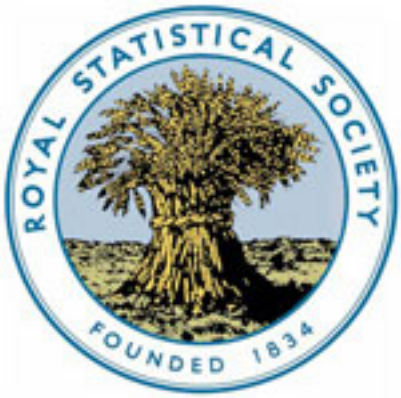

J. R. Statist. Soc. A (1984),
147, *Part* 2, *pp.* 278–292

Present Position and Potential Developments: Some Personal Views

Statistical Theory
The Prequential Approach

By A. P. DAWID

[*The Chairman of the Institute of Statisticians*, Dr W. G. Gilchrist, *in the Chair*]

SUMMARY

The prequential approach is founded on the premiss that the purpose of statistical inference is to make sequential probability forecasts for future observations, rather than to express information about parameters. Many traditional parametric concepts, such as consistency and efficiency, prove to have natural counterparts in this formulation, which sheds new light on these and suggests fruitful extensions.

Keywords: PROBABILITY FORECASTING; PREQUENTIAL PRINCIPLE; CONSISTENCY; EFFICIENCY; LIKELIHOOD

1. INTRODUCTION

An anniversary is naturally a time to look back over the path we followed to arrive where we are, in the hope that the past may shed light on the present. But it is also a time to look ahead, and our Conference Organizers have wisely chosen to emphasize the theme "Present Position and Potential Developments". I shall therefore reverse the usual direction of historical view, and take the opportunity to outline my personal vision of an approach to Statistical Theory which will, I hope, provide a fertile soil in which many ideas that are now current can develop further, and new ideas can germinate. Such a forward view can, I believe, shed as much light on the present and past as a more conventional historical review. Those who still wish to look back over their shoulders are referred to Dawid (1983a), which attempts to highlight the important ideas and philosophies which have driven Statistical Theory up to 1984. In this paper, however, I shall proceed on the basis that it is more interesting to look ahead rather than back.

This last assertion in fact forms the theme of my proposed framework for Statistics, variations on which will be taken up in the rest of this paper.

One of the major purposes of statistical analysis is to make *forecasts for the future*. It seems to me that we can shed much light on our subject by formalizing what is involved in making such forecasts, and by assessing our methods on their empirical success at this task. Another major purpose of statistical analysis is to offer suitable measures of the uncertainty associated with unknown events or quantities. I am persuaded by the arguments of de Finetti (1975) that the only concept needed to express uncertainty is Probability, and that this is most meaningful associated with genuine observables. Consequently, the "forecasts" I shall be considering will be *probability distributions* over future events.

The third main characteristic of the approach to be considered is the sequential nature of the forecasting task. At any time, we can make a forecast for tomorrow. Come tomorrow, we can observe the outcomes of the events which were the subject of today's forecasts and, with this additional experience to draw on, formulate our new forecasts for the following day. And so on, day after day, drawing on accumulating experience. I call this *prequential forecasting*. The name,

Present address: Department of Statistical Science, University College London, London WC1E 6BT, UK.

 0035–9238/84/147278 $2.00

like the subject, combines *pro*bability forecasting with *sequential pre*diction.

In this paper I want to show how prequential spectacles can be used to bring into focus a broad range of problems and concepts of Theoretical Statistics. Given any statistical model, we can convert it into a prequential forecasting system, in various possible ways—for example, by replacing the unknown parameter in the predictive distribution of the next observation by an estimate based on the data collected so far. We can then compare different ways of doing this, or different models, in terms of their prequential performance.

From this standpoint, we shall re-examine some of the traditional concepts of parametric inference, such as likelihood, consistency and efficiency. These all have natural prequential paraphrases, which turn out to be applicable in much greater generality than their classical counterparts. We also discuss the prequential assessment of the goodness of fit of a model to data. I shall try to persuade you that the prequential approach has great scope, both for deepening our understanding of a broad range of current concerns of Statistical Theory, and for suggesting entirely new areas worthy of further study.

2. PROBABILITY FORECASTING

US meteorologists have for a long time issued regular forecasts of the "probability of precipitation" in their own regions for the following (say) 12 hours. This has stimulated a fascinating literature in the meteorological journals on the theory and practice of *Probability Forecasting*: the field is reviewed by Dawid (1983b). An essential concern of this literature is the provision of methods for the empirical assessment and comparison of sequences of probability forecasts in the light of the outcomes of the forecast events.

These meteorological considerations, somewhat remote from current mainstream statistical theory, will provide the basis of the prequential approach to statistics. We shall suppose that data arrive in sequence, and view the statistician's task, at any time, as the appropriate manipulation of the data currently available so as to produce a specific probability distribution for the next observation. His success at this task will be judged by methods borrowed from Probability Forecasting.

This formalism may appear to be an uncomfortable straightjacket into which to squeeze statistical theory. The data may arrive *en bloc*, rather than in a natural order; if they come from a time-series, it may be impossible, or not obviously desirable, to analyse them at every point of time, or to formulate one-step ahead forecasts; and the restriction whereby all uncertainty about the next observation is to be encoded in a probability distribution, while acceptable to Bayesians, may not appeal to others. All these are valid *prima facie* objections; but I would respond by suggesting that, if you will tentatively join me in following through the implications of the prequential approach, you may find that it offers new insights enough to offset such disquiet.

3. PREQUENTIAL FORECASTING SYSTEMS

So now let $\mathbf{X} = (X_1, X_2, \ldots)$ be a sequence of uncertain quantities. At any time n, the prequential forecaster, with the values $\mathbf{x}^{(n)}$ of $\mathbf{X}^{(n)} = (X_1, X_2, \ldots, X_n)$ to hand, must issue a probability forecast distribution P_{n+1} for the next observation X_{n+1}.

A practising forecaster, such as a meteorologist, need only issue his or her forecast P_{n+1} in the light of the actual values, $\mathbf{x}_0^{(n)}$ say, of $\mathbf{X}^{(n)}$ that have materialized, and need only do so on day n. We restrict attention here, however, to *prequential forecasting systems* (PFS's). Such a system is defined by a rule which associates a choice of P_{n+1} with each value of n and with *any* possible set of outcomes $\mathbf{x}^{(n)}$ of $\mathbf{X}^{(n)}$.

One obvious way of constructing a PFS is by specifying a joint distribution P for $\mathbf{X}$, and taking for P_{n+1} the implied conditional distribution for X_{n+1} given $\mathbf{X}^{(n)}$. Thus P might, for example, model an assumed time-series structure for $\mathbf{X}$, or it might express the personal uncertainty of a subjectivist Bayesian. Conversely, any PFS, however constructed, is consistent in this way with a unique joint distribution P.

An important generalization of the above definition of a PFS, of considerable independent

interest, allows the data used in forecasting X_{n+1} to be more extensive than past outcomes of $\mathbf{X}^{(n)}$ only. Thus P_{n+1} might be a function of $(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$, with $\mathbf{y}^{(n)}$ the outcome of further variables $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)$. Almost all of the theory of PFS's extends with little difficulty to this case; however, we shall not be considering this extension here.

4. STATISTICAL FORECASTING SYSTEMS

Suppose we have a parametric family $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ of probability distributions for $\mathbf{X} = (X_1, X_2, \ldots)$. We might use this in a variety of ways to construct a PFS, P^* say, which we could call a *statistical forecasting system* (SFS) for $\mathscr{P}$. For example, at time n we might form an estimate $\tilde{\theta}_n$ of θ based on data $\mathbf{X}^{(n)} = \mathbf{x}^{(n)}$, and predict X_{n+1} as if $\tilde{\theta}_n$ were the true parameter: $P^*_{n+1}(\cdot) = P_{\tilde{\theta}_n}(\cdot \mid \mathbf{X}^{(n)} = \mathbf{x}^{(n)})$. This may be called the estimative or *plug-in* approach.

Alternatively, we might take a *Bayesian* approach and assign a prior distribution Π to θ. The appropriate forecast distribution for X_{n+1} is

$$P^*_{n+1}(\cdot) = \int P_\theta(\cdot \mid \mathbf{X}^{(n)} = \mathbf{x}^{(n)})\, d\Pi_n(\theta),$$

where Π_n is the posterior distribution for θ given data $\mathbf{X}^{(n)} = \mathbf{x}^{(n)}$. This is equivalent to constructing all forecasts by simple conditioning of the fixed joint probability distribution $P^* = \int P_\theta \, d\Pi(\theta)$.

Many other possible ways of constructing a SFS from $\mathscr{P}$ include, for example, fiducial predictive distributions. The quality of the PFS based on any familiar method of estimation or elimination of parameters can be used as a measure of the merit of that method.

Example 1. Under P_θ, let $X_i = 1$ with probability θ, $X_i = 0$ with probability $1 - \theta$, independently. After observing r 1's on n trials, the maximum likelihood plug-in forecast probability for $X_{n+1} = 1$ is r/n. The Bayesian forecast, based on a uniform prior, is $(r + 1)/(n + 2)$.

We can use a SFS to construct more complex k-step ahead forecasts by using the one-step forecasts as building blocks. Thus the forecast probability of $(X_{n+1} = 1, X_{n+2} = 0, X_{n+3} = 1)$ would be

$$(r/n)\{(n-r)/(n+1)\}\{(r+1)/(n+2)\}$$

for the above plug-in rule, rather than $(r/n)\{(n-r)/n\}(r/n)$. (Compare the corresponding Bayesian forecast for the uniform prior, viz.

$$\{(r+1)/(n+2)\}\{(n-r+1)/(n+3)\}\{(r+2)/(n+4)\},$$

which simply involves increasing both r and $n - r$ by 1.)

Example 2. Let $X_i \sim \mathcal{N}(\theta, 1)$ independently under P_θ. The m. l. plug-in forecast distribution for X_{n+1} is $\mathcal{N}(\bar{x}_n, 1)$, where $\bar{x}_n = n^{-1}(x_1 + \ldots + x_n)$. A fiducial forecast distribution could be formed by inversion of the pivotal quantity $X_{n+1} - X_n$, whose distribution is $\mathcal{N}(0, 1 + n^{-1})$, yielding $X_{n+1} \sim \mathcal{N}(\bar{x}_n, 1 + n^{-1})$. This agrees with the Bayesian solution for the improper uniform prior distribution.

Example 3. Suppose that $\theta = (\mu, \rho, \phi)$ $(0 \leqslant \rho \leqslant 1, \phi > 0)$, and that, under P_θ, the (X_i) form a Gaussian process with $E(X_i) = \mu$, var $(X_i) = \phi$, cov $(X_i, X_j) = \rho\phi$ $(i \neq j)$. This is a *non-ergodic* process, with parameters which are not even consistently estimable. The maximum likelihood estimators based on $\mathbf{X}^{(n)}$ are $\hat{\rho}_n \equiv 0$, $\hat{\mu}_n = \bar{X}_n$ (with distribution $\mathcal{N}(\mu, (\rho + n^{-1}(1-\rho))\phi)$), and $\hat{\phi}_n = n^{-1} \Sigma (X_i - \bar{X}_n)^2$ (with distribution $n^{-1}(1-\rho)\phi\, \chi^2_{n-1}$), which appear highly unsatisfactory. Nevertheless, we shall see below that the plug-in SFS based on these estimators is well behaved.

In full generality, we could regard *any* distribution P^* over $\mathbf{X}$ as a SFS for $\mathscr{P}$. The *Fundamental Question of Prequential Statistics* is then seen to be: What does it mean for P^* to be a "good" SFS for the full family $\mathscr{P}$? Before examining this, we must first consider the

Fundamental Question of Prequential Probability: What does it mean for P to be a "good" PFS for empirical data $\mathbf{x}$?

5. EMPIRICAL ADEQUACY

5.1. *The Prequential Principle*

Let P be a PFS for $\mathbf{X}$, $\mathbf{x} = (x_n)$ the realized outcomes of $\mathbf{X}$, and $\mathbf{P} = (P_n)$ the corresponding prequential forecast distributions produced by P. We may wish to assess the overall adequacy of P as a probabilistic explanation for $\mathbf{x}$. It seems desirable that any such assessment should depend on P only through the sequence $\mathbf{P}$ of forecasts that it in fact made. This requirement we shall call the *Prequential Principle*. It has an obvious analogy with the Likelihood Principle, in asserting the irrelevance of hypothetical forecasts that might have been issued in circumstances that did not, in fact, come about.

5.2. *Calibration and Jeffreys's Law*

For the case of binary (X_n), Dawid (1982a) introduced a criterion, *complete calibration*, for comparing the prequential probabilities $\mathbf{p} = (p_n)$ (where $p_n = P_n(X_n = 1) = P(X_n = 1 \mid \mathbf{x}^{(n-1)})$) with the outcomes $\mathbf{x}$. Informally, this requires that an average of the p's over a suitably selected subset of terms should agree, asymptotically, with the corresponding average of the x's. This criterion is justified by the fact that the above property holds with P-probability 1, so that its failure discredits P. In Dawid (1982b) it was further shown that, if P and Q both satisfy such a criterion for sufficiently many subsets, then $p_n - q_n \to 0$ as $n \to \infty$. Consequently, all non-rejected PFS's end up making the same forecasts.

This is an interesting and unexpected boon: in just those cases where we cannot choose empirically between several forecasting systems, it turns out that we have no need to do so! This property has implications for Philosophy of Science, giving some support to Popper's methodology, wherein a number of alternative hypotheses about Nature may be put forward, each being retained until it is refuted because its forecasts depart from observation. In our context, such refutation follows evidence that the complete calibration criterion is violated. This approach need not pick out, even asymptotically, a single "true model". (Indeed, there is no need even to assume the existence of an underlying "true" law generating the data.)Using it, we should, however, eventually be left only with PFS's that can all be expected to continue to make essentially identical predictions. I shall call this finding "Jeffreys's Law", after an admittedly distorted interpretation of Jeffreys (1938): "When a law has been applied to a large body of data without any systematic discrepancy being detected . . . the probability of a further inference from the law approaches certainty whether the law is true or not."

5.3. *Probability Integral Transform*

With continuous quantities (X_n), we can proceed as follows. Let $U_n = F_n(X_n)$, where F_n is the distribution function of P_n (assumed continuous). It is easily seen that, under P, the U_n are independent uniform $U[0, 1]$ variables (Rosenblatt, 1952). Consequently, we can reduce the assessment of P to the question of whether a sequence $\mathbf{u} = (u_n) = (F_n(x_n))$ "looks like" a random sample from $U[0, 1]$. This may be investigated by standard statistical tests, or by an extension of the calibration criterion: we should require that, for all suitably selected infinite subsets of the terms (u_n), the limiting empirical distribution formed from these terms is just $U[0, 1]$. Again, two PFS's both passing this test will be in asymptotic agreement. Note that *all* tests based on the (u_n) alone will be in accord with the Prequential Principle.

If the (u_n) seem to show some systematic departure from the "white noise" $U[0, 1]$ structure (in particular, if they exhibit dependence), it may be possible to find a successful PFS, R say, for predicting $\mathbf{u} = (u_n)$. This may then be used to "tune" P: the appropriate new prequential distribution function for X_n would be $G_n(F_n(\cdot))$, where G_n is that given by R. Of course, there can be no guarantee that such tuning would improve future forecasting performance. Successful applications of this procedure in forecasting software reliability are described by Keiller and

Littlewood (1984).

5.4. *Recursive Residuals*

For assessing the empirical adequacy of a statistical model $\mathscr{P}$, we can replace it by a suitable SFS P^* and proceed as above. (The most satisfactory choice would be an *efficient* SFS, as defined in Section 9 below). The corresponding (U_n) can be regarded as extending the idea of *recursive residuals* (Brown *et al.*, 1975), which have been extensively used for diagnostic checking of Box–Jenkins and other time-series models. Under any $P_\theta \in \mathscr{P}$ the (U_n) will behave (at least asymptotically) like independent $U[0, 1]$ variables, and so provide a diagnostic tool for model inadequacy.

In certain models it is possible to redefine the (U_n) slightly so as to be *exactly* independent $U[0, 1]$ variables under any $P_\theta \in \mathscr{P}$. This may be achieved through predictive fiducial pivotals in models with group-structure, or through the conditional probability integral transform (O'Reilly and Quesenberry, 1973) in those with suitable sufficient statistics. Such (U_n) should provide more reliable small-sample diagnostics.

6. GOODNESS OF FIT

Let (X_n) take values 0 and 1, and consider a PFS P under which $\mathbf{X}$ forms a Markov chain with transition probabilities $p_{00} = p_{11} = \frac{1}{4}, p_{01} = p_{10} = \frac{3}{4}$. The initial sequences of data (x_n) and forecasts (p_n), where $p_n = P_n(X_n = 1)$, might be

$$\left.\begin{array}{lcccccccccc} x_n: & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ p_n: & ? & \frac{3}{4} & \frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{1}{4} \end{array}\right\} \tag{6.1}$$

(We can ignore the first trial, for which p_1 needs further specification.)

The data may be summarized in the table of transition counts $(n_{ij}: i, j = 0, 1)$:

$$\begin{array}{ccc} 1 & 2 & 3 \\ 1 & 5 & 6 \\ 2 & 7 & 9 \end{array}, \tag{6.2}$$

where the rows correspond to $x_n = 0, 1$ and the columns to $x_{n+1} = 0, 1$.

How well does P fit these data? According to the Prequential Principle, the answer to this question should depend only on the values displayed in (6.1) or (6.2), and not on how the (p_n) were produced. In particular, let us consider a PFS Q under which $(X_n: n > 1)$ were *independent*, with $Q(X_n = 1)$ just happening to agree with p_n as displayed in (6.1). Then any assessment of the goodness of fit of Q should, according to the Principle, likewise be a valid assessment for P.

Now, under Q, the (X_n) divide into two sets of Bernoulli trials, Set 1 (containing trials 2, 3 and 9) having $Q(X_n = 1) = \frac{3}{4}$, and Set 2 (trials 4, 5, 6, 7, 8, 10) having $Q(X_n = 1) = \frac{1}{4}$. We might thus summarize the outcomes in (6.1) in the table:

$$\begin{array}{ccc} 1 & 2 & 3 \\ 1 & 5 & 6 \end{array}, \tag{6.3}$$

where the two rows are sets 1, 2 and the two columns correspond to outcomes 0, 1. For data (6.1), the entries in (6.3) happen to agree with those in (6.2). We can then calculate, in the usual way, the "expected frequencies", (m_{ij}) say, based on Q, yielding the table:

$$\begin{array}{ccc} \frac{3}{4} & 2\frac{1}{4} & 3 \\ 4\frac{1}{2} & 1\frac{1}{2} & 6 \end{array} \tag{6.4}$$

arranged in the same way as (6.3).

We could now calculate, for example, the Pearson chi-squared statistic

$$X^2 = \Sigma (n_{ij} - m_{ij})^2 / m_{ij} = 11.$$

Referred to tables of χ^2_2, the observed significance level is about $\frac{1}{2}$ per cent. This is a measure of the inadequacy of Q as an explanation of the data. By the Prequential Principle, it should likewise constitute such a measure for P.

Under Q, the counts in (6.3) are observations on Binomial variables, and standard theory justifies the asymptotic χ^2_2 distribution of X^2. Under P, however, the joint distribution of the counts in (6.2) is much more complicated. In particular (contrary to assertions of Basawa and Prakasa Rao, 1980), we do *not* have n_{01} and n_{11} independently binomial $\mathcal{B}$ $(3; \frac{3}{4})$ and $\mathcal{B}$ $(6; \frac{1}{4})$ after conditioning on $n_{0+} = 3$, $n_{1+} = 6$, since this information would tell us that the number of 0's in (6.1) must be 3 or 4, whence $n_{+0} \leqslant 4$, and so $n_{11} \geqslant 2$. A frequentist statistician might well be unwilling to accept the above "observed significance level" of X^2 calculated under Q as relevant to an assessment of P, without further evidence that the sampling distribution of X^2 under P itself was approximately χ^2_2. Fortunately, and perhaps surprisingly, this evidence is available: Billingsley (1961) shows that the asymptotic distribution of X^2 under P is indeed χ^2_2, thus supplying some frequentist support for the Prequential Principle.

Such support in fact remains available in very great generality. Many recent results on inference from dependent processes can be regarded as extending independence results to their prequential counterparts. Examples are the theorems of Dvoretzky (1972) (central limit theorems), Hill (1982) (strong laws of large numbers) and Shiryayev (1981) (absolute continuity of distributions). The work of Hill is particularly interesting in supplying a "meta-theorem" that justifies extending any theorem of a certain kind, that one might prove assuming independence, to apply equally, in its prequential version, to arbitrarily dependent processes. It may be conjectured that this, in turn, is a special case of a still more powerful meta-theorem. This would then provide strong frequentist support for the procedure of assessing prequential forecast against outcomes always *as if* the variables involved had been *independent* with the assigned distributions. This procedure greatly simplifies the problem of deciding, for example, whether sample departure from perfect calibration, as measured by various summary statistics, is "significant".

7. COMPARISON OF FORECASTING SYSTEMS

Consider a PFS P over $\mathbf{X}$, and suppose that the forecast distribution P_n has density function $p_n = p_n(x_n \mid \mathbf{x}^{(n-1)})$ with respect to some fixed underlying measure μ_n. The *prequential likelihood* of P for data $\mathbf{x}^{(n)}$ is defined as

$$L_{P,n} = \prod_{i=1}^{n} p_i,$$

viz. the joint density for $\mathbf{X}^{(n)}$ at $\mathbf{x}^{(n)}$ under the marginal distribution $P^{(n)}$ for $\mathbf{X}^{(n)}$ implied by P.

(If we were to consider the extended definition of a PFS which allows p_n to depend on the values of variables $\mathbf{Y}^{(n)}$ additional to $\mathbf{X}^{(n)}$, the above prequential likelihood would be a *partial likelihood* in the sense of Cox, 1975.)

Now let Q be another PFS, with prequential densities (q_n). We can compare P and Q, for data $\mathbf{x}^{(n)}$, by means of their *prequential likelihood ratio*

$$\Lambda_n = L_{Q,n}/L_{P,n} = \prod_{i=1}^{n} (q_i/p_i) = (dQ^{(n)}/dP^{(n)}).$$

Now, under P, (Λ_n) forms a non-negative martingale, and so, by standard theory, converges to a finite limit Λ with probability one. If then we find that the realized sequence (λ_n) of likelihood ratios tends to infinity, we shall be observing an event deemed almost impossible by P,

and would thus be justified in regarding P as discredited, in favour of Q—and conversely if $\lambda_n \to 0$. (Howard, 1975 has used this property as the basis of a criterion for the empirical adequacy of a PFS.)

When $P \ll Q$ (P is absolutely continuous with respect to Q), Λ is positive with probability one under both P and Q; in fact Λ^{-1} is a version of the Radon-Nikodym derivative dP/dQ. Consequently, we can only expect to discredit P in favour of Q, and not conversely. (Any attempt to test Q, by specifying an event A, with $Q(A) = 0$, whose occurrence would be taken to discredit Q, would equally discredit P when A occurred). We might thus regard Q as at least as good as P for prequential purposes, at any rate asymptotically. This is borne out by a Theorem of Blackwell and Dubins (1962): if $P \ll Q$ then, with P-probability one, the conditional distributions over the infinite future, given $\mathbf{X}^{(n)}$, made under each of P and Q, will be asymptotically indistinguishable.

If $P \sim Q$, (P and Q are mutually absolutely continuous), then we cannot expect ever to be able to make a definitive choice between P and Q, which, by the above result, will in any case yield asymptotically indistinguishable forecasts. (We see here another instance of Jeffreys's Law at work.) We can therefore regard such P and Q as *equivalent* PFS's.

8. CONSISTENCY

We now return to the statistical set-up of Section 4. Let P^* be a SFS for $\mathscr{P}$. Suppose that P^* dominates $\mathscr{P}$, so that $P_\theta \ll P^*$, all θ. In this case, the theorem of Blackwell and Dubins implies that, with probability one under any P_θ, the forecasts of the infinite future made by P^* will be asymptotically indistinguishable from those made under the correct P_θ. We might then call P^* *completely consistent* for $\mathscr{P}$.

If, for example, $\mathscr{P}$ is finite or countable, then a suitable P^* would be any Bayesian rule $P^* = \Sigma_\theta\, \pi(\theta)P_\theta$, with $\pi(\theta) > 0$ and $\Sigma_\theta\, \pi(\theta) = 1$.

In general, however, $\mathscr{P}$ will not be dominated, and complete consistency will be unattainable. Consider, for instance, the uniform-prior Bayesian PFS, P_B say, for Bernoulli trials (Example 1). Any of the model distributions P_{θ_0} corresponds to a prior distribution concentrated at θ_0, which is singular with respect to the uniform prior. Consequently, P_B and P_{θ_0} are mutually singular, for all θ_0. And, correspondingly, conditional on any data there exist future events whose predictive probabilities under P_B and P_{θ_0} are very different: an example is $A =$ "$r/n \to \theta_0$", for which $P_{\theta_0}(A \mid \mathbf{x}^{(n)}) \equiv 1, P_B(A \mid \mathbf{x}^{(n)}) \equiv 0$.

Nevertheless, we shall generally be able to attain (*simple*) *consistency*, defined as asymptotic equivalence of the one-step prequential forecasts P_n^* and $P_{n,\theta}$, with P_θ-probability 1 for every θ. For instance, in Example 2, $\bar{X}_n \to \theta$ with P_θ-probability 1, so that both prequential distributions there considered, $\mathcal{N}(\bar{x}_n, 1)$ and $\mathcal{N}(\bar{x}_n, 1 + n^{-1})$, will asymptotically agree with the true forecast $\mathcal{N}(\theta, 1)$. Clearly this behaviour will hold in some generality (given suitable regularity conditions) when θ is consistently estimable. We could plug in a consistent estimator of θ, or use a Bayesian SFS with a positive prior density (the posterior then placing the bulk of its mass near the true θ-value).

Example 3 (*continued*). Under P_θ, the forecast distribution $P_{n+1,\theta}$ for X_{n+1} given $\mathbf{X}^{(n)} = \mathbf{x}^{(n)}$ is normal with mean $\{n\rho\bar{x}_n + (1-\rho)\mu\}/\{1 + (n-1)\rho\}$, and variance $\phi[(1-\rho) + \{1 + (n-1)\rho\}^{-1}]$. For large n, this is approximately $\mathcal{N}(\bar{x}_n, (1-\rho)\phi)$ if $\rho > 0$, or $\mathcal{N}(\mu, \phi)$ if $\rho = 0$. The plug-in rule in either case is thus approximately

$$\mathcal{N}\left(\bar{x}_n, n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right).$$

The variance term in this converges, with probability one under P_θ, to the true asymptotic prediction variance $(1-\rho)\phi$. Consequently, although the m.l.e. is not consistent in the usual sense, it yields a consistent SFS. So, too, would a typical Bayesian SFS be consistent.

In order to gain some understanding of the unexpectedly good prequential behaviour in the above example, we can give the following general analysis. For a model $\mathscr{P} = \{P_\theta\}$, define an equivalence relation $\approx$ on values of θ by: $\theta_1 \approx \theta_2$ if $P_{\theta_1} \sim P_{\theta_2}$. The full parameter θ is consistently estimable if and only if $P_{\theta_1} \perp P_{\theta_2}$ (mutually singular) for $\theta_1 \neq \theta_2$, and the equivalence classes are then one-point sets. More generally, it will typically be the case that $P_{\theta_1} \perp P_{\theta_2}$ whenever $\theta_1 \not\approx \theta_2$. This we now assume. We can label the equivalence classes with a parameter α, and the distributions within an equivalence class by a further parameter β. Then $P_{\alpha,\beta_1} \sim P_{\alpha,\beta_2}$, all β_1, β_2, but $P_{\alpha_1,\beta_1} \perp P_{\alpha_2,\beta_2}$ if $\alpha_1 \neq \alpha_2$. This holds in Example 3, with $\alpha = (1-\rho)\phi$, $\beta = (\mu, \rho\phi)$.

An infinite sequence $\mathbf{x}$ of data-values will not be able to discriminate between distributions in a given equivalence class, but can do so between different, mutually singular, equivalence classes. In other words, α is a maximal consistently estimable parameter. We can plug in a consistent estimator $\tilde{\alpha}_n$ for α, but have no obvious way of substituting for β. However, just because $P_{\alpha,\beta_1} \sim P_{\alpha,\beta_2}$ for all β_1, β_2, the forecast distribution $P_{n,\tilde{\alpha}_n,\beta}$ will not depend, asymptotically, on β, so that it will not matter that we cannot estimate β. That is, asymptotically, the forecast distributions will only depend on a parameter α that *can* be consistently estimated. Any reasonable SFS (plug-in, Bayes, etc.) will then be consistent, this thus holding in very much greater generality than required for classical (parametric) consistency. The preceding argument can even be extended to cases where different model distributions are neither mutually absolutely continuous nor mutually singular. In general, Jeffreys's Law comes to our rescue: things we shall never find much out about cannot be very important for prediction.

9. EQUIVALENT STATISTICAL FORECASTING SYSTEMS AND EFFICIENCY

Let Q and R be SFS's for $\mathscr{P}$, and

$$\Lambda_n = \prod_{i=1}^{n} (q_i/r_i)$$

their prequential likelihood ratio for data $\mathbf{X}^{(n)}$. Extending slightly the argument of Section 7, we can regard Q as *no better than* R for forecasting $\mathscr{P}$ if Λ_n converges to $\Lambda < \infty$ with P_θ-probability one for all θ. We call Q and R *equivalent* for $\mathscr{P}$ if each is no better than the other. In this case the forecasts for the whole infinite future, given $\mathbf{X}^{(n)}$, made by Q and R will be asymptotically indistinguishable, with P_θ-probability one for all θ.

Example 2 (*continued*). Let Q and R be the given plug-in and fiducial SFS's respectively. Note that both densities q_1 and r_1 are undefined. We therefore condition on $X_1 = x_1$, so yielding the slightly modified likelihood-ratio comparison

$$\Lambda_n^* = \prod_{i=2}^{n} (q_i/r_i).$$

We have

$$2 \log q_i = -\log(2\pi) - (X_i - \bar{X}_{i-1})^2,$$

$$2 \log r_i = -\log(2\pi) - \{(i-1)/i\}(X_i - \bar{X}_{i-1})^2 - \log i + \log(i-1).$$

So

$$U_n = 2 \log \Lambda_n^* = -\sum_{i=2}^{n} (X_i - \bar{X}_{i-1})^2/i + \log n.$$

Now, under any P_θ (as under R),

$$(X_i - \bar{X}_{i-1})^2/i \sim \chi_1^2/(i-1),$$

independently for each i. Thus

$$E(U_n) = \log n - \sum_{i=1}^{n-1} i^{-1} \to -\gamma$$

where γ is Euler's constant 0.577 . . .,

$$\text{var}\,(U_n) = 2 \sum_{i=1}^{n} i^{-2} \to \pi^2/3.$$

In fact, $U_n \to U$ a.s. $[P_\theta]$, where U has characteristic function $\{\Gamma(1-2it)\}^{\frac{1}{2}}$. Consequently $0 < \Lambda^* = \lim(\Lambda_n^*) < \infty$ a.s. $[P_\theta]$, so that Q and R are equivalent SFS's for $\mathscr{P}$.

Aitchison (1975) has argued strongly that fiducial or Bayesian predictive distributions, as provided by R, are preferable to plug-in type distributions such as Q. In our prequential setting, where the plug-in values are adjusted to take account of each new observation, this superiority turns out to be asymptotically unimportant. Aitchison's argument is nevertheless valid to the extent that large data-sets may be needed to approach the asymptotic limit.

Now let $\mathscr{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$, and let R be a Bayesian SFS based on a prior density $\pi(\boldsymbol{\theta}) > 0$. Let Q be an arbitrary SFS for $\mathscr{P}$. Then

$$1 = R(\Lambda_n \to \Lambda < \infty) = \int P_{\boldsymbol{\theta}}\,(\Lambda_n \to \Lambda < \infty)\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta},$$

whence $P_{\boldsymbol{\theta}}(\Lambda_n \to \Lambda < \infty) = 1$ for almost all $\boldsymbol{\theta}$; i.e. R cannot be improved upon, except perhaps for a null set of parameter-values.

Any SFS R sharing this property may be called *efficient*. Essentially, this holds if and only if R is mutually absolutely continuous with respect to a positive-density Bayesian rule. A sequence of estimators may be called *prequentially efficient* if it yields an efficient plug-in rule. (The possibility of "super-efficiency" cannot be ruled out, but must be confined to a Lebesgue-null set.)

This new definition of estimator efficiency makes almost no assumptions about the family $\mathscr{P}$. In the traditional context of i.i.d. (independent identically distributed) variables (X_i), with suitably differentiable density $p(x \mid \boldsymbol{\theta})$, it may be demonstrated, by an argument somewhat parallel to that of Example 2 (continued), to be equivalent to Fisher's definition. Only recently have attempts been made to generalize the notion of Fisher efficiency to apply to stochastic process models: problems arise here when the distribution of the observed information does not become degenerate (Feigin, 1978). It is conjectured that prequential efficiency in this case is essentially the same as efficiency in the sense of Heyde (1975), a property possessed by maximum likelihood estimation. Indeed, under smoothness conditions, the m.l.e. may well prove to be prequentially efficient in still greater generality, as in cases such as Example 3. But the prequential definition of efficiency applies just as well to "non-regular" problems, for example, where the support of the distribution varies with the parameter value; and it further provides a unifying mechanism with which to probe still more distant departures from the regular i.i.d. case.

10. THE LIKELIHOOD OF A MODEL

Let P^* be an efficient SFS for $\mathscr{P} = \{P_\theta\}$. Then P^* yields, asymptotically, the most effective way of analysing past data for the purpose of predicting the future. The success of $\mathscr{P}$ in explaining a data sequence $\mathbf{x}^{(n)}$ should thus be measured, at least for large n, in terms of the fit of P^* to the data. For purposes of comparing $\mathscr{P}$ with an alternative model $\mathscr{Q} = \{Q_\phi\}$, we might wish to examine the likelihood ratio

$$\lambda_n^* = \prod_{i=1}^{n} p_i^*/q_i^*,$$

where Q^* is efficient for $\mathcal{Q}$ and $p_i^* = p_i^*(x_i \mid \mathbf{x}^{(i-1)})$, etc. We would prefer $\mathcal{P}$ to $\mathcal{Q}$ if λ_n^* is large, with such preference becoming definitive if $\lambda_n^* \to \infty$ as $n \to \infty$.

We thus define the (*prequential*) *likelihood* of the model $\mathcal{P}$ on data $\mathbf{x}^{(n)}$ as

$$L_n(\mathcal{P}) = \prod_{i=1}^{n} p_i^*,$$

and prefer models with large likelihoods, at least for large n. Two models whose likelihood ratio stays bounded (and away from zero) will be, asymptotically, equally acceptable. Note that it does not matter, for this criterion, which efficient SFS is employed to represent a model $\mathcal{P}$: a change from one to another will just multiply the limiting likelihood by a non-zero constant.

If one takes for P^* the m.l. plug-in rule, which, it is conjectured, will be prequentially efficient in very great generality, one obtains the likelihood

$$L_n(\mathcal{P}) = \prod_{i=1}^{n} p_i(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_{i-1}).$$

(Initial terms for which $\hat{\theta}_{i-1}$ is perhaps undefined may be ignored.) The typical term of this measures how well $\mathcal{P}$ did in forecasting X_i when the parameter was replaced by its estimate based on previous data $\mathbf{x}^{(i-1)}$ only. This avoids the over-optimism which would be typical of a pseudo-likelihood such as

$$\hat{L}_n(\mathcal{P}) = \sup_{\theta} p(\mathbf{x}^{(n)} \mid \theta) = \prod_{i=1}^{n} p_i(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_n),$$

wherein x_i itself contributes to the estimation of the value of θ used in "forecasting" it. In this property, prequential likelihood is similar to cross-validatory likelihood (Stone, 1977), but with the advantage of being more amenable to theoretical study. Over-fitting is properly penalized by $L_n(\mathcal{P})$: typically, if $\mathcal{P} \subset \mathcal{Q}$, and the data arise from a distribution in $\mathcal{P}$, then $L_n(\mathcal{P})/L_n(\mathcal{Q}) \to \infty$ as $n \to \infty$ (Q^* being consistent, but not efficient, for $\mathcal{P}$.)

Use of $L_n(\mathcal{P})$ is particularly appropriate if the purpose of modelling is forecasting, since it judges $\mathcal{P}$ precisely on how well it has performed in doing its best at forecasting the future. While there can never be any assurance that past forecasting performance will prove a reliable guide to future performance, what other reasonable guide can we follow? Even when forecasting is not the major aim, $L_n(\mathcal{P})$ forms a naturally appealing measure of the quality of $\mathcal{P}$.

We note, as is to be expected, that $L_n(\mathcal{P}) \leqslant \hat{L}_n(\mathcal{P})$. For let $r_n = L_n(\mathcal{P})/\hat{L}_n(\mathcal{P})$. Then

$$r_{n+1}/r_n = \frac{\prod_{i=1}^{n+1} p(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_{i-1}) \quad \prod_{i=1}^{n} p(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_n)}{\prod_{i=1}^{n+1} p(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_{n+1}) \quad \prod_{i=1}^{n} p(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_{i-1})}$$

$$= \prod_{i=1}^{n+1} p(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_n) \Big/ \prod_{i=1}^{n+1} p(x_i \mid \mathbf{x}^{(i-1)};\ \hat{\theta}_{n+1})$$

$$= p(\mathbf{x}^{(n+1)} \mid \hat{\theta}_n)/p(\mathbf{x}^{(n+1)} \mid \hat{\theta}_{n+1})$$

$$\leqslant 1.$$

Also, $L_1(\mathscr{P}) = p(x_1 \mid \hat{\theta}_0) \leqslant p(x_1 \mid \hat{\theta}_1) = \hat{L}_1(\mathscr{P})$, however we define $\hat{\theta}_0$. The result follows by induction.

To investigate $L_n(\mathscr{P})/\hat{L}_n(\mathscr{P})$ further, suppose that $\mathscr{P} = \{P_{\boldsymbol{\theta}}\}$, where $\boldsymbol{\theta}$ ranges over $\mathbb{R}^d$ (or some suitable subset thereof), and let $L_n(\boldsymbol{\theta}) = \log p(\mathbf{x}^{(n)} \mid \boldsymbol{\theta})$. We suppose L_n to be suitably differentiable, and suppose further that $\hat{\boldsymbol{\theta}}_n$ converges to a limit $\hat{\boldsymbol{\theta}}_0$, and that the matrix of second derivatives $L_n''(\hat{\boldsymbol{\theta}}_n)$ is asymptotically of the form $-n\mathbf{J}$ with $\mathbf{J}$ finite and non-singular. These properties will hold, with probability 1, under the usual assumptions of independent identically distributed variables, but also more generally. For example, we do not require $\mathbf{J}$ to be non-random, thus allowing the sort of model termed "non-ergodic" by Basawa and Scott (1983).

Instead of a plug-in rule, we use the (efficient) Bayesian PFS P^* based on a continuous positive prior density $\pi(\boldsymbol{\theta})$. Then

$$L_n(\mathscr{P}) = p^*(\mathbf{x}^{(n)}) = \int p(\mathbf{x}^{(n)} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

(Incidentally, this form of $L_n(\mathscr{P})$ does not depend on the order in which the observations are presented. Consequently, the effect of this order on any efficient plug-in rule can be of no asymptotic importance.)

We have

$$L_n(\mathscr{P})/\hat{L}_n(\mathscr{P}) = \int \exp\{L_n(\boldsymbol{\theta}) - \hat{L}_n(\hat{\boldsymbol{\theta}}_n)\}\, \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$\sim \int \exp\{-\tfrac{1}{2} n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\mathbf{J}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\}\, \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$\sim \pi(\hat{\boldsymbol{\theta}}_n)\, (2\pi)^{\frac{1}{2}d} \{\det(n\mathbf{J})\}^{-\frac{1}{2}}.$$

Thus $\log L_n(\mathscr{P}) = \log \hat{L}_n(\mathscr{P}) - \frac{1}{2} d \log n + k$, where k is bounded. Consequently, for large n at least, the prequential log-likelihood of a model may be approximated by penalizing the maximized value, $\log p(\mathbf{x}^{(n)} \mid \hat{\boldsymbol{\theta}}_n)$, by a term $\frac{1}{2} d \log n$. This is basically the Bayesian argument and conclusion given by Jeffreys (1936); see also Schwarz (1976) and Akaike (1978). However, the prequential argument justifies using this, even if one does not take a Bayesian approach, as an approximation to any reasonable likelihood assessment of predictive performance. (Note that the AIC is also based on an assessment of predictive performance, but for an entirely hypothetical new, independent problem with the same structure as that providing the data (Akaike, 1969, 1970, 1974). Our predictive assessment is confined to the data actually being collected, and deliberately eschews purely conceptual repetitions of the whole process.)

11. FURTHER APPLICATIONS

The scope of the prequential approach to Statistics is not limited to the theoretical aspects surveyed here, but also takes in a wide variety of practical problems of data analysis. It can, for example, be used as a very general alternative to procedures such as AIC or cross-validation for assessing strategies of analysis, and selecting between rival models. Applications that have been investigated by students at University College London include choice of symptom variables for medical diagnosis (Seillier, 1982), order selection for Markov models of daily rainfall occurrence (Edwards, 1983; Jain, 1983; Dawid, 1984) and the vetting of items proposed for inclusion in a

battery of educational tests (Opie, 1983). Other promising applications, out of many further possibilities, might be the choice of kernel width in density estimation (Hall, 1982), or of the number of groups in cluster analysis. There is a good deal of common ground with, and much to be gleaned from, the whole field of Stochastic Control Theory; see, for example, Maybeck (1979-82). There is particular scope for related computational methods of recursive estimation, such as stochastic approximation and the Kalman-Bucy filter.

Further insight from a prequential approach is foreseen into a host of theoretical topics which are the subject of current research activity, such as asymptotic and conditional inference, particularly in relation to general stochastic processes. The possible developments and applications are manifold, and my sincere hope is that, by making this personal forecast of a prequential future for Statistical Theory, I may stimulate others to apply themselves to these problems, and so prove the forecast empirically valid.

ACKNOWLEDGEMENTS

I am grateful to Martin Crowder and an anonymous referee for helpful comments on an earlier draft of this paper.

REFERENCES

Aitchison J. (1975) Goodness of prediction fit. *Biometrika*, **62**, 547–554.

Akaike, H. (1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, **21**, 243–247.

——— (1970) Statistical predictor identification. *Ann. Inst. Statist. Math.*, **22**, 203–217.

——— (1974) A new look at the statistical model identification. *IEEE Trans. Auto. Control*, **AC-19**, 716–723.

——— (1978) A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, **30**, 9–14.

Basawa, I. V. and Prakasa Rao, B. L. S. (1980) *Statistical Inference for Stochastic Processes.* New York: Academic Press.

Basawa, I. V. and Scott, D. J. S. (1983) *Asymptotic Optimal Inference for Non-Ergodic Models.* Lecture Notes in Statistics Vol. 17, New York: Springer-Verlag.

Billingsley, P. (1961) *Statistical Inference for Markov Processes*. University of Chicago Press.

Blackwell, D. and Dubins, L. E. (1962) Merging of opinions with increasing information. *Ann. Math. Statist.*, **33**, 882–886.

Brown, R. L., Durbin, J. and Evans, J. M. (1975) Techniques for testing the constancy of regression relationships over time (with Discussion). *J. R. Statist. Soc.* B, **37**, 149–192.

Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.

Dawid, A. P. (1982a) The well-calibrated Bayesian (with Discussion). *J. Amer. Statist. Ass.*, 77, 605–613.

——— (1982b) Objective probability forecasts. Research Report 14, Department of Statistical Science, University College London.

——— (1983a) Inference, statistical: I. In *Encyclopedia of Statistical Sciences*, Vol. 4 (S. Kotz, N. L. Johnson and C. B. Read, eds), pp. 89–105. New York: Wiley.

——— (1983b) Probability forecasting. Research Report 30, Department of Statistical Science, University College London. To appear in *Encyclopedia of Statistical Sciences*. New York: Wiley.

——— (1984) In discussion of Stern and Coe (1984). *J. R. Statist. Soc.* A, **147**, 22.

de Finetti, B. (1975) *Theory of Probability*. (English translation.) Two volumes. New York: Wiley.

Dvoretsky, A. (1972) Asymptotic normality for sums of dependent random variables. *Proc. 6th Berkeley Symp.* Vol. II, 513–535.

Edwards, A. S. (1983) Probabilistic weather forecasting. BSc Dissertation, Department of Statistical Science, University College London.

Feigin, P. D. (1978) The efficiency criteria problem for stochastic processes. *Stoch. Proc. Appl.*, **6**, 115–127.

Hall, P. (1982) Cross-validation in density estimation. *Biometrika*, **69**, 383–390.

Heyde, C. C. (1975) Remarks on efficiency in estimation for branching processes. *Biometrika*, **62**, 49–55.

Hill, T. P. (1982) Conditional generalizations of strong laws which conclude the partial sums converge almost surely. *Ann. Prob.*, **10**, 828–830.

Howard, J. V. (1975) Computable explanations. *Zeitschr. f. Math. Logik und Grundlagen d. Math.*, **21**, 215–224.

Jain, R. (1983) Probabilistic weather forecasting. MSc Dissertation, Department of Statistical Science, University College London.

Jeffreys, H. (1936) Further significance tests. *Proc. Camb. Phil. Soc.*, **32**, 416–445.

——— (1938) Science, logic and philosophy. *Nature*, **141**, 716–719.

Keiller, P. A. and Littlewood, B. (1984) Adaptive software reliability modelling. Research Report, Centre for Software Reliability, The City University.

Maybeck, P. S. (1979–82) *Stochastic Models, Estimation and Control.* Three volumes. New York: Academic Press.
Opie, G. A. (1983) Educational scaling. BSc Dissertation, Department of Statistical Science, University College London.
O'Reilly, F. J. and Quesenberry, C. P. (1973) The conditional probability integral transformation and applications to obtain composite chi-square goodness-of-fit tests. *Ann. Statist.*, **1**, 74–83.
Rosenblatt, M. (1952) Remarks on a multivariate transformation. *Ann. Math. Statist.*, **23**, 470–472.
Schwarz, G. (1976) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Seillier, F. M.-L. G. (1982) Selection aspects in medical diagnosis. MSc Dissertation, Department of Statistical Science, University College London.
Shiryayev, A. N. (1981) Martingales: recent developments, results and applications. *Int. Statist. Rev.*, **49**, 199–233.
Stern, R. D. and Coe, R. (1984) A model fitting analysis of daily rainfall data (with Discussion). *J. R. Statist. Soc.* A, **147**, 1–34.
Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B, **39**, 44–47.

INVITED DISCUSSION

O. E. Barndorff-Nielsen (Aarhus University): Prediction is a subject that has not attracted the attention from theoretical statisticians that it deserves, and Professor Dawid's interesting paper is a welcome contribution towards rectifying this state of affairs.

The conceptual difficulties in establishing a satisfactory theory of prediction begin already with the implicit possible situation, that in which it is desired to predict the outcome of a single experiment governed by a fully known distributional law. Suppose the outcome is a scalar variate x with a probability (density) function $p(x)$. What value of x should be considered the most "likely"? Or, if we want to delineate a 99 per cent prediction region how should this be done? Standard proposals for a predictate are the mean or the mode of $p(x)$. But, outside a decision-theoretic framework, what rational reasons might be given for any particular such choice?

Professor Dawid's definitions of a "prequential principle" and of "Jeffreys's law", as well as those of prequential efficiency and prequential likelihood of a model, must constitute essential elements for a theory of prequential statistics and these conceptual clarifications seem of very considerable value. I was also much interested to read the author's discussion of "Jeffreys's law" in its relation to models with parameters that cannot all be estimated consistently and to non-ergodic models.

Ideas related to those of the prequential approach have been discussed by P. Martin-Löf and S. L. Lauritzen (see Lauritzen, 1974, 1984, and references there), and it would be useful to have this connection commented on.

An obvious reason why the literature on prediction is rather limited is that in statistical practice problems of forecasting are not very commonly met. For this same reason, I have reservations as to the usefulness for statistical theory in general of the approach proposed here by Professor Dawid. On the other hand, the development of a stringently thought out theory of prediction seems bound to reflect fruitfully on other parts of statistics.

Since we are here not only to discuss prediction theory but also to make some actual predictions on the future developments, as well as the present state of statistical theory, it will be in order, I trust, if I outline here some areas of statistical theory that seem to me of particular importance and in which we are witnessing or are likely in the near future to see some substantial advances.

Parametric statistical models and likelihood constitute the backbone of statistics, and will undoubtedly continue to do so.

Due both to the application of advanced mathematical methods, such as those of differential and integral geometry or of graph theory and combinatorics, and to the use of computers, we are able to handle more and more structured models, and to take better advantage of their structures in the inference. Important classes of such models are full and curved exponential models, transformation models, the *CG* models developed by Lauritzen and Wermuth (1983, 1984), the intricate normal multifactor models presently being studied by R. A. Bailey and T. P. Speed and others, the Cox-type models for survival analysis, the developments of Nelder and Wedderburn's generalized linear models, models for multivariate time series and models for spatial variation.

Hand in hand with these advances go further developments of the ideas and techniques of separate inference. A breakthrough in this area came with David Cox's idea of partial likelihood

(Cox, 1975a) and we have not, I believe, seen the full potentials of this yet. There are connections to S-ancillarity and cuts and to the exogeneity concepts of econometrics, cf. Engle *et al.* (1983), Johansen (1983). Some very recent and interesting results on sufficiency have been obtained by Remon (1984) and McCullagh (1984). Remon defines a statistic t to be sufficient with respect to a parameter of interest ψ if the distribution of t depends on ψ only and if the profile likelihood for ψ depends on the data through t only. Surprisingly, this simple idea encompasses not only the ordinary sufficiency concept but also the seemingly disparate definitions of S-sufficiency and G-sufficiency due, respectively, to Fraser (1956) and Barnard (1963). The title "Local sufficiency" of McCullagh (1984) echoes the "Local ancillarity" of Cox (1980). McCullagh bases his arguments on a statistic t of fixed dimension p, such as a vector of lower order derivatives of the log likelihood function, and he shows by asymptotic analysis, as the underlying sample size n tends to infinity, that there exists a one-to-one transformation $t \to (s, a)$ such that s has the same dimension d as the parameter ω of the model while a, of dimension $p-d$, is approximately ancillary and independent of s locally at the governing value of ω, to a degree of approximation that makes conditioning on a unnecessary. As a consequence one is, in a specific sense, able to make "conditional inference without conditioning". The tools of the derivation are cumulants and Edgeworth expansions. A rather different approach (Barndorff-Nielsen, 1984b, c) uses mixed derivatives of the log-model function and expansions derived from the formula $c\,|\hat{j}|^{\frac{1}{2}}\,\bar{L}$ for the conditional distribution of the maximum likelihood estimator. This approach allows "conditional inference without conditioning and without integrations over the sample space". The "Problem of the Nile" is an element in much of this work. Although its solution is known for transformation models and to a large extent for exponential models (Amari, 1984; Barndorff-Nielsen, 1980, 1984a), much remains to be done.

It would seem of some interest to explore the relations between the works of McCullagh and Remon referred to above and the concept of modified profile likelihood (Barndorff-Nielsen, 1983).

For a particular use of the latter concept, in a study of distributional shape, see Barndorff-Nielsen *et al.* (1983).

The various results discussed above, as well as other recent investigations, strongly underline the fundamental character of R. A. Fisher's concept of likelihood.

In recent years there has been considerable activity, comprising much of the above, in the area of what might be called "intermediate asymptotics". This is concerned with sharpening the usual kind of "0th order" asymptotic results, based on the central limit theorem, so as to include the correction terms of order $O(n^{-\frac{1}{2}})$ and $O(n^{-1})$. This, among other things, allows a unified formulation of certain exact and asymptotic results, the formula $c\,|\hat{j}|^{\frac{1}{2}}\,\bar{L}$ being a case in point. The use of Bartlett adjustment factors has a natural place in this framework, cf. Barndorff-Nielsen and Cox (1984). There are also interesting and illuminating connections to differential geometry; see Amari (1984), Barndorff-Nielsen (1984c). We are likely to see this activity continue in substantial measure.

In addition to prediction, another fairly underdeveloped area that seems ripe for advances is that of parametric robustness studies, with the associated concept of robust pivotals, cf. Barnard (1981, 1983). Many of the above-mentioned results should be instrumental here and, again, likelihood must be of central importance; in this connection, see Cox (1975b), and Barndorff-Nielsen (1981).

Some further fields where we are likely to see exciting progress are: (i) sample surveys, (ii) stereology, (iii) inference for stochastic processes and (iv) the connections to statistical physics and chemistry.

In the meantime, the more foundational problems should not be left out of sight. For instance, the method of setting confidence regions, in spite of its great usefulness, rests on a somewhat unsatisfactory foundation.

All in all, however, we are entitled to feel, I think, that statistics has come of age as a mature and highly vigorous and substantial science, due indeed in large measure to the efforts of the Royal Statistical Society.

ADDITIONAL REFERENCES

Amari, S. (1984) *Differential Geometry in Statistics*. Lecture Notes in Statistics. Springer, Heidelberg. (To appear. Title preliminary.)

Barnard, G. A. (1963) Some logical aspects of the fiducial argument. *J. R. Statist. Soc.* B, **25**, 111–114.
——— (1981) The conditional approach to robustness. In *Statistics and Related Topics* (A. K. Md. E. Saleh, M. Csörgö, D. A. Dawson and J. N. K. Rao, eds). Amsterdam: North-Holland.
——— (1983) Pivotal inference and the conditional view of robustness (Why have we for so long managed with normality assumptions?). In *Scientific Inference, Data Analysis, and Robustness*. (G. E. P. Box, T. Leonard and C.-F. Wu, eds.) pp. 1–8. New York: Academic Press.
Barndorff-Nielsen, O.E. (1980) Conditionality resolutions. *Biometrika*, **67**, 293–310.
——— (1981) Likelihood prediction. *Symposia Mathematica*, **XXV**, 11–24.
——— (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.
——— (1984a) On conditionality resolution and the likelihood ratio ancillary for curved exponential models. *Scand. J. Statist.*, **11** (to appear).
——— (1984b) Confidence limits from $c\,|\hat{j}|\sqrt{L}$. Research Report 104, Department of Theoretical Statistics, Aarhus University.
——— (1984c) Differential and integral geometry in statistical inference. Research Report 106, Department of Theoretical Statistics, Aarhus University.
Barndorff-Nielsen, O. E., Blaesild, P., Jensen, J. L. and Sørensen, M. (1983) The fascination of sand. With three appendices by R. A. Bagnold. Research Report 93, Department of Theoretical Statistics, Aarhus University. (To appear in *A Celebration of Statistics*, Centenary Volume of the ISI.)
Barndorff-Nielsen, O. E. and Cox, D. R. (1984) Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. R. Statist. Soc.* B, **46**, (to appear).
Cox, D. R. (1975a) Partial likelihood. *Biometrika*, **62**, 269–276.
——— (1975b) Prediction intervals and Bayes confidence intervals. In *Perspectives in Probability and Statistics* (J. Gani, ed.) London: Academic Press.
——— (1980) Local ancillarity. *Biometrika*, **67**, 273–278.
Engle, R. F., Hendry, D. F. and Richard, J.-F. (1983) Exogeneity. *Econometrica*, **51**, 277–304.
Fraser, D. A. S. (1956) Sufficient statistics with nuisance parameters. *Ann. Math. Statist.*, **27**, 838–842.
Johansen, S. (1983) An extension of Cox's regression model. *Internat. Statist. Rev.*, **51**, 165–174.
Lauritzen, S. L. (1974) Sufficiency, prediction and extreme models. *Scand. J. Statist.* **1**, 128–134.
——— (1984) Extreme point models in statistics. *Scand. J. Statist.*, **II**. (To appear.)
Lauritzen, S. L. and Wermuth, N. (1983) Graphical and recursive models for contingency tables. *Biometrika*, **70**, 537–552.
——— (1984) Mixed interaction models. Unpublished manuscript.
McCullagh, P. (1984) Local sufficiency. *Biometrika*, **71**. (To appear.)
Remon, M. (1984) On a concept of partial sufficiency: *L*-sufficiency. *Internat. Statist. Rev.* **52**. (To appear.)