{l},\ M,\ N\} \times Pr\{\ \underline{L} = \underline{l} \mid M, N\ \}. \qquad (3)$$

Clearly all that must be specified in order to make further evaluations are simply the three components of the prior distribution on the composition of the population, namely

$$Pr\{\ \underline{L} = \underline{l} \mid M,\ N\ \}. \qquad (4)$$

$$Pr\{\ X_{(j\,1)} = x_{(1)}, \ldots, X_{(j\,m)} = x_{(m)} \mid \underline{L} = \underline{l},\ M,\ N\ \}, \qquad (5)$$

and

$$Pr\{\ M \mid N\ \} \times Pr\{N\}. \qquad (6)$$

Although our primary interest in this article is the specification of (5) in such a way as to express diffuse or vague knowledge about the underlying population of values, we note that our formulation is sufficiently general so as to include conventional parametric specifications as well. For example, we may be of the opinion that the population distribution is approximately normal, in which case the distribution of $\underline{X}$ can be chosen so that the $X_{(i)}$ are order statistics of a sample from a normal population, and similarly for any other parametric distribution. We shall not pursue this idea here, however, since the most basic case is the nonparametric one.

We shall specify (4) and (5) as follows:

$$Pr\{\ \underline{L} = \underline{l} \mid M, N\ \} = \binom{N-1}{M-1}^{-1}, \qquad (7)$$

while, for each possible $x_{(1)}, \ldots, x_{(m)}$,

$$Pr\{\ X_{(j_1)} = x_{(1)}, \ldots, X_{(j_m)} = x_{(m)} \mid \underline{L} = \underline{l},\ M,\ N\ \} \qquad (8)$$

does not depend upon $\underline{j}$.

Any specification of (4) and (5) is equivalent to a specification of the prior distribution for the empirical distribution of the population, given M and N. Obviously this can be done in infinitely many ways, any one of which might be appropriate in a specific real world situation. But it is of value to single out those specifications that are of special significance, such as for example correspond to a diffuse prior distribution (as is commonly done with improper prior distributions on conventional parameters), and also those that are known to be compatible with much real world data. The specification (7) that I originally chose was to take $Pr\{\ \underline{L} = \underline{l} \mid M,\ N\ \} = \binom{N-1}{M-1}^{-1}$, which is the Bose-Einstein distribution for non-empty cells, as in Feller (1968, p. 40). Thus the results in Hill (1968) are based upon this choice, while those in Hill (1980a) discuss the robustness of this choice within the class of exchangeable distributions for $\underline{L}$. My doctoral student, Wen-Chen Chen, in his Ph. D. dissertation (1978) and Chen (1980) generalized this choice to include arbitrary symmetrical Dirichlet-multinomial distributions, and argued that for some data it is desirable to choose a Dirichlet prior other than the Bose-Einstein, which of course is a Dirichlet-multinomial corresponding to a uniform Dirichlet distribution. See also Lewins and Joanes (1984) and Boender and Kan (1987) who use the same model. My primary motivation for the Bose-Einstein distribution (which I still regard as the single most appropriate choice) is the connection with Zipf's Law. This law represents more real world data than any other known law, including the Gaussian. It is shown in my articles Hill (1970, 1974a, 1975a, 1979, 1980a, 1981), and in Hill and Woodroofe (1975) and Woodroofe and Hill (1975), that the Bose-Einstein

choice yields Zipf's Law. This is why I singled it out as of special significance within the class of exchangeable prior distributions for $\underline{L}$. See also Ijiri and Simon (1975) for discussion of the Bose-Einstein distribution. Of course it is mathematically straightforward to replace the Bose-Einstein distribution by any other Dirichlet-multinomial distribution, and sometimes this may be of value in modelling the data. The logic underlying my model would only at best suggest that the distribution of $\underline{L}$ should be chosen to be exchangeable, and even this is not really necessary. See also Hill (1987b) for the relationship between my model for Zipf's Law and the random discrete distributions of Kingman (1975).

Next, one must also make some specification for the prior distribution of M and N. The most basic case for inference is simply where N is known to be large, and M has a uniform distribution, given N. This was the case considered in Hill (1968). Hill (1979) then considered the case where M has a truncated negative binomial distribution, of which the uniform is a special case. Although the specification of the distribution for M, given N, is of lesser importance here than the specification of (4) and (5), it does play a crucial role in obtaining Zipf's Law, as in the cited articles by myself and by Chen.

Even more important than the choice of the Bose-Einstein distribution for $\underline{L}$ is the choice of (8). Here we directly confront the problem of formulating a diffuse prior distribution on the empirical distribution of the population. Note that if M = N, so that all $L_i = 1$, then we obtain the case where ties have probability 0, and must then only express vagueness of opinion about the jump-points $X_{(i)}$ in the population. Thus the problem of expressing a diffuse distribution for the jump-points is logically independent of that of expressing one for $\underline{L}$. It was shown by Lane and Sudderth (1978, Theorem 1), defining A_n for the case where ties have probability 0 and where the sequence is exchangeable, that (8) is equivalent to A_n.

Consider then the specification (8). What it says is that no matter what the distinct values $x_{(i)}$ may be, they contain no information whatsoever about the ranks J_i of these values in the population. Clearly this is not always appropriate. For example, if one believed that the population was approximately Gaussian in form, then one would favor some $\underline{j}$ vectors over others. Or if one knew sufficiently much about the set of values in the population, then one might know, for example, that $x_{(m)}$ was in fact the largest value in the population. Or again, if the $X_{(i)}$ are necessarily integers, and if two are consecutive, then one knows that the corresponding J_i are also consecutive. To understand the force of the argument for (8) however, consider the following example.

Suppose for the sake of argument that there are 100,000 adult male emperor penguins, and that their weights can be measured sufficiently precisely so that no two agree exactly. (This is assumed only to make the essential point clear. My model H_n, with ties, can deal with any degree of rounding.) Consider your apriori subjective opinions about the population of weights of these penguins. Suppose now that I were to give you all but one of these weights as the data $x_{(i)}$, i. e., 99999 positive numbers, no two of which are equal. The question I wish

you to think about concerns your opinions about the vector $\underline{J}$, which specifies the ranks of these 99999 numbers in the population of all 100000 numbers. Condition (8) here would require that you be aposteriori indifferent as to which ranks these observations have in the population with M = N = 100,000. Note that this is meant to apply no matter what the $x_{(i)}$ values are, provided that only possible values are included, so that negative values are excluded, as well as weights that are known to be impossibly large or impossibly small. For example, if (8) holds, then you are indifferent as to which of the 100000 possible values is missing in the data. It could just as well be the largest as the smallest, or any other member of the population. Thus it would be the largest that is missing if it were the case that $\underline{J}$ consists of the ranks $1, \ldots, 99999$ in the population, and it would be the smallest that is missing if $\underline{J}$ consists of the ranks $2, \ldots, 100,000$ in the population. Are you so indifferent?

A fairly natural first reaction is to say that you might or might not be indifferent, depending upon what the numbers $x_{(i)}$ that I give you are. And you might feel that for lots of such sets of 99999 numbers you might be, and for others you might not be. But think again. Suppose, to take an extreme case that might seem to speak against $A_{(99999)}$, that the $x_{(i)}$ that I give you are such that there is an enormous gap between the largest, $x_{(99999)}$, and all the others. In fact, suppose that $x_{(99999)}$ is an extremely large value, say 1000 pounds, one that (although perhaps not impossible), seems highly improbable, while the other sample weights are all less than 100 pounds. You do not appreciate the full force of A_n until you realize that if the largest weight in the sample were in fact 1000 pounds, then there might well be another penguin that weighs even more than this! Thus the naive reaction, which would be that no penguin weighs anything like 1000 pounds, is immediately dispelled once one fully appreciates the fact that you have already seen one such (in the scenario of the problem) and may therefore well see another one. Still another example of this type concerns human age. One might well regard it as extraordinarily improbable that any human being has lived to the age 500 years. But if one such could be demonstrated, then you might well think that another might also, and even find that your opinions were roughly in accord with A_n.[2]

What condition (8) is expressing is a completely pragmatic attitude towards the population. Such an attitude is not only a subjectively Bayesian coherent attitude but in the case at hand even seems quite compelling; and this is for the case of weights on a ratio scale, which is the worst type of example for A_n, as opposed to data on a merely ordinal scale, such as the Mohs scale for hardness of rocks, where the hardness values are more or less meaningless. See, for example, Whitrow (1980, p. 216). And yet I think, after reflection, you may find it compelling even in the extreme example I have given. It would of course

[2]The Encyclopedia Americana, 1981, referring to penguins, states "In size they range from the gigantic emperor penguins, standing about 40 inches high, and weighing up to 90 pounds, to the diminutive fairy penguin of the Australian region that attains a length of just over a foot.

be even more compelling if the largest weight in the data were say, 120 pounds, rather than 1000 pounds. The general argument that I would give is that (8) with m = M - 1, and N = M (so there are no ties), is a highly compelling subjective evaluation, and this implies A_{M-1}. Note that there is no possibility of a mathematical proof that (8) is 'correct,' just as there is never any way of proving that one ordinary prior distribution is more appropriate than any other. All prior distributions are possible, and each is to be given 'equal rights,' as de Finetti says. But just as some prior distributions are sometimes regarded as more appropriate than others, for example, a uniform prior distribution on the parameter of a Bernoulli process is sometimes regarded as particularly appropriate, so too I claim that (8) is quite compelling, and I personally regard it is the most generally appropriate specification. My reasons are perhaps not entirely unrelated to those of Bayes (1764), and the fiducial intuitions by 'Student' and Fisher, .

That A_n for large n should be highly compelling also agrees with certain frequentistic ideas in conventional nonparametric statistics. Very few statisticians use parametric models when dealing with large samples from some underlying population F. The reason is that one is nearly certain that the true distribution is not of any specific parametric form, for example, Gaussian, and that with a sufficiently large sample the discrepancies will almost certainly appear and be serious. This is part of the approach to hypothesis testing of J. Berkson (1938), for example, who pointed out that with a sufficiently large sample you will certainly reject most conventional null hypotheses. Thus for a sufficiently large sample one might be nearly certain that the data will allow rejection of any prespecified fixed dimensional parametric model, even using a subjective Bayesian test of the hypothesis, for which it is more difficult to reject the null hypothesis. On the other hand, if the sample from the very same population F were sufficiently small, then one might well use the Gaussian or some other parametric model. Because of the relationship of A_n to the empirical distribution function, as in Berliner and Hill (1988, p. 773), it is clear that the same considerations that make conventional statisticians prefer the empirical distribution function when dealing with large samples should also apply to A_n.

Now we come to a rather strange and interesting fact. Suppose I have managed to convince you of the appropriateness of A_{M-1}. But it is a mathematical fact, proved in Hill (1968, p.688), that A_k implies A_j for $j \leq k$. Thus if you accept A_{M-1} as appropriate exactly, then you are forced into A_1 as well. Of course, both A_1 and A_2 correspond to conventional Bayesian and frequentistic procedures, with a diffuse prior on location, or on location and scale parameters, respectively, and they are certainly sometimes appropriate as an approximation. But it is equally clear that they are not always appropriate. How are we to explain this? My argument for A_{M-1}, which I regard as extremely compelling when M is large, if accepted, then implies A_1 as well, which is not always compelling. I believe that the explanation is as follows. In my proof that A_k implies A_j for $j \leq k$ there is a backwards induction. In carrying the

argument backwards, it is possible that slight discrepancies from A_{M-1} may build up, yielding a possibly much larger discrepancy for A_1. I should also point out that even A_{M-1} need not hold literally. For example, suppose that you knew a great deal about the average weight in the population of penguins. Then if I gave you all but one of these weights, you would have a good idea about the missing weight. Indeed, you would know it exactly if you knew the average weight exactly. Similarly, one might observe that a particular weight that one knows occurs in the population is missing in the sample. Thus one might have a discrepancy even from A_{M-1}, and this could build up even more in reaching down to A_1. It is considerations such as these which point out the importance of recognizing, once and for all, that we are at best only dealing with approximations. These approximations can nonetheless be very useful. It is my opinion that the nonparametric formulation, as in H_n, although itself only an approximation, is ordinarily the most important way to perform predictive inference, with parametric representations, such as the Gaussian, being useful primarily for inference and prediction when the sample size is small.

I remarked earlier that A_n is exactly appropriate for merely ordinal data, such as hardness of rocks, in the absence of ties. The argument is as follows. Suppose one draws a simple random sample of n rocks from a population of N rocks in which no two are of the same composition or of the same hardness. (Here, as is usual, one rock is said to be harder than another if it scratches the other rock.) Before the data is taken you are surely of the opinion that $\underline{J}$ is equally likely to be any of the $\binom{N}{n}$ possible $\underline{j}$ vectors, since this is precisely what sampling without replacement means. In the present case, however, even after the sample is drawn you must still be of the same opinion, since no 'data' becomes available other than the relative orderings of the rocks in hardness, i. e., there are no values. (Even if some arbitrary scale is used, such as the Mohs scale, it means nothing, and its 'values' are totally uninformative, as in Whitrow (1980, p. 216).) Thus in this situation one is forced to make the evaluation (8). Furthermore, this provides a justification for the original fiducial intuition, which also ignores the 'values' of the observations. Finally, if we now consider the case of ties, as for example if we draw a simple random sample from the rocks on some mountain, then it can easily be seen that H_n rather than A_n applies, using an appropriate exchangeable distribution for $\underline{L}$.

Finally, in the case of 'colors' or 'species,' the natural way to proceed is as follows. Suppose that we go to a new region and, taking samples, find n living creatures that we decide belong to m distinct species. We can number these species in any way we like, for example, we can take species 1 to be the first type caught, etc., with species m the last type caught in our sample. Define the quantity γ_i, i = $1, \ldots, m$, to be the proportion of the unsampled population belonging to the same species as the i^{th} sample species, and θ_i to be the proportion of the unsampled population with value strictly between $x_{(i)}$ and $x_{(i-1)}$, just as in Hill (1968, p. 682). In this case the θ_i are necessarily 0 for $i \leq m-1$, since any creature belonging to a new species must then be given a

number larger than m. Then although neither A_n nor H_n is exactly appropriate, it is shown in Hill (1980a) that the posterior distribution of the quantities γ_i, $i = 1, \ldots, m$, is exactly as under H_n, and that the posterior distribution of M and the posterior probability of catching a new species is as in Hill (1968, p. 681, p. 691; 1979)

3 On the meaning of parameters

The role and meaning of parameters in the de Finetti theory is quite different from that in conventional statistics. Consider a finite exchangeable sequence of 0-1 valued observations, X_i, for $1 = 1, \ldots, N$. In the de Finetti approach, there need not be any pre-existing 'true' probability, p, for a success, i. e., for $X_i = 1$. However, according to de Finetti's theorem, if the sequence were infinite, then one would implicitly be acting as though there were such a p, and the prior distribution for such a p would be simply the de Finetti measure π for the sequence, i. e., it is as though the a priori distribution for p was π, and one's opinions about the observable X_i were such that conditional on p, the observations formed a Bernoulli sequence. If the sequence is only finite, but N is sufficiently large, to a good approximation the same thing is true. In this case, p is simply the average of the N random quantities, $p = \bar{X} = (\sum_{i=1}^{N} X_i)/N$. Conditional upon this p, one no longer has exact independence of the X_i, but some degree of dependence. The difference between the infinite case and the finite case amounts to the difference between sampling with replacement versus without replacement from an urn. See Heath and Sudderth (1976) and Diaconis and Freedman (1980). Of course all real world sequences are necessarily finite, but for moderate N the difference between the infinite and finite case is of little importance, and one uses the infinite case as a convenient approximation to the finite case. This is also the spirit in which I originally proposed A_n.

In this formulation note that before the sequence is actually determined, for example, before the coin is flipped, there is no pre-existing p, and what p actually represents is the random average $\bar{X}$ of the N observables, which is as yet to be determined. The a priori distribution for p is merely the prior distribution for $\bar{X}$, and this is in fact a useful way to elicit opinions about the conventional Bernoulli parameter p. Although p is usually thought of as a quantity with an objective existence even before the coin is tossed, this is not really the case. Of course, one can imagine if one likes, that the coin has already been tossed N times, so that the X_i have already been determined, but that one has not yet observed them. In this case there would be an existing quantity, which is as yet unknown, and is simply $\bar{X}$ for the realized sequence X_i. Provided there is no additional information, in the subjective Bayesian framework it is then precisely as though the tosses had not yet been made. See Hill (1988, Sec. 3) for futher discussion. It is then largely immaterial, for practical purposes, which point of view one takes as to the 'objective' existence of p. (Note, however, that

even in the case where $\bar{X}$ has already been realized, the interpretation of this quantity as the 'true' probability cannot be made without assumptions as to the sampling mechanism.) In this framework the distinction between 'inference' and 'prediction' becomes blurred. On the one hand, if the sequence has not yet been determined, one would view p as a random quantity which one might want to predict, i. e., it is the future proportion of heads. On the other hand, if the sequence has been determined, but is as yet unobserved, then p might be thought of as a parameter in the conventional sense. The upshot of this discussion is that the usual sharp distinction between prediction and parametric inference is largely illusory.

The situation with regard to parameters in A_n is more subtle. Given the data, X_i, for $i = 1, \ldots, n$, I have defined the 'parameters' θ_i and γ_i to be the proportions of observations in the unsampled population, between and at the order statistics of the data. Such parameters are defined in terms of the data, and so are not the usual kinds of parameters. Nonetheless, they are unknown quantities, and so in the de Finetti theory one can deal with them just as with any other unknown or 'random' quantities. Because the sequence of observable random quantities, X_i, is viewed as exchangeable, it follows from the general form of the theorem of de Finetti, that one is acting as though one had a distribution π on the space of all possible distribution functions, F. In principle the situation is as follows. Given the data, the prior distribution π is updated, as usual, to become a posterior distribution π^*, and posterior predictive probabilities for future observables can be obtained by taking expectations with respect to π^*. For example, if ties have probability 0, then

$$Pr\{X_{n+1} \in I_i \mid data\} = E[\theta_i \mid data] = E[F(x_{(i)}) - F(x_{(i-1)}) \mid data],$$

where in this equation F is the empirical distribution function for the unsampled population[3], and where the expectation is taken with respect to the posterior distribution of F. Thus despite the fact that the θ_i depend upon the data for their definition, in principle their posterior expectation can be defined in terms of the 'parameter' F, just as in conventional parametric statistics. Here, practically speaking, F is simply the empirical distribution for the entire finite population of the X_i, for large N, and plays the same role as does $\bar{X}$ for a Bernoulli sequence. The only aspect that is more subtle is the fact that because we are dealing with the huge space of all possible empirical distributions, it is difficult analytically to specify the prior and posterior distributions π and π^*, respectively. However, the parametric model of Hill (1987b) makes it clear just what these distributions are.

The two cases we have considered here, namely the 0-1 case, and the fully nonparametric case, are in fact the extreme cases with respect to complexity of the underlying model. Much of statistical inference and prediction takes place in

[3]This distinction is necessary when N is finite

an intermediate case, namely of a conventional parametric model for real-valued observations. However, such intermediate cases can be considered in much the same way. Consider, for example, the case of an exponential model for data. Let the parameter be taken to be, say, α, the expectation of the exponential distribution. Again imagine only a finite population of values, X_i, for $i = 1, \ldots, N$, and consider $\bar{X}$ for this population. Then one's apriori distribution for α is approximately one's prior distribution for $\bar{X}$, and conditional upon $\bar{X}$, the observations are approximately independently distributed according to an exponential distribution with 'parameter' $\bar{X}$.

The final point I wish to discuss concerns the role of Bayesian data analysis with respect to A_n. In Hill (1987b) two theorems are proved. The first gives a simple parametric model, called a nested splitting process, that gives rise to A_n exactly. The second shows that from a subjectivistic point of view, A_m holds in sampling from complex mixtures of distributions, where here m represents the number of groups or types formed from the n observations via data analysis. For example, in sampling from the population of cetaceans (whales, porpoises, dolphins) the sample animals may be classified according to species (or other variables) into m groups. From my point of view, such classification should be done by a form of data analysis. After performing such classification, the statistical problem can be reduced to one concerning the random effects model in the analysis of variance. See Box and Tiao (1973), Hill (1965, 1967, 1977, 1980b), and Lindley and Smith (1972) for Bayesian analysis of such models.

In Hill (1988) a theory of Bayesian data analysis is put forth in which, because of computational complexity, or because of thoughts that are triggered off during the analysis of the data, a departure is made from the classical Bayesian theory in which models and prior distributions are all specified before seeing the data. I believe that this modification is essential in order to make the classical Bayesian approach more realistic in applications. Any scientist worth his salt would play with his data, analysing it in a variety of ways, and giving free rein to his imagination and creativity. As argued in Hill (1985), classical non-Bayesian theory breaks down completely in connection with such data analysis, since all probabilities would have to be conditional on the exact procedures employed, including their order, and even the thoughts that cross one's mind. This also poses a challenge for the Bayesian approach. However, one can view the data-analytic procedures as occurring prior to the point at which the Bayesian analysis, proper, begins; and it should be observed that the Bayesian theory has always had an arbitrary element in it as to the time point at which one proceeds to make a formal Bayesian analysis. I have suggested that often the appropriate point is following the process of data analysis. After one has reformulated old models, or formulated entirely new models, by means of data analysis, one can proceed with the classical form of Bayesian reasoning, including robustness and sensitivity analysis, as in Berger (1984), Hill (1980b). See also Hacking (1967) and Smith (1986).

In the context of A_n and H_n this means that the 'parameters' θ_i and γ_i are

actually the results of such Bayesian data analysis, as discussed in Hill (1987b). This need not, however, change the basic interpretation of these quantities. As shown in Hill (1988), a generalization of the restricted likelihood principle of Hill (1987a) remains valid in the context of data analysis. Of course classical non-Bayesian reasoning, for example conventional asymptotic theory, becomes entirely irrelevant. However, in low dimensional problems one can still plot likelihood functions, and these may turn out to be sharp relative to 'apriori' distributions for the parameters introduced following the data analysis. The force of such a Bayesian analysis of data must depend upon an agreement amongst scientists that specific prior distributions and likelihood functions are pertinent to the problem, and can be considered on their own merits, even after the data has been observed. In high dimensional problems one must learn new techniques for the analysis and display of likelihood functions, as in Hill (1975) with an example concerning the tails of distributions. See Mosteller and Wallace (1964), and Hill (1987b, 1988) for examples of Bayesian data analysis.

4 CONCLUDING REMARKS

The initial intuition as regards A_n seems to be due to 'Student,' or at least Fisher (1939) implies that this is the case. Fisher then generalized the idea and interpreted it in a fiducial spirit, which Dempster (1963) crystrallized and called 'direct probability.' Note that for all three of these authors the justification for A_n seems to be purely intuitive. Thus none give anything vaguely representing a 'proof' for A_n, or suggest a way in which its coherency or rationale can be discussed, or even indicate in what circumstances it might or might not be appropriate. While I believe that sound intuition (or inspiration) is what all scientific progress ultimately comes from, it is nonetheless the case that a critical attitude is necessary, and that one must ask when and why A_n is sensible, and whether there are any qualifications and pitfalls associated with it. For example, one must ask immediately, when, if ever, should one be using conventional parametric models as opposed to A_n. It is in helping to understand such questions that I believe the subjective Bayesian approach plays a fundamental role.

Consider, for example, the case of a normal model with known standard deviation of 1 and unknown mean, θ. The fiducial argument of Fisher suggests that the pivotal quantity $(X - \theta)$ should continue to have the N(0,1) distribution even after X is replaced by its observed value x, which is a number. (The confidence argument of Neyman does not assume this, but provides no way of telling whether there is anything peculiar about the particular x for which the confidence is quoted. As is well known, there are many examples, such as the Fieller-Creasy example, in which such a procedure is patently absurd, in that the whole real line may have confidence 95 percent. More generally, such confidence procedures do not provide a way to allow one to deal with data for

which the conventional confidence level is obviously inappropriate based upon prior knowledge that is generally accepted. Thus the confidence argument, as applied in practice by sensible statisticians, is instead a conditional argument, i. e., it is conditional upon not getting data that is wildly contrary to prior knowledge. As shown in Hill (1985), the 'true' confidence coefficients, when adjusted to be conditional upon not getting such data, are necessarily both unknown and unknowable.)

The Bayesian argument goes far beyond this. It first tells one that if one has a prior distribution for θ which is sufficiently diffuse relative to the likelihood function, then in fact Fisher's fiducial conclusion is justified. (This fact, which Harold Jeffreys had been telling Fisher for years, seems finally to have been accepted by Fisher, as the previously quoted footnote of Fisher (1959, p. 51) seems to indicate.) Next, the Bayesian argument tells you that there are many situations in which instead the prior distribution may be sharp relative to the likelihood function, in which case the appropriate conclusions are quite different; and still again, there are important cases in which the prior and the likelihood are of comparable magnitude. Thus one sees clearly in what situations the fiducial argument is relevant, and what the nature of its limitations are. See Hill (1974, p.570) for a mathematical discussion of the behavior of a posterior distribution for various kinds of extreme data.

The situation with regard to A_n is of the same general nature as that for a normal mean, except it is much more complicated. Thus the primary basis for A_n or H_n is (8), which really says that the observed values x_i are totally uninformative about $\underline{J}$. Once again the initial intuition comes from a form of Fisher's fiducial argument. Even if one finds his argument compelling, however, one would presumably want to put it into a broader context, including at least sampling without replacement and the case of ties. Thus the generalization from A_n to H_n, and from sampling with replacement to sampling from a finite population without replacement, is important, since the case of ties and of sampling without replacement is both more fundamental and more realistic. The Bayesian approach does not stop here, however, for the underlying assumptions, especially (8), are themselves only approximations. They are extremely valuable, because without such approximations there is nothing that one can do in a rational and logical way. But as with all things, they too are only approximations, and the trick is to learn when they are appropriate.

Finally, because A_n is a de Finetti coherent procedure, one knows that there are no operationally meaningful ways in which one can be made a loser by using A_n or H_n. Obviously, this property is a desirable one, but it is not sufficient to justify use of these procedures. Thus in addition to the internal coherency property, one wants also to know whether the procedure is 'reasonable,' that is to say, whether it corresponds to prior knowledge that is generally considered to be appropriate for the situation at hand, or for which a reasonable case can be made. In my opinion, it ordinarily is, but with a few qualifications.

Let me conclude by observing that A_n is supported by all of the major

approaches to statistical inference. It is Bayesian, fiducial, and even a confidence/tolerance procedure. It is simple, coherent, and plausible. It can even be argued, I believe, that when viewed in the context of Bayesian data analysis, A_n, along with H_n, constitute the best solution we now have to the problem of induction as formulated by Hume.

REFERENCES

Aitchison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis,* Cambridge University Press.

Berger, J. (1984), "The robust Bayesian viewpoint," in *Robustness of Bayesian Analysis,* J. Kadane, ed., North-Holland: Amsterdam, (with discussion), 321-372.

Berliner, L. Mark., and Hill, Bruce M. (1988), "Bayesian nonparametric survival analysis," *Journal of the American Statistical Association, 83,* 772-784 (with discussion).

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics,* 1, 353-355.

Boender, C. G. E., and Kan, A. H. G. Rinnooy (1987), "A multinomial Bayesian approach to the estimation of population and vocabulary size," *Biometrika, 74,* 849-856.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis,* Addison-Wesley, Reading.

Chang, C. (1988), *Bayesian Nonparametric Prediction Based on Censored Data,* University of Michigan Doctoral Dissertation.

Chen, Wen-Chen (1978), *On Zipf's Law*, University of Michigan Doctoral Dissertation.

Chen, Wen-Chen (1980), "On the weak form of Zipf's law," *Journal of Applied Probability,* 17, 611-622.

Csörgö, S., Deheuvels, P., and Mason, D. M. (1985), "Kernel estimates of the tail index of a distribution," *Annals of Statistics,* 13, 1050-1077.

de Finetti, B. (1937), "La prévision: ses lois logiques, ses sources subjectives," *Annales de l'Institut Henri Poincaré,* 7, 1-68.

de Finetti, B. (1974). *Theory of Probability, Vol. 1,* London: John Wiley & Sons, Inc.

Dempster, A. P. (1963), "On Direct Probabilities," *Journal of the Royal Statistical Society B, 25,* 100-114.

Diaconis, P., and Freedman, D. (1980), "Finite exchangeable sequences," *The Annals of Probability, 8,* 745-764.

Diaconis, P., and Freedman, D. (1981), "Partial exchangeability and sufficiency," *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions,* 205-236.

Dickey, J., and Kadane, J. (1980), "Bayesian decision theory and the simplification of models," in *Evaluation of Econometric Models,* J. Kmenta and J. Ramsey, eds., Academic Press, 245-268.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications, Third Edition, Revised Printing,* New York: John Wiley Sons.

Feller, W. (1971), *An Introduction to Probability Theory and its Applications, Volume 2, Second Edition,* New York: John Wiley Sons.

Ferguson, T. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics,* 1, 209-230.

Fisher, R. A. (1939), "Student,"*Annals of Eugenics, 9,* 1-9.

Fisher, R. A. (1948), "Conclusions Fiduciare,"*Annales de l'Institut Henri Poincaré,* 10, 191-213.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference,* Second Edition, New York: Hafner Publishing Co.

Good, I. J. (1965), *The Estimation of Probabilities,* MIT Research Monograph No. 30.

Hacking, I. (1967), "Slightly more realistic personal probability," *Philosophy of Science, 34,* 311-325.

Hartigan, J. (1983), *Bayes Theory,*, New York: Springer-Verlag.

Heath, D., and Sudderth, W. (1976), "de Finetti's theorem for exchangeable random variables," *The American Statistician, 30,* 188-189.

Hewitt, E., and Savage, L. J. (1955), "Symmetric measures on cartesian products," in *The Writings of Leonard Jimmie Savage-A Memorial Selection,* Published by The American Statistical Association and The Institute of Mathematical Statistics, 1981, 244-275.

Hill, B. M. (1965), "Inference about variance components in the one-way model," *Journal of the American Statistical Association,* 58, 918-932.

Hill, B. M. (1967), "Correlated errors in the random model," *Journal of the American Statistical Association,* 62, 1387-1400.

Hill, B. M. (1968), "Posterior distribution of percentiles: Bayes theorem for sampling from a finite population," *Journal of the American Statistical Association,* 63, 677-691.

Hill, B. M. (1969), "Foundations for the theory of least squares," *Journal of the Royal Statistical Society, Series B, 31,* 89-97.

Hill, B. M. (1970), "Zipf's law and prior distributions for the composition of a population," *Journal of the American StatisticalAssociation,* 65, 1220-1232.

Hill, B. M. (1974a), "The rank frequency form of Zipf's law," *Journal of the American Statistical Association,* 69, 1017-1026.

Hill, B. M. (1975), "A simple general approach to inference about the tail of a distribution," *Annals of Statistics,* 3, 1163-1174.

Hill, B. M. (1977), "Exact and approximate Bayesian solutions for inference about variance components and multivariate inadmissibility," in *New Developments in the Application of Bayesian Methods,* ed. by A. Aykac and C. Brumat, North Holland, Chapter 9, 129-152.

Hill, B. M. (1979), "Posterior moments of the number of species in a finite population, and the posterior probability of finding a new species," *Journal of the American Statistical Association, 74,* 668-673.

Hill, B. M. (1980a), "Invariance and robustness of the posterior distribution of characteristics of a finite population, with reference to contingency tables and the sampling of species." In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, ed A. Zellner, North-Holland, 383-395.

Hill, B. M. (1980b), "Robust analysis of the random model and weighted least squares regression," in *Evaluation of Econometric Models*, ed. by J. Kmenta and J. Ramsey, Academic Press, 197-217.

Hill, B. M. (1980c), "On finite additivity, non-conglomerability, and statistical paradoxes," (with discussion) in *Bayesian Statistics*, J. M. Bernardo, M. H. Degroot, D. V. Lindley, A. F. M. Smith, eds., University Press: Valencia, Spain, 39-66.

Hill, B. M. (1981), "A theoretical development of the Zipf (Pareto) law," in *Studies on Zipf's Laws*, ed. by H. Guiter, Sprachwissenschaftliches Institute, Ruhr-Universitat, Bochun.

Hill, B. M. (1985), "Some subjective Bayesian considerations in the selection of models," *Econometric Reviews 4, No. 2*, 191-288 (with discussion).

Hill, B. M. (1987a), "The validity of the likelihood principle," *The American Statistician, 41*, 95-100.

Hill, B. M. (1987b), "Parametric models for A_n: Splitting Processes and Mixtures," Unpublished, Department of Statistics, The University of Michigan.

Hill, B. M., (1988a), "A theory of Bayesian data analysis," to appear in *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, J. Hodges, A. Zellner, eds., North-Holland, 1989.

Hill, B. M. (1988b), "De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference," in *Bayesian Statistics 3*, J. M. Bernardo, M. H. Degroot, D. V. Lindley, and A. F. M. Smith, eds., Oxford University Press, 211-241 (with discussion).

Hill, B. M. and Lane, David (1985), "Conglomerability and countable additivity," *Sankhyá, 47, Series A*, 366-379.

Hill, B. M., Lane, David, and Sudderth, William (1980), "A strong law for some generalized urn processes," *The Annals of Probability*, 8, 214-226.

Hill, B. M., Lane, David, and Sudderth, William (1987), "Exchangeable urn processes," *The Annals of Probability*, 15, 1586-1592.

Hill, B. M. and Woodroofe, M. (1975), "Stronger forms of Zipf's law," *Journal of the American Statistical Association*, 70, 212-219.

Hoppe, F. (1987), "The sampling theory of neutral alleles and an urn model in population genetics," *Journal of Mathematical Biology*, 25, 123-159.

Hume, David (1748), *An Enquiry Concerning Human Understanding*, London.

Ijiri, Y., and Simon, H. A. (1975), "Some distributions associated with Bose-Einstein statistics," Proceeding of the National Academy of Science, USA, 72, 1654-1657.

Jeffreys, H. (1957), *Scientific Inference*, Second Edition, Cambridge University Press.

Jeffreys, H. (1961), *Theory of Probability*, Third Edition, Oxford at the Clarendon Press.

Johnson, W. E. (1932), "Probability: the deductive and inductive problems," *Mind, 49,* 409-423. [Appendix on pages 421-423 edited by R. B. Braithwaite].

Kingman, J. F. C. (1975), "Random discrete distributions," *Journal of the Royal Statistical Society, Series B, 37,* 1-22 (with discussion).

Lane, D., and Sudderth, W. (1978), "Diffuse models for sampling and predictive inference," *Annals of Statistics, 6,* 1318-1336.

Lane, D., and Sudderth, W. (1984), "Coherent predictive inference," *Sankhyá Ser. A, 46,* 166-185.

Lenk, P. (1984), *Bayesian Nonparametric Predictive Distributions,* Doctoral Dissertation, The University of Michigan.

Lewins, W. A., and Joanes, D. N. (1984), "Bayesian estimation of the number of species," *Biometrics, 40,* 323-328.

Lindley, D., and Smith, A. F. M. (1972), "Bayes estimates for the linear model," *Journal of the Royal Statistical Society, Series B,* 34, 1-41.

Luce, R. D., Narens, L. (1987), "Measurement scales on the continuum," *Science, 236,* 1527-1531.

Mandelbrot, B. B. (1982), *The Fractal Geometry of Nature,* W. H. Freeman and Co., New York.

Poincaré, H. (1912), *Calcul des Probabilités,* Deuxiéme Edition, Gauthier-Villars.

Ramakrishnan, S. and Sudderth, W. (1988), "A sequence of coin- toss variables for which the strong law fails," *American Mathematical Monthly, 95,* 939-941.

Rényi, A. (1970), *Probability Theory,* New York : American Elsevier.

Savage, L. J. (1972), *The Foundations of Statistics*, Second Revised Edition. New York: Dover.

Smith, A. F. M., (1986), "Some Bayesian thoughts on modelling and model choice," *The Statistician,* 35, 97-102.

Whitrow, G. J. (1980), *The Natural Philosophy of Time, Second Edition,* Oxford University Press.

Woodroofe, M. and Hill, B. M. (1975), "On Zipf's law," *Journal of Applied Probability, 12,* 425-434.

Zabell, S. L. (1982), "W. E. Johnson's sufficientness postulate," *The Annals of Statistics, 10,* 1091-1099.