Revista Brasileira de Probabilidade e Estatística (1993), **7**, pp. 65–82.

PREDICTIVE LIKELIHOOD IN FINITE POPULATIONS

Pilar Iglesias, Mônica C. Sandoval and Carlos Alberto de Bragança Pereira

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo, SP, Brasil

Summary

Superpopulation models are transformed in predictive models in order to permit the use of standard classical statistics techniques. Confidence intervals based on predictive models replace the predictive intervals based on superpopulation models. The ideas are illustrated by various examples and the normal case turns out to produce intervals that are also obtained by the standard classical survey sampling techniques.

Key words: Maximum likelihood predictor, nuisance parameter, pivotal quantity, prediction, predictive interval, predictive model, minimal sufficient reduction, specific sufficiency.

1. Introduction

Prediction of unknown quantities in parametric statistics has been focused from different points of view and predictive intervals (for these quanti-

ties) of various types have been developed by many authors (see, for example, Thatcher (1964), Hahn (1969), Lawless (1972), and Vit, (1973)). Generally, methods for deriving such intervals either make use of pivotal quantities and normal approximations or are based on hypothesis testing approaches. Faulkenberry (1973) proposed an alternative approach that was based on the study of the conditional distribution of the original random quantities given a sufficient statistics. In connection with this and under a non-Bayesian perspective, we advocate the use of a likelihood approach in order to obtain maximum likelihood predictors for the quantities of interest. The idea is to derive conditional probability functions producing the predictive likelihoods that do not depend on the value of the unknown parameter (a quantity of no interest) which is in fact replaced by the sufficient statistics (the quantity of interest). For more details on the various forms of predictive likelihood we refer to the recent work of Bjørnstad(1990). Other important references about choice of likelihoods are Bayarri, De Groot, and Kadane (1986) and de Finetti (1977).

The related definitions of predictive likelihoods introduced by Lauritzen (1974), Hinkley (1979), and Butler (1985) that consist of conditioning on minimal sufficient statistics, may differ in some situations. Consequently, the maximum likelihood predictors obtained under these situations may differ significantly (Bjørnstad, 1990). The aim of the present paper is to consider a definition that combine the former ones and have interesting justifications under the survey sampling or finite population context. The main idea consists of looking for a family of probability for the data to be observed, indexed by the quantity of interest, to be predicted. This quantity, that in the context of finite populations is a function of both the observed and the unobserved quantities, must act as the parameter in a standard statistical model.

Let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be a random vector with distribution indexed by a parameter θ (scalar or vector). The quantity to be observed, the sample, is represented without loss of generality by

$$\mathbf{Y}_s = (Y_1, \ldots, Y_n)$$

where $n < N$. The remaining part of $\mathbf{Y}$, the unobserved quantity, is represented by $\mathbf{Y}_U = (Y_{n+1}, \ldots, Y_N)$.

The problem to be solved consists of making a predictive statement about a function of $\mathbf{Y}$, $\boldsymbol{\tau}(\mathbf{Y})$, based on the observed value, $\mathbf{y}_s$ of $\mathbf{Y}_s$.

From a Bayesian point of view, the problem is solved in a straightforward manner. Once the prior probability (density) function, pdf, for the model parameter, θ, is considered, the predictive distribution of $\boldsymbol{\tau}(\mathbf{Y})$ given $(\mathbf{Y}_s = \mathbf{y}_s)$ is obtained. For instance, if $p(\theta)$ is the prior pdf, then the predictive pdf at point $\boldsymbol{\tau}(\mathbf{Y}) = \boldsymbol{\tau}$ is

$$f(\boldsymbol{\tau}|\mathbf{y}_s) = \int f(\boldsymbol{\tau}|\mathbf{y}_s, \theta) p(\theta|\mathbf{y}_s)\, d\theta \ , \tag{1.1}$$

where $p(\theta|\mathbf{y}_s)$ is the posterior pdf of θ. That is, the predictive pdf is the average of $f(\boldsymbol{\tau}|\mathbf{y}_s, \theta)$ under the posterior distribution. Also note that alternatively we could write

$$f(\boldsymbol{\tau}|\mathbf{y}_s) \propto f(\boldsymbol{\tau}) f(\mathbf{y}_s|\boldsymbol{\tau}) \ , \tag{1.2}$$

which can be interpreted as the Bayes' operation when $\boldsymbol{\tau}$ is considered as the parameter and $f(\mathbf{y}_s|\boldsymbol{\tau})$ defines the likelihood of $\boldsymbol{\tau}$. Here also, $f(\boldsymbol{\tau})$ and $f(\mathbf{y}_s|\boldsymbol{\tau})$ are obtained by integration using adequate distributions of θ. Note that, if τ is sufficient under the full model of $\mathbf{Y}$, then integration to obtain $f(\mathbf{y}_s|\boldsymbol{\tau})$ is unnecessary.

If prior information is not supposed to be used, the distributions that could be used for prediction, $f(\boldsymbol{\tau}|\theta)$, $f(\boldsymbol{\tau}|\mathbf{y}_s, \theta)$, and $f(\mathbf{y}_s|\boldsymbol{\tau}, \theta)$, do involve θ. Consequently, a satisfactory frequentist solution for the prediction of $\boldsymbol{\tau}$ may not be obvious. A tentative could be to replace, in these functions, the parameter θ by its maximum likelihood estimate. Such an approach implicitly assumes that the true value of θ is its estimate and would not take in account the uncertainty about θ.

Hinkley (1979) and Butler (1986) introduced predictive likelihood functions that neither involves the replacement of an estimate for θ nor requires the use of prior distributions. Consequently, standard inferential methods could

be used. The different definitions presented by Hinkley (1979) and Butler (1986) may lead to different solutions for the prediction problem. The definition of predictive likelihood that will be used to solve the problem stated above, the prediction of $\boldsymbol{\tau}$, is more general than the former ones and produces no inconsistency for the particular case of finite populations.

Several aspects of a prediction problem are discussed in Section 2. Section 3 presents solutions for finite population problems. It is interesting to notice that the standard normal produce the same results obtained under the standard survey sampling techniques.

2. Prediction

Based on the model described in Section 1, the problem to be solved is the prediction of $\boldsymbol{\tau}(\mathbf{Y})$ using the observation $\mathbf{y}_s$ of $\mathbf{Y}_s$. The starting point in a frequentist context consists in associating a distribution to $\mathbf{Y}_s$ indexed by $\boldsymbol{\tau}$. The natural procedure is to consider the original random variables $Y_1, \ldots, Y_N$, as independent with a common distribution indexed by a unknown parameter θ. With this model we obtain the conditional distribution of the sample $\mathbf{Y}_s$ given $\boldsymbol{\tau}$ and θ. Hence the new model is indexed by two parameter, $\boldsymbol{\tau}$ and θ. In this manner and according to the main purpose of this study, the problem may involve inference about $\boldsymbol{\tau}$ in the presence of a nuisance parameter, θ (Basu (1977)).

The parameter of the modified model is denoted by $\pi = (\boldsymbol{\tau}, \theta)$ and the modified likelihood by $l(\pi|\mathbf{y}_s)$. The definitions and results presented in the sequel will be the basis of the solution. Note that there is a lack of independence between the sample elements, $Y_1, \ldots, Y_n$, in the modified model.

We are using, in a general notation, f (or g) and l for probability density and likelihood functions, respectively. For simplicity, they do not show neither differences in dimension nor in distribution since they are implicit in each case.

Definition 2.1. The function $l(\pi|\mathbf{y}_s)$ is called predictive likelihood. We also call predictive functions the ones that are proportional to it, following the principle of sufficiency.

Definition 2.2. If there exists a point $\widehat{\pi} = (\widehat{\boldsymbol{\tau}}, \widehat{\theta})$ such that

$$\sup_{\pi} l(\pi|\mathbf{y}_s) = l(\widehat{\pi}|\mathbf{y}_s) \ ,$$

then $\widehat{\boldsymbol{\tau}}$ is the maximum likelihood predictor of $\boldsymbol{\tau}(\mathbf{Y})$.

Definition 2.3. If there exist functions $\mathbf{T}_1 = \mathbf{T}_1(\mathbf{Y}_s)$ and $\mathbf{T}_2 = \mathbf{T}_2(\mathbf{Y}_s)$, with observed values $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$, such that $\mathbf{T}_1 < \mathbf{T}_2$ a.s. and $\Pr\left[(\mathbf{T}_1, \mathbf{T}_2) \ni \boldsymbol{\tau}(\mathbf{Y})|\theta\right] = \gamma$, then $(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$ is called the predictive interval of $\boldsymbol{\tau}(\mathbf{Y})$ with $100\gamma\%$ of confidence.

The following result characterizes the role $\boldsymbol{\tau}(\mathbf{Y})$ when it is a minimal sufficient statistics under the original model, $f(\mathbf{y}|\theta)$.

Lemma 2.1. *If there exist functions $T = T(\mathbf{Y}_s)$ and $U = U(\mathbf{Y}_s)$ such that, (with respect to θ) S, U, and $\boldsymbol{\tau} = \boldsymbol{\tau}(\mathbf{Y})$ are minimal sufficient reductions of $\mathbf{Y}_s$, $\mathbf{Y}_U$, and $\mathbf{Y}$, respectively, then*

i) $f(\mathbf{y}_s|\pi)$ depends on π only through $\boldsymbol{\tau}$ and

ii) $f(\mathbf{y}_s|\pi) = g(t|\boldsymbol{\tau})h(\mathbf{y}_s)$, where g is the pdf of T.

The proof is straightforward.

It should be noted that, if in this Lemma, $\mathbf{Y}_s$ and $\mathbf{Y}_U$ are statistically independent in the original model and U is uniquely defined by T and $\boldsymbol{\tau}$, then g defines the predictive likelihood of Hinkley (1979). That is, Hinkley's predictive likelihood is $l^*(u|t)$ which is equal to $g(t|\boldsymbol{\tau})$ at the observed point t. Moreover, under the conditions of the Lemma, the maximum likelihood

predictor of $\boldsymbol{\tau}$ agrees with the predictor defined in Lauritzen (1974). Also, in this case there is no nuisance parameter to be eliminated.

Another feature of the modified likelihood when $f(\mathbf{y}_s|\pi) = f(\mathbf{y}_s|\boldsymbol{\tau})$ is that, under the Bayesian point of view, formula (1.2) may be immediately applied without the need of integration to obtain its factors. In this case, comparison between Bayesian and frequentist methods follows the standard procedures since there is no need for elimination of nuisance parameters.

To illustrate Lemma 2.1, we present the following simple example.

Example 2.1. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be independent exponential random quantities with common unknown mean equal to θ^{-1}. The population total $\boldsymbol{\tau}(\mathbf{Y}) = Y_1 + \cdots + Y_N$ and the sample total $T = Y_1 + \cdots + Y_n$ are minimal sufficient reductions of $\mathbf{Y}$ and $\mathbf{Y}_s$, with respect to the original model. The predictive likelihood may be written as

$$l(\pi|\mathbf{y}_s) \propto l(\boldsymbol{\tau}|t) \propto \frac{(\boldsymbol{\tau} - t)^{N-n-1} t^{n-1}}{\boldsymbol{\tau}^{N-1}} \mathbf{I}(t, \boldsymbol{\tau}) \ , \tag{2.1}$$

where $\mathbf{I}(\cdot, \boldsymbol{\tau})$ is the indicator function of the interval $(0, \boldsymbol{\tau})$.

The maximum likelihood predictor obtained from (2.1) is $\widehat{\boldsymbol{\tau}} = \frac{T}{n}(N-1)$. To obtain the predictive interval, we notice that $Z = \frac{T}{\boldsymbol{\tau}}$ is a pivotal quantity with distribution Beta with parameters $N - n$ and n. Hence a predictive interval with $100\gamma\%$ of confidence is of the form

$$\left[\frac{T}{b} \ ; \ \frac{T}{a}\right] \ ,$$

where a and b are chosen in such a way that $a < b$ and $\Pr(a < Z < b) = \gamma$. (Note that $\Pr(a < Z < b)$ is the incomplete Beta function divided by the Beta function, both calculated at point $(N - n, n)$.] To obtain the shortest interval we choose a and b that makes $(1/a) - (1/b)$ minimum. In particular if $N = n + 1$, the shortest predictive interval is

$$\left[T, \frac{T}{\sqrt[n]{(1-\gamma)}}\right] \ .$$

Unfortunately, there are many situations where the quantity of interest, $\boldsymbol{\tau}$ is not a minimal sufficient reduction of $\mathbf{Y}$ for the original model. Consequently, elimination of the nuisance parameter, θ, would be necessary. However, interesting results can be obtained under other kinds of simplifications. In the sequel we discuss a common situation that allows simple solutions.

Suppose that the original parameter has the representation $\theta = (\theta_S, \theta_U)$ and that, for any fixed value of θ_U, $\boldsymbol{\tau}(\mathbf{Y})$ is specific sufficient relatively to θ_S (see Basu (1977)); that is, the nuisance parameter to be eliminated is simply θ_U. In this case, the predictive model depends on the modified parameter $\pi = (\boldsymbol{\tau}, \theta)$ only through $\pi^* = (\boldsymbol{\tau}, \theta_U)$. As before we represent the maximum likelihood predictor by $\widehat{\boldsymbol{\tau}}$.

A dual situation is when there exists a minimal sufficient reduction of $\mathbf{Y}$ (in relation to θ), $\boldsymbol{\eta}(\mathbf{Y})$, such that $\boldsymbol{\eta}(\mathbf{Y}) = (\boldsymbol{\tau}(\mathbf{Y}), \boldsymbol{\lambda}(\mathbf{Y}))$. The function $l(\boldsymbol{\eta}|\mathbf{y}_S) = f(\mathbf{y}_S|\boldsymbol{\eta})$ may be considered as the predictive likelihood where $\boldsymbol{\lambda}$ is the nuisance parameter to be eliminated and $\boldsymbol{\tau}$ is the quantity to be predicted. The maximum likelihood predictor in this case is represented by $\widetilde{\boldsymbol{\tau}}$.

The next example is very standard and incorporates both situations described above, besides the fact that there is a choice of $\boldsymbol{\lambda}$ such that $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ are statistically independent. Consequently, to obtain the predictor we can use indifferently either likelihood $l(\pi^*|\mathbf{y}_S)$ or $l(\boldsymbol{\eta}|\mathbf{y}_S)$ since $\widetilde{\boldsymbol{\tau}} = \widehat{\boldsymbol{\tau}}$.

Example 2.2. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be independent normal random variables with unknown common mean and variance represented by $\theta_S = \mu$ ($\in \mathbb{R}$) and $\theta_U = \sigma^2$ ($\in \mathbb{R}^+$), respectively. Here, we take

$$T = Y_1 + \cdots + Y_n\,,\ \ V = Y_1^2 + \cdots + Y_n^2\,,\ \ \boldsymbol{\tau} = Y_1 + \cdots + Y_N\,,\ \text{ and } \boldsymbol{\lambda} = Y_1^2 + \cdots Y_N^2\,.$$

Recall that $\boldsymbol{\tau}$ is specific sufficient in relation to μ. To obtain the first predictive likelihood, $l(\pi^*|\mathbf{y}_S)$, we note that $\mathbf{Y}_S|\pi^*$ is distributed as a n-variate normal with mean $\left(\frac{\boldsymbol{\tau}}{N}\mathbf{j}_n\right)$ and covariance matrix $\sigma^2\left(\mathbf{I}_n - \frac{1}{N}\mathbf{J}_n\right)$, where $\mathbf{j}_n$ is the n-variate vector with all components equal to the unity, $\mathbf{I}_n$ is the identity matrix

of order n and $\mathbf{J}_n$ is the squared matrix of order n with all components equal to the unity. Hence the maximum likelihood predictor of $\boldsymbol{\tau}$ is $\widehat{\boldsymbol{\tau}} = \frac{N}{n}T$. On the other hand, the alternative likelihood is based on the (conditional) distribution of $(T, V)|\boldsymbol{\eta}$, since (T, V) is a sufficient reduction (in the original model) of the sample, $\mathbf{Y}_S$, in relation to (μ, σ^2). After some calculations we obtain the likelihood at point (t, v)

$$l(\boldsymbol{\eta}|t, v) \propto \left(\boldsymbol{\lambda} - v - \frac{(\boldsymbol{\tau} - t)^2}{N - n}\right)^{(N-n-3)/2} \left(\boldsymbol{\lambda} - \frac{\boldsymbol{\tau}^2}{N}\right)^{-(N-3)/2} . \qquad (2.2)$$

The maximum likelihood predictor, $\widetilde{\boldsymbol{\tau}}$, coincides with $\widehat{\boldsymbol{\tau}} = \frac{N}{n}T$. If $\boldsymbol{\lambda}$ is replaced by

$$\boldsymbol{\lambda}^* = \sum_{i=1}^{N} (Y_i - (\boldsymbol{\tau}/N))^2 \ ,$$

then we would obtain the same maximum likelihood predictor for $\boldsymbol{\tau}$ since there is a one-to-one correspondence between $(\boldsymbol{\tau}, \boldsymbol{\lambda})$ and $(\boldsymbol{\tau}, \boldsymbol{\lambda}^*)$. Note that, in the original model, $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}^*$ are statistically independent.

We end this section by noticing that all the discussion presented here can be extended for the case where $Y_1, \ldots, Y_N$ are vectors.

3. Finite population examples

Prediction in finite population has been studied till now through the standard classical sampling theory and the superpopulation model approach. Advantages and restrictions for these two methods have been presented in the literature (see Basu (1969) and Cassel, Särndal, & Wretman (1967) for interesting discussions and for a large list of references). The restrictions for the standard sampling theory leans upon the probabilistic model used that is not related to the quantity of interest. For the superpopulation model, the restrictions can be stated from the fact that ad hoc methods must be constructed. The problems come from the fact that there is an unknown

parameter of no interest that must be estimated in order to use it for the prediction of the unknown quantity of interest. The approach discussed in the present paper eliminates one of the steps by transforming the quantity of interest in a parameter of an alternative model obtained from the original superpopulation model. In this way only standard statistical techniques needed to be used.

As in most finite population situations, in this section we consider the population total as the quantity of interest and the sample total as the relevant statistic. Hence, we use the following notation through this section:

$$T = Y_1 + \cdots + Y_n \quad \text{and} \quad \boldsymbol{\tau} = Y_1 + \cdots + Y_N \ .$$

The observed value of T is represented by t.

Next we present thee predictive likelihood for the exponential family.

Lemma 3.1. *Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent random quantities with common density function*

$$f(y|(\theta, \phi)) \propto \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \ ,$$

where $a(\cdot), b(\cdot)$ and $c(\cdot, \cdot)$ are known functions. The predictive likelihood in this case is

$$l(\boldsymbol{\tau}, \theta, \phi|\mathbf{y}_S) \propto l(\boldsymbol{\tau}, \phi|t) \propto \exp\left[h_1(t, \phi) + h_2(\boldsymbol{\tau} - t, \phi) - h_3(t, \phi)\right]$$

where h_1, h_2 and h_3 are known functions.

The proof is simple and uses stronglyu the specific sufficiency of $\boldsymbol{\tau}$ (for $\mathbf{Y}$) and t (for $\mathbf{Y}_S$) and the fact that t and $\boldsymbol{\tau} - t$ are statistically independent, in the original model.

In the sequel we present standard examples that will permit the readers to compare with the solutions obtained under the superpopulation approach.

Example 3.2. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent Bernoulli random variables with common unknown parameter (sucess probability) θ. The predictive likelihood is obtained from the conditional distribution of $T|\boldsymbol{\tau}$ which is hipergeometric with parameter $\boldsymbol{\tau}$, population size N and sample size n. The maximum likelihood predictor, $\widehat{\boldsymbol{\tau}}$, is an integer that satisfies

$$(N+1)\frac{T}{n} - 1 \leq \widehat{\boldsymbol{\tau}} \leq (N+1)\frac{T}{n} \ .$$

The construction of the predictive with $100\gamma\%$ confidence for $\boldsymbol{\tau}$, presents the same numerical difficulties of the confidence interval for the parameter of the hipergeometric model.

Example 3.3. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent Poisson random variables with common unknown mean θ. The predictive likelihood is obtained from the conditional distribution of $T|\boldsymbol{\tau}$ which is binomial with the parameters $\boldsymbol{\tau}$ and n/N. For this binomial model, $\boldsymbol{\tau}$ represents the sample size and n/N is the probability of success. The maximum likelihood predictor, $\widehat{\boldsymbol{\tau}}$, is an integer that satisfies

$$N\frac{T}{n} - 1 \leq \widehat{\boldsymbol{\tau}} \leq N\frac{T}{n} \ .$$

The construction of the predictive set for $\boldsymbol{\tau}$ presents the same numerical difficulties of the former example.

Example 3.4. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent geometric random variables with common unknown parameter θ, having common mean equal to $(1-\theta)/\theta$. The predictive likelihood is obtained from the conditional distribution of $T|\boldsymbol{\tau}$ and is equal to

$$l(\boldsymbol{\tau}|t) \propto \frac{\binom{n+t-1}{t}\binom{N-n+\boldsymbol{\tau}-t-1}{\boldsymbol{\tau}-t}}{\binom{N+\boldsymbol{\tau}-1}{\boldsymbol{\tau}}} \ ,$$

where t and $\boldsymbol{\tau}$ are integers satisfying $0 \leq t \leq \boldsymbol{\tau}$. The maximum likelihood

predictor, $\widehat{\boldsymbol{\tau}}$, is an integer that satisfies

$$(N-1)\frac{T}{n} - 1 \leq \widehat{\boldsymbol{\tau}} \leq (N-1)\frac{T}{n} \ .$$

Again, we have the same difficulties to built the predictive set for $\boldsymbol{\tau}$.

Note that the above three examples have in common the fact that the quantity of interest is an integer quantity which brings the standard difficulties of constructing confidence sets. The following examples are related with continuous random quantities and present nice and simple analytical solutions.

Example 3.5. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent Gamma random variables with common positive parameters a (known) and β (unknown) having common mean equal to a/β. The predictive likelihood is obtained from the distribution of $T|\boldsymbol{\tau}$ and is equal to

$$l(\boldsymbol{\tau}|t) \propto \left(1 - \frac{t}{\boldsymbol{\tau}}\right)^{(N-n)a-1} \left(\frac{t}{\boldsymbol{\tau}}\right)^{na} \ .$$

The maximum likelihood predictor is $\widehat{\boldsymbol{\tau}} = (N - a^{-1})\frac{T}{n}$, To obtain the predictive interval for $\boldsymbol{\tau}$ with confidence $100\gamma\%$ we notice the fact that the pivotal quantity $Z = T/\boldsymbol{\tau}$ has a beta distribution with parameter $(na; (N-n)a)$. Hence, the predictive interval for $\boldsymbol{\tau}$ is

$$\left[\frac{T}{d} \ ; \ \frac{T}{c}\right] \ ,$$

where c and d are chosen in such a way that

$$\frac{1}{c} - \frac{1}{d} \text{ is minimum under } \int_c^d f(z)\, dz = \gamma \ .$$

Example 3.6. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent Normal random variables with common unknown mean μ and known variance c^2. The predictive likelihood is obtained from the conditional distribution of $T|\boldsymbol{\tau}$ which is normal with mean $\left[\frac{n}{N}\right]\boldsymbol{\tau}$ and variance $\left[1 - \frac{n}{N}\right] nc^2$. Consequently,

the maximum likelihood predictor of $\boldsymbol{\tau}$ is $\widehat{\boldsymbol{\tau}} = \frac{N}{n}T$ and the predictive interval for $\boldsymbol{\tau}$ is obtained using the pivotal quantity

$$Z = \frac{T - \frac{n}{N}\boldsymbol{\tau}}{c\sqrt{n\left[1 - \frac{n}{N}\right]}} ,$$

that is distributed as a standard normal variable. The predictive interval for $\boldsymbol{\tau}$ is defined by

$$\widehat{\boldsymbol{\tau}} \pm Nzc\sqrt{\frac{1}{n} - \frac{1}{N}} ,$$

where z is such that $\Pr(-z < Z < z) = \gamma$.

The above examples did not involved any nuisance parameter. The following ones are multiparametric cases.

Example 3.7. This is the continuation of Example 2.2. If the quantity of interest is only the population total, $\boldsymbol{\tau}$, then, in order to obtain the predictive interval, the following pivotal quantity can be used:

$$W = \frac{T - \frac{n}{N}\boldsymbol{\tau}}{S\sqrt{n\left[1 - \frac{n}{N}\right]}} ,$$

where $S^2 = \frac{1}{n-1}\left[V - \frac{T^2}{n}\right]$. Since the distribution of W is Student t with $n-1$ degrees of freedom, the predictive interval for $\boldsymbol{\tau}$ is defined by

$$\widehat{\boldsymbol{\tau}} \pm NwS\sqrt{\frac{1}{n} - \frac{1}{N}} ,$$

where w is such that $\Pr(-w < W < w) = \gamma$.

Note that the last example produces formulas that are equal to those obtained by using the classical sampling techniques which is not based on any superpopulation model.

Example 3.8. If in Example 3.7 we also consider the population variance as a quantity of interest, then the second predictive likelihood (expression 2.2) of

Example 2.6 is the one to be used. Suppose that the population variance is defined by

$$\boldsymbol{\phi} = \frac{1}{N-3}\sum_{i=1}^{N}\left(Y_i - \frac{\boldsymbol{\tau}}{N}\right)^2$$

(for simplicity we consider $N-3$ in the place of N). The maximum likelihood predictor of $\boldsymbol{\phi}$ is given by

$$\widehat{\boldsymbol{\phi}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{T}{n}\right)^2 - \frac{1}{n}\left(\frac{\widehat{\boldsymbol{\tau}}^2}{N} - \frac{T^2}{n} - \frac{(\widehat{\boldsymbol{\tau}} - T)^2}{N-n}\right)$$

if this expression is positive and zero otherwise. This predictor can receive interesting interpretations since is a function of the sample variance corrected by a function of the population total, the sample total and the total of the unobserved part of the population. It may not be simple to decide which pivotal quantity to be used to obtain a predictive interval for the population variance. However, a good start could be the fact that the ratio between

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{T}{n}\right)^2 \quad \text{and} \quad \frac{1}{N-n}\sum_{i=n+1}^{N}\left(Y_i - \frac{\boldsymbol{\tau} - T}{N-n}\right)^2$$

is a pivotal quantity with a well known distribution.

Example 3.9. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent Normal random variables with known common variance c^2 and unknown mean of Y_i, for $i = 1, 2, \ldots, N$, equal to $\beta_0 + \beta_1 x_i$ where x_i is a fixed value of a covariance x and β_0 and β_1 are unknown parameters. In this case, $\boldsymbol{\tau}$ is specific sufficient for β_0. If the only quantity of interest is $\boldsymbol{\tau}$, then we use the likelihood obtained from the conditional distribution of $\mathbf{y}_S|\boldsymbol{\tau}$, β_1 which is n-variate normal with mean equal to

$$\frac{\boldsymbol{\tau}}{N}\mathbf{j}_n + \beta_1(\mathbf{x} - \overline{x}_N\mathbf{j}_n)$$

and variance equal to

$$c^2(\mathbf{I}_n - \frac{1}{N}\mathbf{J}_n)\ ,$$

where $\mathbf{j}_n$, $\mathbf{I}_n$ and $\mathbf{J}_n$ are defined as in Example 2.2 and $\mathbf{x} = (x_1, \ldots, x_n)$ and

$$\overline{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is the population mean of the x_i's (note that, equivalently, $\overline{x}_n$ is the sample mean of the x_i's).

The maximum likelihood predictors of β_1 and $\boldsymbol{\tau}$ are, respectively:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_N)\left(Y_i - \frac{T}{n}\right)}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$

and

$$\widehat{\boldsymbol{\tau}} = \frac{N}{n}T - N\widehat{\beta}_1(\overline{x}_n - \overline{x}_N)\ . \tag{3.1}$$

To obtain the predictive interval for $\boldsymbol{\tau}$ we use the following standard normal pivotal quantity

$$\frac{\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}}{Nc\sqrt{\frac{R(\mathbf{x})}{n} - \frac{1}{N}}}\ ,$$

where

$$R(\mathbf{x}) = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_N)^2}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\ .$$

Example 3.10. In Example 3.9 suppose that besides $\boldsymbol{\tau}$,

$$\boldsymbol{\rho} = \sum_{i=1}^{N} Y_i x_i$$

is also a quantity of interest. The vecto $(\boldsymbol{\tau}, \boldsymbol{\rho})$ is a minimal sufficient statistic for the original model in relation to (β_0, β_1). For $Z = \sum_{i=1}^{n} Y_i x_i$ and since (T, Z) is a minimal sufficient reduction of the sample, the predictive likelihood may be obtained from the conditional distribution of $(T, Z)|(\boldsymbol{\tau}, \boldsymbol{\rho})$ which is also normal. To describe the mean and the variance of this distribution we introduce the following notation:

$$s_n \ = \ \frac{1}{n}\left(x_1^2 + \cdots + x_n^2\right)\ ,$$

$$\begin{aligned} s_N &= \frac{1}{N}\left(x_1^2 + \cdots + x_N^2\right) , \\ \Sigma_{11} &= \Sigma_{12} = n \begin{pmatrix} 1 & \overline{x}_n \\ \overline{x}_n & s_n \end{pmatrix} , \\ \Sigma_{22} &= N \begin{pmatrix} 1 & \overline{x}_N \\ \overline{x}_N & s_N \end{pmatrix} \quad \text{and} \\ \Sigma &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12} . \end{aligned}$$

The conditional distribution of $(T, Z)|(\boldsymbol{\tau}, \boldsymbol{\rho})$ is normal with mean μ and variance Σ where

$$\mu = \frac{n}{N(s_N - \overline{x}_N^2)} \begin{pmatrix} s_N - \overline{x}_N\overline{x}_n & \overline{x}_n - \overline{x}_N \\ \overline{x}_n s_N - \overline{x}_N s_n & s_n - \overline{x}_N\overline{x}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\rho} \end{pmatrix} = B \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\rho} \end{pmatrix} .$$

Clearly, the maximum likelihood predictor of $(\boldsymbol{\tau}, \boldsymbol{\rho})$ is given by $B^{-1}(T, Z)'$. It is not difficult to check that the first component of this vector coincides with the expression (3.1). Also by a proper standard transformation, as we present in the next example, we obtain the pivotal quantity that will produce the predictive region for $(\boldsymbol{\tau}, \boldsymbol{\rho})$.

Except for the last example we have been working with univariate random variables. We end this section with a multivariate normal distribution where the quantity of interest is the vector of population totals.

Example 3.11. Let $Y_1, \ldots, Y_N$, the elements of $\mathbf{Y}$, be statistically independent Normal random vector of order k (> 1) with unknown common mean vector μ and known common covariance matrix Σ. The population total, $\boldsymbol{\tau}$, which is also a vector of order k, is a minimal sufficient reduction of $\mathbf{Y}$ with respect to μ. The predictive likelihood can be obtained from the conditional distribution of the sample total T given $\boldsymbol{\tau}$. This distribution is multivariate normal with mean vector $\left[\frac{n}{N}\right]\boldsymbol{\tau}$ and covariance matrix $\left[1 - \frac{n}{N}\right] n\Sigma$. Hence, the maximum likelihood predictor of $\boldsymbol{\tau}$ is $\widehat{\boldsymbol{\tau}} = \frac{N}{n}T$ and the predictive region for $\boldsymbol{\tau}$

is given by

$$\left\{\boldsymbol{\tau} : \frac{N}{(N-n)n}\left(T - \frac{N}{n}\boldsymbol{\tau}\right)' \Sigma^{-1}\left(T - \frac{N}{n}\boldsymbol{\tau}\right) \leq \chi^2\right\} ,$$

where χ^2 is the value of a qui-squared with k degrees of freedom that gives $100\gamma\%$ of confidence.

4. Final remarks

Since we presented cases of smooth likelihoods, to obtain the maximum likelihood predictors in the discrete cases, we compared the likelihood ratios $l(\boldsymbol{\tau}|t)/l(\boldsymbol{\tau} - 1|t)$ and $l(\boldsymbol{\tau}|t)/l(\boldsymbol{\tau} + 1|t)$ with the unity. For the continuous cases sove the equation obtained by making the partial derivative of the log-likelihood equal to zero.

It is important that we understand the relevant role that sufficiency and specific sufficiency play in the method discussed in this article. In order to illustrate this role, let us consider two simple cases.

Again, consider the population total, $\boldsymbol{\tau}$, as the quantity of interest. First we go back to Examples 2.2 and 3.7 where the original model is normal with mean μ and variance σ^2. Suppose that we receive an additional information saying that the mean is known to be zero. The model now has only one unknown parameter, σ^2. The predictive likelihood based on the conditional distribution of the sample given the population total is exactly the same as before, $l(\pi^*|\mathbf{y}_S)$. Hence the information that the parameter μ, of no interest, is known to be zero, would not improve the prediction of $\boldsymbol{\tau}$. We believe that this is not reasonable, since $\boldsymbol{\tau}$ and μ are strongly related.

The second case considered here is the regression case of Examples 3.9 and 3.10. The information that the intercept parameter, β_0, is null also would not change the predictive likelihood and no improvement in the prediction is attained with such an important information. Using Bayesian methods, where

all relevant information is processed, this problem would not occur (Datta & Gosh (1991)).

We end this article by emphasizing that the method discussed here is based on a proper model that is consequence of standard suppositions. In no place, additional suppositions or restrictions were considered. The real question to be discussed is an old one: *what is the likelihood function?* We believe that Bayesians would answer this question saying that, after elimination of nuisance parameters by integration, $l(\boldsymbol{\tau}|\mathbf{y}_S)$ is the correct likelihood function.

(Received January 1993. Revised September 1993.)

References

Bayarry, M.J., De Groot, M.H. and Kadane, J.B. (1986). What is the likelihood function? In: Gupta, S.S. and Berger, J.O. (Eds.). *Proceedings of the Fourth Purdue Symposium of Statistical Decision Theory and Related Topics*. Springer-Verlag, New York, pp. 3–27.

Basu, D. (1969). Role of sufficiency and likelihood principles in sample survey theory. *Sankhyã A*, **31**, 441–54.

Basu, D. (1975). Statistical information and likelihood. *Sankhyã A*, **37**, 1–71.

Basu, D. (1977). On the elimination of nuisance parameters. *JASA*, **72**, 355–66.

Bjørnstad, J.F. (1990). Predictive likelihood: a review (with discussion). *Statistical Science*, **5**, 242–65.

Butler, R.W. (1986). Predictive likelihood inference with applications (with discussion). *JRSS*, **B 47**, 1–38.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1977). *Foundations of inference in survey sampling*. John Wiley, New York.

Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics*, **73(4)**, 1748–1770.

Davidson, A.C. (1986). Approximate predictive likelihood. *Biometrika*, **73**, 323–32.

De Finetti, B. (1987). Probabilities of probabilities: a real problem or a misunderstanding? In: Aykaç, A. and Brumat, C. (Eds.). *New developments in the applications of Bayesian methods*. North Holland, Amsterdam, pp. 1–10.

Faulkenberry, G.D. (1973). A method of obtaining prediction intervals. *JASA*, **68**, 433–5.

Hahn, J.G. (1969). Factor for calculating two sided prediction intervals for samples from a normal distribution. *JASA*, **64**, 878–88.

Hinkley, D.V. (1979). Predictive likelihood. *Ann. Statist.*, **7**, 718–28.

Lauritzen, S.L. (1974). Sufficiency, prediction and extreme models. *Scand. J. Statist.*, **1**, 128–3.

Lawless, J.F. (1972). On prediction intervals for samples from the exponential distribution and prediction limits for system survived. *Sankhyã B*, **34**, 1–14.

Royall, R.M. (1968). An old approach to finite population sampling theory. *JASA*, **63**, 1269–79.

Royall, R.M. (1971). Linear regression models in finite population sampling theory. In: Godambe, V.P. and Sprott, D.A. (Eds.). *Foundations of statistical inference*. Holt, Rinehar and Winston, Toronto, 259–74.

Thatcher, A.R. (1964). Relationships between Bayesian and confidence limits for prediction. *JRSS*, **B 26**, 176–92.

Vit., P. (1973). Interval prediction for a Poisson process. *Biometrika*, **60**, 667–8.