Some History of the Hierarchical Bayesian Methodology

I.J. GOOD
Virginia Polytechnic Institute and State University

SUMMARY

A standard technique in subjective "Bayesian" methodology is for a subject ("you") to make judgements of the probabilities that a physical probability lies in various intervals. In the hierarchical Bayesian technique you make probability judgements (of a higher type, order, level, or stage) concerning the judgements of lower type. The paper will outline *some* of the history of this hierarchical technique with emphasis on the contributions by I.J. Good because I have read every word written by him.

Keywords: HIERARCHICAL BAYES; PARTIALLY-ORDERED PROBABILITIES; UPPER AND LOWER PROBABILITIES; EMPIRICAL BAYES; SPECIES FREQUENCIES; MULTINOMIAL ESTIMATION; PROBABILITY ESTIMATION IN CONTINGENCY TABLES; PROBABILITY DENSITY ESTIMATION; MAXIMUM ENTROPY; ML/E METHOD; TYPE II LIKELIHOOD RATIO; INFORMATION IN MARGINAL TOTALS; KINDS OF PROBABILITY; BAYES/NON-BAYES SYNTHESIS; HYPER-RAZOR OF DUNS AND OCKHAM.

1. PHILOSOPHY

In 1947, when few statisticians supported a Bayesian position, I had a non-monetary bet with M.S. Bartlett that the predominant philosophy of statistics a century ahead would be Bayesian. A third of a century has now elapsed and the trend supports me, but I would now modify my forecast. I think the predominant philosophy will be a Bayes/non-Bayes synthesis or compromise, and that the Bayesian part will be mostly hierarchical. But before discussing hierarchical methods, let me "prove" that my philosophy of a Bayes/non-Bayes compromise or synthesis is necessary for human reasoning, leaving aside the arguments for the specific axioms.

Proof. Aristotelean logic is insufficient for reasoning in most circumstances, and probabilities must be imcorporated. You are therefore forced to make probability judgements. These subjective probabilities are

more directly involved in your thinking than are physical probabilities. This would be even more obvious if you were an android (and you cannot prove you are not). Thus subjective probabilities are required for reasoning. The probabilities cannot be sharp, in general. For it would be only a joke if you were to say that the probability of rain tomorrow (however sharply defined) is 0.3057876289. Therefore a theory of partially ordered subjective probabilities is a necessary ingredient of rationality. Such a theory is "a compromise between Bayesian and non-Bayesian ideas. For if a probability is judged merely to lie between 0 and 1, this is equivalent to making no judgment about it at all" (Good, 1976b, p.137). Therefore a Bayes/non-Bayes compromise or synthesis is an essential ingredient of a theory of rationality. *Quod erat demonstrandum*.

The notion of a hierarchy of different types, orders, levels, or stages of probability is natural (i) in a theory of physical (material) probabilities, (ii) in a theory of subjective (personal) probabilities, and (iii) in a theory in which physical and subjective probabilities are mixed together. I shall not digress to discuss the philosophy of kinds of probability. (See, for example, Kemble, 1941; Good, 1959; 1965, Chapter 2.) It won't affect what I say whether you believe in the real existence of physical (material) probability or whether you regard it as defined in terms of de Finetti's theorem concerning permutable (exchangeable) events.

I shall first explain the three headings leaving most of the elaborations and historical comments until later.

(i) *Hierarchies of physical probabilities.* The meaning of the first heading is made clear merely by mentioning populations, superpopulations, and super-duper-populations, etc. Reichenbach (1934/1949, Chapter 8) introduced hierarchies of physical probabilities in terms of random sequences, random sequences of random sequences, etc.

(ii) *Hierarchies arising in a subjective theory.* Most of the justifications of the axioms of subjective probability assume sharp probabilities or clear-cut decisions, but there is always some vagueness and one way of trying to cope with it is to allow for the confidence that you feel in your judgements and to represent this confidence by probabilities of a higher type.

(iii) *Mixed hierarchies.* The simplest example of a mixed hierarchy is one of two levels wherein a subjective or perhaps logical distribution is assumed for a physical probability. But when there are only two levels it is somewhat misleading to refer to a "hierarchy".

In case (i), Bayes's theorem is acceptable even to most frequentists; see, for example, von Mises (1942). He made the point, which now seems obvious, that if, in virtue of previous experience, something is "known" about the distribution of a parameter θ, then Bayes's theorem gives information about

the final probability of a random variable x whose distribution depends on θ. Presumably by "known" he meant "judged uncontroversially". In short he emphasized that a "non-Bayesian" can use Bayes's theorem in some circumstances, a point that was also implicit in Reichenbach's Chapter 8. The point was worth making in 1942 because statisticians had mostly acquired the habit of using Fisherian techniques which nearly always ignore the possibility that there might sometimes be uncontroversial approximate prior distributions for parameters. F.N. David (1949, pp. 71 & 72) even said that Bayes's *theorem* "is wholy fallacious except under very restrictive conditions" and "... at the present time there are few adherents of Bayes' theorem". von Mises (1942, p.157) blew it by saying that the notion that prior probabilities are non-empirical "cannot be strongly enough refuted". He certainly failed to refute them strongly enough to stem the expansion of modern forms of subjectivistic Bayesianism.

Some people regard the *uncontroversial* uses of Bayes's theorem, that is, those uses acceptable to von Mises, as a case of the empirical Bayes method. Others, such as R.G. Krutchkoff, use the expression "empirical Bayes" only for the more subtle cases where the prior is assumed to exist but drops out of the formula for the posterior expectation of θ. It was in this sense that A.M. Turing used the empirical Bayes method for a classified application in 1941. I applied his method with many elaborations in a paper published much later (Good, 1953) which dealt with the population frequencies of species of animals or plants or words. If, in a sample of N animals, there are n_r species each represented r times, we may call n_r the frequency of the frequency r. Of course $\Sigma r n_r = N$. Let q_r be the population probability of such a species. Turing argued that

$$\mathrm{E}(q_r) = \frac{(r+1)n_{r+1}}{\mathrm{N}n_r} \tag{1}$$

and I modified this formula to $(r+1)n'_{r+1}/(Nn'_r)$ where $n'_1, n'_2, \ldots$, is a smoothing of $n_1, n_2, \ldots$, and I generalized the argument to give formulae for the moments of the posterior distribution of q_r. It follows that, in another sample of size N, the total expected frequency of the set of species that were each represented r times (in the first sample) is about $(r+1)n'_{r+1}$, not rn_r as would be suggested by a naive application of the method of maximum likelihood. In particular the probability that the next animal or word that you meet will be one that you have not met before is approximately n_1/N and not the maximum likelihood estimate which is zero. The formula (1) was later obtained by Robbins (1956, p. 159) in relation to the almost identical problem of sampling a large collection of Poisson distributions. In fairness to Robbins

it should be noted that he had some of the philosophical ideas of the empirical Bayes method in Robbins (1951) though he did not name the method at that time.

Perhaps a statistical argument is not fully Bayesian unless it is subjective enough to be controversial, even if the controversy is between Bayesians themselves. Any subjective idea is bound to be controversial in spite of the expression "de gustibus non disputandum est" (concerning taste there is no dispute). Perhaps most disputes *are* about taste. We can agree to differ about subjective probabilities but controversies arise when communal decisions have to be made. The controversy cannot be avoided, though it may be decreased, by using priors that are intended to represent ignorance, as in the theories of Jeffreys and of Carnap. (Of course "ignorance" does not here mean ignorance about the prior.) All statistical inference is controversial in any of its applications, though the controversy can be negligible when samples are large enough. Some anti-Bayesians often do not recognize this fact of life. The controversy causes difficulties when a statistican is used as a consultant in a legal battle, for few jurymen or magistrates understand the foundations of statistics, and perhaps only a small fraction even of statisticians do. I think the fraction will be large by 2047 A.D.

Now consider heading (ii), in which at least two of the levels are logical or subjective. This situation arises naturally out of a theory of partially ordered subjective probabilities. In such a theory it is not assumed, given two probabilities p_1 and p_2, that either $p_1 \geq p_2$ or $p_2 \geq p_1$. Of course partial ordering requires that probabilities are not necessarily numerical, but numerical probabilities can be introduced by means of random numbers, shuffled cards etc., and then the theory comes to the same thing as saying that there are upper and lower probabilities, that is, that a probability lies in some interval of values. Keynes (1921) emphasized such a theory except that he dealt with logical rather than subjective probabilities. Koopman (1940a, b) developed axioms for such a theory by making assumptions that seemed complex but become rather convincing when you think about them. I think the simplest possible acceptable theory along these lines was given by Good (1950), and was pretty well justified by C.A.B. Smith (1961). (See also Good, 1962.) Recently the theory of partially-ordered probability has often been called the theory of qualitative probability, though I think the earlier name "partially ordered" is clearer. When we use sharp probabilities it is for the sake of simplicity and provides an example of "rationality of type 2" (Good, 1971c).

If you can say confidently that a logical probability lies in an interval (a,b) it is natural to think it is more likely to be near to the middle of this interval than to the end of it; or perhaps one should convert to log-odds to

express a clear preference for the middle. (Taking the middle of the log-odds interval is an invariant rule under addition of weight of evidence.) At any rate this drives one to contemplate the notion of a higher type of probability for describing the first type, even though the first type is not necessarily physical. This is why I discuss hierarchies of probabilities in my paper on rational decisions, Good (1952). Savage (1954, p.58) briefly discusses the notion of hierarchies of subjective probabilities, but he denigrates and dismisses them. He raises two apparent objections. The first, which he heard from Max Woodbury, is that if a primary probability has a distribution expressed in terms of secondary probabilities, then one can perform an integration or summation so as to evaluate a composite primary probability. Thus you would finish up with a sharp value for the primary probability after all. (I don't regard this as an objection.) The second objection that he raises is that there is no reason to stop at secondary probabilities, and you could in principle be led to an infinite hierarchy that would do you no good.

In Good, (1950, p. 41) I had said that higher types of probability might lead to logical difficulties but in Good (1952) I took the point of view that it is mentally healthy to think of your subjective probabilities as estimates of credibilities, that is, of logical probabilities (just as it is healthy for some people to believe in the existence of God). Then the primary probabilities might be logical but the secondary ones might be subjective, and the composite probability obtained by summation would be subjective also. Or the secondary ones might also be logical but the tertiary ones would be subjective. This approach does not deny Max Woodbury's point; in fact it might anticipate it. I regard the use of hierarchical chains as a technique helping you to sharpen your subjective probabilities. Of course if the subjective probabilities at the top of the hierarchy are only partially ordered (as they normally would be if your judgements were made fully explicit), the same will be true of the composite primary or type I probabilities after the summations or integrations are performed. Another development of the hierarchical approach in my 1952 paper is in relation to minimax decision functions. Just as these were introduced to try to meet the difficulty of using ordinary Bayesian decisions, one can define a minimax decision function of type II, to avoid using Bayesian decision functions of type II. (The proposal was slightly modified in Good, 1955.) Leonid Hurwicz (1951) made an identical proposal simultaneously and independently. I still stand by the following two comments in my paper: "... the higher the type the woollier the probabilities ... the higher the type the less the wooliness matters provided [that] the calculations do not become too complicated". (The hierarchical method must often be robust, otherwise, owing to the wooliness of the higher levels, scientists would not agree with one another as often as they do. This is

why I claimed that the higher wooliness does not matter much.) Isaac Levi (1973, p. 23), says "Good is prepared to define second order probability distributions..., and third order probability distributions over these, etc., until he gets tired". This was funny, but it would be more accurate to say that I stop when the guessed expected utility of going further becomes negative if the cost is taken into account.

Perhaps the commonest hierarchy that deserves the name comes under heading (iii). The primary probabilities, or probabilities of type I, are physical, the secondary ones are more or less logical, and the tertiary ones are subjective. Or the sequence might be: physical, logical, logical, subjective. In the remainder of my paper I shall discuss hierarchies of these kinds.

2. SMALL PROBABILITIES IN LARGE CONTINGENCY TABLES

I used a hierarchical Bayesian argument in Good (1956) (original version rejected in 1953 I am proud to say) for the estimation of small frequencies in a large pure contingency table with entries (n_{ij}). By a *pure* table I mean one for which there is no clear natural ordering for the rows or for the columns. Let the physical probabilities corresponding to the cells of the table be denoted by p_{ij}, and the marginals by $p_{i.}$ and $p_{.j}$. Then the amount of information concerning a row provided by the observation of a column can be defined as $\log [p_{ij}/(p_{i.}p_{.j})]$ and it seemed worth trying the assumption that this has approximately a normal distribution over the table as a whole. This turned out to be a readily acceptable hypothesis for two numerical examples that were examined. In other words it turned out that one could accept the loglinear model

$$\log p_{ij} = \log p_{i.} + \log p_{.j} + \epsilon$$

where ϵ has a normal distribution whose parameters can be estimated from the data. (This was an early example of a loglinear model. Note that if ϵ is replaced by ϵ_{ij} and its distribution is not specified, then the equation does not define a model at all.) If then a frequency n_{ij} is observed it can be regarded as evidence concerning the value of p_{ij}, where p_{ij} has a lognormal distribution. Then an application of Bayes's theorem gives a posterior distribution for p_{ij}, even when $n_{ij} = 0$. This seemed to me an interesting example of estimating the probability of an event that had never occurred, but the referee discouraged me from saying this, possibly because it sounded philosophical. As Jimmie Savage once remarked "'Philosophy' is a dirty ten-lettered word". The lognormal distribution was used as a prior for the parameter p_{ij} and the parameters in this distribution would now-a-days often be called hyperparameters. Perhaps this whole technique could be regarded as a non-

controversial use of Bayes's theorem. Incidentally, if it is assumed that $p_{ij}/(p_{i.}p_{.j})$ has a Pearson Type III distribution, the estimates turn out to be not greatly affected, so the method appears to be robust. (The calculation had to be iterative and was an early example of the EM method as pointed out by Dempster *et al*, 1977, p. 19.)

3. MAXIMUM LIKELIHOOD/ENTROPY FOR ESTIMATION IN CONTINGENCY TABLES

For ordinary and multidimensional *population* contingency tables, with some marginal probabilities known, the method of maximum entropy for estimating the probabilities in the individual cells leads to interesting results (Good, 1963). [The principle of maximum entropy was interpreted by Jaynes (1957) as a method for selecting prior distributions. Good (1963) interprets it as a method for formulating hypotheses; in the application it led to hypotheses of vanishing interactions of various orders. Barnard mentions that an early proposer of a principle of maximum entropy was Jean Ville in the Paris conference on the history and philosophy of science in 1949 but I have not yet been able to obtain this reference.] When there is a sample it is suggested in Good (1963, p. 931) that one might find the estimates by maximizing a linear combination of the log-likelihood and the entropy, that is, in the two-dimensional case, by maximizing an expression of the form $\Sigma(n_{ij} - \lambda p_{ij}) \log p_{ij}$, subject to constraints if the marginal probabilities are assumed. [Here (n_{ij}) is the sample and (p_{ij}) the population contingency table.] This technique could be adopted by a non-Bayesian who would think of λ as a "procedure parameter". A Bayesian might call it a hyperparameter because the ML/E method, as we may call it, is equivalent to the maximization of the posterior density when the prior density is proportional to $\Pi p_{ij}^{-\lambda p_{ij}}$. This method has been investigated by my ex-student Pelz (1977). I believe that the best way to estimate the hyperparameter λ is by means of the method of cross-validation or predictive sample reuse, a method that could also be used for comparing the ML/E method with other methods (Good, 1979c). We intend to try this approach.

4. MULTINOMIAL DISTRIBUTIONS

Some hierarchical models that have interested me over a long period are concerned with multinomials and contingency tables, and these models received a lot of attention in my monograph on the estimation of probabilities from a Bayesian point of view (Good, 1965). (See also Good, 1964.) To avoid controversy about purely mathematical methods I there used the terminology of distributions of types I, II and III, without committing myself about whether the probabilities were physical, logical, or subjective. But, in a

Bayesian context, it might be easiest to think of these three kinds of probability as being respectively of the types I, II and III. My next few hundred words are based on Good (1965) where more details can be found although the present discussion also contains some new points.

The estimation of a binomial parameter dates back to Bayes and Laplace, Laplace's estimate being known as "Laplace's law of succession". This is the estimate $(r + 1)/(N + 2)$, where r is the number of successes and N the sample size. This was the first example of a shrinkage estimate. It was based on the uniform prior for the binomial parameter p. The more general conjugate prior of beta form was proposed by the actuary G.F. Hardy (1889). De Morgan (1847) (cited by Lidstone, 1920) generalized Laplace's law of succession to the multinomial case where the frequencies are (n_i) $(i = 1, 2,\ldots,t)$. (I have previously attributed this to Lidstone.) De Morgan's estimate of the i^{th} probability p_i was $(n_i + 1)/(N + t)$ which he obtained from a uniform distribution of $(p_1, p_2,\ldots,p_t)$ in the simplex $\Sigma\, p_i = 1$ by using Dirichlet's multiple integral. The estimate is the logical or subjective expectation of p_i and is also the probability that the next object sampled will belong to the i^{th} category. The general Dirichlet prior, proportional to $\Pi p_i^{k_i - 1}$, leads to the estimate $(n_i + k_i)/(N + \Sigma k_j)$ for p_i. But if the information concerning the t categories is symmetrical it is adequate, at the first Bayesian level, to use the prior proportional to Πp_i^{k-1} which leads to the estimates $(n_i + k)/(N + tk)$. In fact we can formulate the Duns-Ockham hyper-razor as "What can be done with fewer (hyper)parameters is done in vain with more". ("Ockham's razor" had been emphasized about twenty years before Ockham by the famous medieval philosopher John Duns Scotus.) We can regard k both as a flattening constant or as the hyperparameter in the symmetric Dirichlet prior. The proposal of using a continuous linear combination of Dirichlet priors, symmetric or otherwise, occurs in Good (1965, p.25). Various authors had previously proposed explicitly or implicitly that a single value of k should be used but I am convinced that we need to go up one level. (Barnard tells me he used a combination of two beta priors in an unpublished paper presented at a conference in Bristol in about 1953 because he wanted a bimodal prior.)

The philosopher W.E. Johnson (1932) considered the problem of what he called "multiple sampling", that is, sampling from a t-letter alphabet. He assumed permutability of the N letters of the sample (later called "exchangeability" though "permutability" is a slightly better term). Thus he was really considering multinomial sampling. He further assumed what I call his "sufficientness postulate", namely that the credibility (logical probability) that the next letter sampled will be of category i depends only on n_i, t, and N, and does not depend on the ratios of the other $t - 1$ frequencies. Under these assumptions he proved that the probability that the next letter sampled will be

of category i is $(n_i + k)/(N + tk)$, but he gave no rules for determining k. His proof was correct when $t \geq 3$. He was presumably unaware of the relationship of this estimate to the symmetric Dirichlet prior. The estimate does not merely follow from the symmetric Dirichlet prior; it also implies it, in virtue of a generalization of de Finetti's theorem. (This particular generalization follows neatly from a purely mathematical theorem due to Hildebrandt and Schoenberg; see Good, 1965, p. 22.) De Morgan's estimate is the case $k = 1$. Maximum Likelihood estimation is equivalent to taking $k = 0$. The estimates arising out of the invariant priors of Jeffreys (1946) and Perks (1947) correspond to the flattening constants $k = ½$ and $k = 1/t$.

Johnson's sufficientness assumption is unconvincing because if the frequencies $n_2, n_3,\ldots,n_t$ are far from equal it would be natural to believe that p_1 is more likely to be far from $1/t$ than if $n_2, n_3,\ldots,n_t$ are nearly equal. Hence it seemed to me that the "roughness" of the frequency count (n_i) should be taken into account. Since roughness can be measured by a scalar I felt that k could be estimated from the sample (and approximately from its roughness), or alternatively that a hyperprior could be assumed for k, say with a density function $\phi(k)$. This would be equivalent to assuming a prior for the p_i's, with density

$$\int_0^\infty \frac{\Gamma(tk)\Pi p_i^{k-1}\phi(k)dk}{[\Gamma(k)]^t}$$

Those who do not want to assume a hyperprior could instead estimate k say by Type II Maximum Likelihood or by other methods in which the estimate of k is related to $X^2 = \frac{t}{N}\Sigma\,(n_i - N/t)^2$. These methods were also developed by Good (1965, 1966, 1967). Good (1967) was mainly concerned with the Bayes factor, provided by a sample (n_i), against the null hypothesis $p_i = 1/t$ $(i = 1, 2,\ldots,t)$. The estimation of the cell probabilities p_i was also covered. (It seems to me to be usually wrong in principle to assume distinct priors, given the non-null hypothesis, according as you are doing estimation or significance testing, except that I believe that more accurate priors are required for the latter purpose.) The null hypothesis corresponds to the complete flattening $k = \infty$ and we may denote it by H_∞. Let H_k denote the non-null hypothesis that the prior is the symmetric Dirichlet with hyperparameter k. Let $F(k)$ denote the Bayes factor in favour of H_k as against H_∞, provided by a sample (n_i). (See Good, 1957. p. 862; or 1967, p. 406.) If k has a hyperprior density $\phi(k)$, then the Bayes factor F against H_∞ is

$$F = \int_0^\infty F(k)\phi(k)dk\ ,$$

$\phi(k)$ must be a proper density, otherwise F would reduce to 1, in other words the evidence would be killed. This is an interesting example where impropriety is a felony. One might try to be noncommittal about the value of k and the usual way of being noncommittal about a positive parameter k is to use the Jeffreys-Haldane density $1/k$ which is improper. This can be approximated by the log-Cauchy density which has the further advantage that its quantiles are related in a simple manner to its hyperhyperparameters (Good, 1969, pp. 45-46). One can determine the hyperhyperparameters by guessing the upper and lower quartiles of the repeat rate Σp_i^2, given the non-null hypothesis, and thereby avoid even a misdemeanour. The Bayes factor F is insensitive to moderate changes in the quartiles of the log-Cauchy hyperprior, and the estimates of the p_i's are even more robust. If you prefer not to assume a hyperprior then a type II or second order or second level Maximum Likelihood method is available because $F(k)$ has a unique maximum F_{max} if $X^2 > t - 1$. This was conjectured by Good (1965, p. 37) largely proved by Good (1975) and completely proved by Levin and Reeds (1977). Other methods of estimating k are proposed by Good (1965, pp. 27, 33, 34) and by Bishop, Fienberg and Holland (1975, Chapter 12). When a hyperparameter is estimated the latter authors call the method "pseudo-Bayesian". It is an example of a Bayes/non-Bayes compromise.

F_{max} is an example of a Type II (or second order or second level) Likelihood Ratio defined in terms of the hyperparametric space which is one-dimensional. Hence the asymptotic distribution of F_{max} is proportional to a chi-squared with one degree of freedom. In 1967 the accuracy of this approximation was not known but it was found to be fairly accurate in numerous examples in Good and Crook (1974), even down to tail-area probabilities as small as 10^{-16}. We do not know why it should be an adequate approximation in such extreme tails.

5. INDEPENDENCE IN CONTINGENCY TABLES

Good (1965) began the extension of the multinomial methods to the problem of testing independence of the rows and columns of contingency tables, and this work was continued in Good (1976a) where extensions to three and more dimensions were also considered. But I shall here consider only ordinary (two-dimensional) tables with r rows and s columns. The case $r = s = 2$ is of special interest because 2×2 tables occur so frequently in practice.

As is well known outside Bayesian statistics, there are three ways of sampling a contingency table, known as sampling Models 1, 2 and 3. In Model 1, sampling is random from the whole population; in Model 2, the row totals (or the comumn totals) are fixed in advance by the statistician; and in Model 3 both the row and column totals are fixed. Model 3 might seem unreasonable

at first but it can easily arise. Denote the corresponding Bayes factors against the null hypothesis H of independence by F_1, F_2, and F_3. But in our latest model it turns out that $F_1 = F_2$ because in this model the fixing of the row totals alone provides no evidence for or against H. The model also neglects any evidence that there might be in the order of rows or of columns; in other words we restrict our attention in effect to "pure" contingency tables. This is of course, also done when X^2 or the likelihood-ratio statistic is used.

The basic assumption in the analysis is that, given the non-null hypothesis $\bar{H}$, the prior for the physical probabilities p_{ij} in the table is a mixture of symmetric Dirichlet's. (Previous literature on contingency tables had discussed symmetric Dirichlet distributions but not mixtures.) From this assumption F_1 and F_3 can be calculated. We can deduce FRACT (the factor against H provided by the row and column totals alone, in Model 1) because FRACT $= F_1/F_3$. A large number of numerical calculations have been done and were reported in Crook and Good (1980). We found that FRACT usually lies between ½ and 2½ when neither of the two sets of marginal totals is very rough and the two sets are not both very flat, and we gave intuitive reasons for these exceptions. We did not report the results for 2×2 tables in that paper but we have done the calculations for this case with the sample size $N = 20$. We find, for example, with our assumptions, that FRACT = 1.48 for the table with margins [5,15;7,13]; FRACT = 2.53 for [10,10;10,10]; FRACT = 2.56 for [1,19;2,18]; and FRACT = 8.65 for the extreme case [1,19;1,19].

If the mixture of Dirichlet's is replaced by a single symmetrical Dirichlet with hyperparameter k, then F_3 is replaced by $F_3(k)$, and $\max_k F_3(k)$ is a Type II Likelihood Ratio. Its asymptotic distribution again turns out to be fairly accurate in the extreme tail of the distribution, even down to tail-area probabilities such as 10^{-40}. The unimodality of $F_3(k)$ when $X^2 > (r - 1)(s - 1)$ has yet to be proved, but is well supported by our numerical results.

I noticed only as recently as May 1978 that the consideration of contingency tables sheds light on the hyperprior ϕ for multinomials. This was first reported in Good (1979b). We write $\phi(t,k)$ instead of $\phi(k)$ to indicate that it might depend on t as well as k. The prior for a t-category multinomial is then $D^*(t)$ where

$$D^*(t) = \int_0^\infty D(t,k)\phi(t,k)dk$$

and where $D(t,k)$ denotes the symmetric Dirichlet density. Our assumption of $D^*(rs)$, given $\bar{H}$ and Model 1, implies the prior $\int_0^\infty D(r,sk)\phi(rs,k)dk$ for the row totals. But, if the row totals alone contain no evidence concerning H, this must be mathematically independent of s and it can be deduced that $\phi(t,k)$ must be of the form $\psi(tk)/k$. Strictly therefore some of the calculations in

Good and Crook (1974) should be repeated, but of course the distribution of the Type II Likelihood Ratio is unaffected, and we have reason to believe the remaining results are robust. This example shows how logical arguments can help to make subjective probabilities more logical. Logical probabilities are an ideal towards which we strive but seldom attain.

A spin-off from the work on contingency tables has been the light it sheds on the classical purely combinatorial problem of the enumeration of rectangular arrays of integers (Good and Crook, 1977; Good, 1979a). This problem had not previously been treated by statistical methods as far as I know.

T. Leonard has used hierarchical Bayesian methods for analyzing contingency tables and multinomial distributions, but since he has attended this conference I shall leave it to him to reference his work in the discussion of the present paper.

6. PROBABILITY DENSITY ESTIMATION AND BUMP HUNTING

Probability density estimation has been a popular activity since at least the nineteenth century, but bump-hunting, which is closely related to it, is I think comparatively recent. There is a short discussion of the matter in Good (1950, pp. 86-87) where the "bumpiness" of a curve is defined as the number of points of inflexion, though half this number is a slightly better definition. The number of bumps was proposed as one measure of complexity, and the greater the number the smaller the initial probability of the density curve *ceteris paribus*.

In the 1970 Waterloo conference, Orear and Cassel (1971) said that bump-hunting is "one of the major current activities of experimental physicists". In the discussion Good (1971a) suggested the idea of choosing a density function f by maximizing $\Sigma \log f(x_i) - \beta R$, that is, log-likelihood minus a roughness penalty proportional to a measure R of roughness of the density curve. (Without the penalty term one gets $1/N$ of a Dirac function at each observation.) It was pointed out that the problem combines density estimation with significance testing. In Good (1971b) the argument is taken further and it is mentioned that $\exp(-\beta R)$ can be regarded as the prior density of f in function space. In this Bayesian interpretation β is a hyperparameter. (There were 21 misprints in this short article, owing to a British dock strike.) The method was developed in considerable detail by Good and Gaskins (1971, 1972) and applied to two real examples, one relating to high-energy physics and the other to the analysis of chondrites (a common kind of meteorite containing round pellets) by Good and Gaskins (1979). In the latter work, the estimation of the hyperparameter was made by means of non-Bayesian tests of goodness of fit so as to avoid controversies arising out of the use of

hyperpriors.

Leonard (1978, p. 129) mentions that he hopes to report a hierarchical form of his approach to density estimation. Also he applies his method to the chondrite data, and he brought this data to my attention so that our methods could be compared.

7. INFERENCE ABOUT NORMAL DISTRIBUTIONS AND LINEAR MODELS

In 1969 I suggested to my student John M. Rogers that he might consider analogies of the multinomial work for the estimation of the parameters of multivariate normal distributions. It turned out that even the univariate problems were complicated and he completed his thesis without considering the multivariate problems. He considered the estimation of a (univariate) normal mean when the prior contains hyperparameters. The priors were of both normal and Cauchy form (Rogers, 1974) and the hyperparameters were estimated by type II maximum likelihood.

Meanwhile hierarchical Bayesian models with three or four levels or stages had been introduced for inferences about normal distributions and linear models by Lindley (1971) and by Lindley and Smith (1972.) A survey of these matters could be better prepared by Lindley so I shall say no more about them.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health, Grant Number R01 GM18770.

REFERENCES

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis* Harvard, Mass: M.I.T. Press.

CROOK, J.F. and GOOD, I.J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, II. *Ann. Statist.* (to be published).

DAVID, F.N. (1949). *Probability Theory for Statistical Methods*. Cambridge: University Press.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39**, 1-38 (with discussion).

GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

— (1952). Rational decisions. *J.Roy Statist. Soc. B* **14**, 107-114

— (1953). On the population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237-264.

— (1955). Contribution to the discussion on the Symposium on Linear Programming. *J. Roy. Statist. Soc. B.* **17**, 194-196.

— (1956). On the estimation of small frequencies in contingency tables. *J. Roy. Statist. Soc. B*, **18**, 113-124.

— (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28**, 861-881.

— (1959). Kinds of probability. *Science* **127**, 443-447.

— (1962). Subjective probability as the measure of a non-measurable set. In *Logic, Methodology and Philosophy of Science* (Nagel, E., Suppes, P., and Tarski, A. eds), 319-329.

— (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist,* **34**, 911-934.

— (1964). Contribution to the discussion of A.R. Thatcher Relationships between Bayesian and confidence limits for predictions. *J. Roy. Statist. Soc. B,* **26**, 204-205.

— (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Harvard, Mass: M.I.T. Press.

— (1966). How to estimate probabilities. *J. Inst. Math. Applics.* **2**, 364-383.

— (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. B* **29**, 399-431.

— (1969). A subjective analysis of Bode's law and an 'objective' test for approximate numerical rationality. *J. Amer. Statist. Assoc.* **64**, 23-66 (with discussion).

— (1971a). Contribution to the discussion of Orear and Cassel (1971), 284-286.

— (1971b). Nonparametric roughness penalty for probability densities. *Nature Physical Science* **229**, 29-30.

— (1971c). Twenty-seven principles of rationality. In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott. ed.) 123-127, Toronto: Holt, Rinehart and Winston.

— (1975). The Bayes factor against equiprobability of a multinomial population assuming a symmetric Dirichlet prior. *Ann. Statist.* **3**, 246-250.

— (1976a). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**, 1159-1189.

— (1976b). The Bayesian influence or how to sweep subjectivism under the carpet. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* **2** (C.A. Hooker and W. Harper, eds.) 125-174, Dordrecht, Holland: D. Reidel.

— (1979a). A comparison of some statistical estimates for the numbers of contingency tables, item C26 in "Comments, Conjectures, and Conclusions". In *J. Statist. Comput. Simul.* **8**, 312-314.

— (1979b). The contributions of Jeffreys to Bayesian statistics. In *Studies in Bayesian Econometrics and Statistics in Honor of Harold Jeffreys.* (A. Zellner, ed.), 21-34. Amsterdam: North Holland.

— (1979c). Predictive sample reuse and the estimation of probabilities. *J. Statist. Comput. Simul.* **9**, 238-239.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

— (1977). The enumeration of arrays and a generalization related to contingency tables. *Discrete Mathematics* **19**, 23-45.

GOOD, I.J. and GASKINS, R.A. (1971). Non-parametric roughness penalties for probability densities *Biometrika* **58**, 255-277.

— (1972). Global nonparametric estimation of probability densities. *Virginia J. of Science* **23**, 171-193

— (1979). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 42-73 (with discussion).

HARDY, G.F. (1889). In correspondence in Insurance Record. Reprinted in *Trans. Fac. Actuaries* **8** (1920), 174-182.

HURWICZ, L. (1951). Some specification problems and applications to econometric models, *Econometrics* **19**, 343-344 (abstract).

JAYNES, E.T. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106**, 620-630.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. (London), A.* **186**, 453-461.

JOHNSON, W.E. (1932). Appendix (ed. R.B. Braithwaite) to Probability: deductive and inductive problems. *Mind* **41**, 421-423.

KEMBLE, E.C. (1941). The probability concept. *Philosophy of Science* **8**, 204-232.

KEYNES, J.M. (1921). *A Treatise on Probability*. London: Macmillan.

KOOPMANM, B.O. (1940a). The basis of probability. *Bull. Amer. Math. Soc.* **46**, 763-764.

— (1940b). The axioms and algebra of intuitive probability. *Ann. Math.* **41**, 269-292.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc, B,* **40**, 113-146 (with discussion).

LEVI, I. (1973). Inductive logic and the improvement of knowledge. *Tech. Rep,* Columbia University.

LEVIN, B. and REEDS, J. (1977). Compound multinomial likelihood functions: proof of a conjecture of I.J. Good., *Ann. Statist.* **5**, 79-87.

LIDSTONE, G.J. (1920). Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Trans. Fac. Actuar.* **8**, 182-192.

LINDLEY, D.V. (1971). The estimation of many parameters. In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott; eds.) 435-455, (with discussion). Toronto: Holt, Rinehart and Winston.

LINDLEY, D.V. and SMITH, A.F.M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. B.* **34**, 1-41 (with discussion).

DE MORGAN, A. (1847). Theory of probabilities. *Encyclopaedia Metropolitana*, **2**, 393-490.

OREAR, J. and CASSEL, D. (1971). Applications of statistical inference to physics. In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott. eds.) 280-288 (with discussion). Toronto: Holt, Rinehart and Winston.

PELZ, W. (1977). *Topics on the estimation of small probabilities.* Ph D thesis, Virginia Polytechnic Institute and State University.

PERKS, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuar* **73**, 285-312.

REICHENBACH, H. (1949). *The Theory of Probability*. Berkeley: University of California Press.

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. 2nd Berkeley Symp.* 131-148. Berkeley: University of California Press.

— (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.* **1**, 157-163. Berkeley: University of California Press..

ROGERS, J.M. (1974). *Some examples of compromises between Bayesian and non-Bayesian statistical methods*. Ph. D. Thesis, Virginia Polytechnic Institute and State University.

SAVAGE, L.J. (1954). *The Foundations of Statistics*. New York: Wiley.

SMITH, C.A.B. (1961). Consistency in statistical inference and decision. *J. Roy. Statist. Soc. B.* **23**, 1-37 (with discussion).

VON MISES, R. (1942). On the correct use of Bayes's formula. *Ann. Math. Statist.* **13**, 156-165.