BAYESIAN STATISTICS AS HONEST WORK

by
Persi Diaconis

TECHNICAL REPORT NO. 204
SEPTEMBER 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
BAYESIAN STATISTICS AS HONEST WORK

by
Persi Diaconis

TECHNICAL REPORT NO. 204
SEPTEMBER 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ABSTRACT

A review of recent work on exchangeability is offered in the spirit of tolerance and respect for competing views that characterized the work of Jack Kiefer and Jerzy Neyman. The theme is this: No matter what paradigm you work in, if you do a piece of honest work, people will find uses for it. This will be illustrated by tracing the fruits of de Finetti's efforts to understand Hume's problem of induction. His resolution involves exchangeability and his well known theorem that every exchangeable sequence is a mixture of coin tossing. Subsequent generalizations and an application to the study of human vision are sketched.
1. Neyman, Keifer, and Bayes

The organizers of this conference have been kind enough to invite a presentation surveying recent progress in Bayesian statistics. Of course, neither Neyman nor Keifer was a Bayesian. I found them ever interested in discussing foundational issues. Let me illustrate.

Neyman and Bayes: When he began working in statistics, Neyman used the Bayesian formulation—most everyone did. Neyman (1977) gives a charming account of his change of view. He was teaching a course in statistics and trying to isolate just what properties of a prior distribution were critical in various applications. Often Bayes procedures only depend on the first few moments of the prior, or, in the case of asymptotic properties, on the support of the prior. As he continued to wrestle with how little could be assumed, Churchill Eisenhart, suggested that he forget completely about the prior and work on things that did not depend on any prior assumptions. As Neyman says, "This remark proved inspiring".

I first met Neyman when I began teaching statistics at Stanford in 1974. It took him several years to recognize me as more than a passing face. One evening, at a dinner after one of the joint Berkeley-Stanford Colloquia, I chanced to be seated next to Neyman. It was finally time to get to know each other. He turned to me and said, "so, what do you do?"

How on Earth does one answer such a question? I wanted to have a serious conversation with Neyman on something I knew about. After a few seconds
of calculation, I managed to answer "Bayesian statistics". His face fell and he answered with a clearly dissappointed "oh." We managed to find something else to talk about.

A few days later, I received a note from Neyman, and a copy of a paper he had written containing a vigorous defense of his foundational position. It is titled "Frequentist Probability and Frequentist Statistics." Clearly Neyman had not in any sense given up trying to communicate. The paper is in a special foundations issue of the philosophy journal "Synthese"; my first paper on de Finettis theorem appears in the same volume. I was proud to have a paper in the same journal as Neyman; this gave me courage to read his paper in a critical light. Believing I detected much concealed subjectivity, I wrote back. He subsequently sent me other papers clarifying his position and never again failed to recognize me. "Well, how is Mister Bayesian."

Over the years, I have continued to read Neyman's papers. For me, his great achievement is the intellectual clarity he brought to our field. There is never any vague talk about the mysteries of inference. He says what he means in clear, unequivocal mathematical language. "Talk" is held to a minimum and "honest work" is clearly delineated. I think that the best of Bayesian statistics, the work of de Finetti and Savage, for example, has this sort of intellectual clarity.

Keifer and Bayes: Jack Keifer always worked at the forefront of the paradigm that Neyman founded. He was an unequivocal frequentist. Yet he never stopped wrestling with the problems and paradoxs of inference. He has a long thoughtful paper in that same volume of Synthese. His papers on conditional inference represents years of work; dealing with one of the
most difficult challenges to the Neyman-Pearson-Wald Theory. I talked to Jack extensively during the tail end of his preparation of these papers. His erudition and care in studying the Bayesian position serves as my model of what it means to be an academic.

I would like to end this recollection of Keifer and Bayes by proposing an inferential problem. Jack would often reward me with limericks. For example, after I talked on statistical problems in ESP at his seminar, he wrote the following:

A young statistician named Persi
Proclaimed very loudly a curse, he
*  
said "DAMN," and with glee

"I'll DESTROY ESP -
And believe me, I'll show them no mercy!"

My inference problem concerns a limerick that Jack handed me after a talk on Bayesian statistics at Berkeley. He wrote:

Our speaker was P. Diaconis
On Him was the Bayesian Onus

The problem is, how did it end? The limerick is recorded on a slip of paper, safely put where I'd never lose it. I'm sure I'll find it someday, but for now, I invite guesses on the ending based on the evidence * and on the rest of this paper!
2. Bayesian Statistics as Honest Work

The controversies at the foundations of statistics seem different than other controversies in science. The Bayes/non-Bayes controversy has been actively fought for several hundred years with no suggestion for resolution currently in view. We have no potential crucial experiment that can decide between the subjectivists and objectivists. Indeed, we have theorems proving that "all roads lead to Rome."

The Neyman-Pearson lemma can be regarded as the first complete class theorem. In one version, it states that the only admissible tests for a simple hypothesis versus a simple alternative with zero-one loss are Bayes for some prior. Wald, Lecam, Stein, and Larry Brown have sweeping generalizations of this result. Roughly, to actually choose a procedure in any reasonable problem amounts to choosing a prior distribution. This means that computations performed by the Bayesian school are of potential use for frequentists.

The inter-relation goes deeper than that. Consider the estimation of vector of normal means with squared error loss. Complete class theorems say you have to choose a Bayes rule. Common sense seems to suggest that if we have to choose a prior, we may as well choose one to reflect whatever we know, or can guess about the mean. The work of Berger, Efron, and Morris, emphasizing the practical interplay between admissibility and prior knowledge has clearly played a large role in the popularity of Stein-like shrinkage estimators.

The above paragraphs point out that frequentists can make use of computations done by Bayesians. In joint work with David Freedman, I have argued thatBayesians can make use of frequentist properties of procedures such as consistency and robustness through a device we call the
"What if Method". The idea is simple, in trying to actually write down a prior distribution, subjective Bayesians have to go through various "elicitation procedures". When considering a specific prior \( \pi \), consider some possible data \( x \), compute the posterior \( \pi | x \) and Bayes rule \( \delta_{\pi}(x) \), and see if they are satisfactory. The mental exercise goes something like this: "what if the data comes out to be \( x \), this prior \( \pi \) would lead me to believe \( \pi | x \) and take action \( \delta_{\pi}(x) \) is that really satisfactory?" If not, then \( \pi \) is probably not a reasonable quantification of what is known.

The "what if method" is a variation of what Good (1950) has called, "the device of imaginary results." To some of us, this seems like a perfect characterization of objectivist statistics: a procedure is evaluated through its behavior on data that might have occurred.

The device of imaginary results might seem like the most useless kind of philosophizing to some. In its defense, let me say that it can suggest very concrete computations. Freedman and I have carried these out in the problem of estimating the center of symmetry in a location problem with unspecified symmetric errors. These computations are reviewed in Diaconis and Freedman (1983a) (1983b). It turned out that a straightforward use of the Dirichlet prior in this problem leads to an inconsistent Bayes rule. This (frequentist) inconsistency in a problem where a large number of standard procedures get the right answer suggests (yet again) that the Dirichlet is a strange prior. The inconsistency suggested that certain widely used frequentist procedures—m-estimators with re-descending \( \Psi \) functions—must also be inconsistent. This turned out to be true for some m-estimators, such as Tukeys biweight, without careful scaling.
The point of all of this is that philosophical questions can lead to honest work—to actually looking at data from a fresh point of view or doing a non-trivial piece of mathematics. When one has done an honest piece of work, one can depend on others, perhaps working in very different paradigms, putting that work to use. Here, I am making an inductive inference.

A Review of Recent Work on Exchangeability

I believe that de Finetti's original work on exchangeability was motivated by thinking about a problem in philosophy which I will call Hume's problem.

Problem: When is it reasonable to believe that the future will be like the past? How can we learn from experience?

de Finetti identifies reasonable belief with probability. Now of course it is not always reasonable to believe that the future will be like the past; past successes may force future failures. Still, in many circumstances we have no awareness of a hidden mechanism and our opinions are symmetric or exchangeable. If \( e_1, e_2, \ldots, e_n \) is a given pattern of future successes or failures (take \( e_i = 0 \) or \( 1 \)) then symmetry becomes \( p\{e_1, \ldots, e_n\} = p\{e_{\pi(1)} e_{\pi(2)} \ldots e_{\pi(n)}\} \) for all permutations \( \pi \). de Finetti proved that any such probability could be represented as a mixture of binomial distributions:

Theorem: If \( p \) is an exchangeable probability on coin tossing space then there exists a unique probability \( \mu \) on \([0,1]\) such that

\[
(1) \quad p\{e_1, \ldots, e_n\} = \int p^k (1-p)^{n-k} \mu(dp) \text{ with } k = \Sigma e_i
\]

After observing \( k \) successes in the first \( n \) trials, the mixing measure for
predictions about the future is proportional to

\[ p^k(1 - p)^{n-k} \mu(dp) \]

Calculus shows that as \( n \) increases, this becomes sharply peaked at \( k/n \). Thus, predictions about the future will be approximately consistent with independent identically distributed trials with parameter \( k/n \).

For simplicity, de Finetti's theorem has been stated for two valued variables. The result holds very generally. While the theorem was proved to answer a philosophical question, the result has applications throughout statistics and probability. Diaconis and Freedman (1983c) survey uses and extensions in Bayesian statistics. Kingman (1978) surveys uses in genetics, and Aldous (1983) surveys uses in probability. Koch and Spizzichino (1982) contains several further surveys. For de Finetti's views, I recommend de Finetti (1964) and Chapter 9 of de Finetti (1972).

Variations of de Finetti's theorem are available to characterize mixtures of many of the usual parametric models of statistics. Here is an example due to Adrian Smith (1981): When is an observed sequence a location scale mixture of normal distributions? If it is invariant under all rotations fixing the line through the vector of all ones. Freedman (1962), and Diaconis and Freedman (1979) give theorems characterizing mixtures of Markov chains. Very general theorems have been developed by Scandinavian statisticians under the leadership of P. Martin-lof and S. Lauritzan. A convenient references is Lauritzan (1982). When the theorems get sufficiently general, they merge with theorems of modern statistical mechanics, see Dynkin (1978), Georgii (1979), or Ruelle (1978). The interplay between physics and statistics is just starting to be sorted out. A full length survey of these and other developments is in Diaconis and Freedman (1983c).
The problem of deriving an appropriate version of de Finetti's theorem for two-dimensional arrays arose in the context of applied Bayesian statistics. We wanted to know how to characterize the usual normal models in a two-way analysis of variance. The simplest version of the problem involved random variables \( X_{ij} \) taking values zero or one. Such an array is called row exchangeable if the Law of \( X_{ij} \) is the same as the law of \( X_{\pi(i)j} \) for any permutation \( \pi \). Column exchangeability is defined similarly. The set of all row-column exchangeable probabilities is clearly a convex set and an appropriate version of de Finetti's theorem would be a neat description of the extreme points.

Observe that de Finetti's theorem for exchangeable, two-valued variables can be rephrased in the language of convex sets—the set of exchangeable variables is a convex simplex with coin-tossing measures as the extreme points. The integral representation (1) follows from the fact that every point in a compact convex subset of a metric space can be represented as a mixture of extreme points.

The problem for two-dimensional arrays has been beautifully solved by David Aldous (1981). In his solution, a typical extreme point is indexed by a function

\[
\phi: [0,1]^2 \rightarrow [0,1]
\]

To produce \( X_{ij} \) for a specified \( \phi \), generate independent uniform variables \( U_1, U_2, \ldots, V_1, V_2, \ldots \). Fill in \( X_{ij} \) as the result of flipping a \( \phi(U_i, V_j) \) coin. As \( \phi \) varies, it passes through all extreme points, and thus every row-column exchangeable process is a mixture of such \( \phi \)-processes.

David Freedman and I were hard at work on this problem while I was visiting Bell Labs in 1977. We could show that \( \phi \)-processes were extreme-points, but didn't know if there were any other extreme points. In the
midst of this, I heard about a problem in the psychology of perception. It turned out that φ-processes provided insight and counter examples to 20 year old conjectures.

The perception research was being carried out by Edgar Gilbert and Bella Julesz at Bell Labs. They were trying to isolate when and why the eye sees patterns such as background-foreground in nature. As part of this research, they generated random patterns from different distributions, put them side-by-side, and asked if people could see the difference. Many examples are presented in Julesz (1975).

Clearly, two patterns will be visually distinguishable if they have different densities (so one is dark and the other light). Patterns are also visually distinguishable if they have the same density but different correlations (so one is smooth and the other clustered). Julesz has been claiming for many years that all the eye can see in these experiments is captured by first and second ordered statistics.

Zero-one processes $X_{ij}$ can be used to generate black and white patterns like a checkerboard. The conjecture translates to the following: If $X_{ij}$ and $Y_{ij}$ are row-column exchangeable and

$$
E[X_{ij}] = E[Y_{ij}] 
$$

Then the patterns formed by $X_{ij}$ and $Y_{ij}$ will be visually indistinguishable.

In our work on this problem, Freedman and I took $Y_{ij}$ as fair coin-tossing. We showed that simple properties of φ gave processes with the same low order moments as coin-tossing:

**Theorem**

a) A φ-processes has $E[X_{ij}] = \frac{1}{2}$ iff $\int \phi(x,y)dx \, dy = \frac{1}{2}$

b) A φ-processes has $E[X_{ij}X_{kl}] = \frac{1}{4}$ iff $\int \phi(x,y)dy = \frac{1}{2} = \int \phi(x,y)dx$ a.e.
c) A $\phi$-process satisfying condition (b) has the same third-order probabilities as fair coin tossing—the chance of any three $X_{ij}$ being one is $1/8$.

d) A $\phi$-process with the same fourth-order probabilities as fair coin tossing is fair coin tossing.

The theorem implies that counter-examples had to be found by searching for $\phi$-functions satisfying (b) of the theorem. An example is shown in the figure below.

A COUNTEREXAMPLE TO THE JULESZ CONJECTURE

The two pictures are clearly visually distinguishable. The underlying
\( \phi \)-function can be represented by giving its values at each point of the unit square:

\[
\begin{array}{cc}
     & 0 & 1 \\
\hline
0 & 1 & 0 \\
1 & 0 & 1
\end{array}
\]

This clearly satisfies \( b \) (and thus \( c \)) so the patterns have the same first, second, and third order statistics as fair coin tossing. An alternative description of the process is this: Fill out the first row of the matrix by fair coin tossing. To fill out successive rows, flip a coin once for each row, if heads, copy the first row. If tails, copy the opposite of the first row mod 2. A third description, given earlier than ours, is in Gilbert (1980).

My paper with Freedman shows dozens of other pictures that run from being easy to distinguish through difficult to distinguish. The pictures allow research to begin on why the conjecture fails.

The point of the example for this discussion is the interplay between problems in philosophy (de Finetti's Theorem), applied statistics (two-way analysis of variance), mathematics (convex sets in infinite dimensions), and psychology. Many other areas of Bayesian statistics can be shown in a similar light, as honest work, yielding results of interest in other areas of statistics and science. Perhaps because of this, the controversy between Bayesians and frequentists seems to have lost its ability to polarize. This gives us time to focus on the coming controversy—the one between those who think the computer has taken over and those who only proceed theoretically. I am sorry that we don't have Keifer and Neyman on board to help resolve it.
Addenda

For those having read this far, I conclude with a report that the speaker following me (Erich Lehman) opened his talk with:

Our speaker was P. Diaconis

On him was the Bayesian Onus

though his logics confusing

he's always amusing

this limerick here is his bonus
REFERENCES


