Discussion Papers No. 608, February 2010
Statistics Norway, Division for Statistical Methods and Standards

Jan F. Bjørnstad

Survey sampling: A necessary journey in the prediction world

Abstract:
The design-approach is evaluated, using a likelihood approach to survey sampling. It is argued that a model-based approach is unavoidable from a scientific point of view. Estimating population quantities can then be regarded as a prediction problem. Predictive likelihood methods are considered in various cases, and evaluated by properties of related confidence intervals and asymptotic consistency.

Keywords: likelihood; predictive likelihood; asymptotic consistency.

JEL classification: C10

Address: Jan F. Bjørnstad, Statistics Norway, Division for Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo, Norway. E-mail: jab@ssb.no

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1. Introduction

The traditional approach to survey sampling, primarily based on Neyman (1934), has several shortcomings discussed in the literature the last 40 years. Already in 1966, Godambe discovered the rather strange effect of likelihood considerations on survey sampling and the humorous elephant example in Basu (1971) put the topic at the forefront.

To fix the ideas, let the finite population for the study be denoted by $U = \{1,2, \dots, N\}$ and let y be a variable of interest with population values $\mathbf{y} = (y_1,..., y_N)$. The typical problem is to estimate the total t or population mean t/N. A sample is a subset s of the population, and is selected according to some sampling design $p(s)$, a known probability distribution for all possible subsets of U assumed to be non-informative about **y**. The design-based inference has only s as the stochastic element and considers **y** as a constant. Some of the shortcomings and problems with design-based inference are:

- Design-based inference is with respect to *hypothetical* replications of sampling for a *fixed* population vector **y**
- Variance estimates may fail to reflect information in a *given sample*
- Difficult to combine with models for nonsampling errors like nonresponse
- If we want to measure how a certain estimation method does in quarterly or monthly surveys, then **y** will vary from quarter to quarter or month to month, and we need to assume that **y** is a realization of a random vector.

We shall use likelihood and the likelihood principle as a guideline on how to deal with these matters. Section 2 discusses the design approach from a likelihood perspective and argues for the necessity of modelling the population. Section 3 considers likelihood in model-based survey sampling as a special case of prediction and Section 4 deals with predictive likelihood methods and asymptotic consistency features in general prediction problems. Section 5 applies the predictive likelihood approach in model-based survey sampling and consider three different cases. Predictive likelihood is a general non-Bayesian likelihood approach to prediction; see Hinkley (1979) and Butler (1986). A review is given in Bjørnstad (1990,1998).

Bolfarine and Zacks (1992) consider methods based on predictive likelihood in survey sampling.

2. Discussion of design-approach from the likelihood perspective

That there is something strange about the purely design-model approach, is the nonexistence of optimal estimators. First discovered by Godambe (1955) for linear unbiased estimators and then by Godambe and Joshi (1965) for the general case, we have the following theorem:

Theorem

Let $p(s)$ be any nontrivial sampling design, i.e., $p(U) < 1$. Assume each y_i has at least two possible values. Then there exists no uniformly best (minimum variance) design-unbiased estimator for the total t.

No matter how small a population is and how simple the sampling design is we cannot find any uniformly best estimator. This negative fact should really make every survey statistician take notice and do some serious reflections about the design-model. Godambe (1966) was first to consider the likelihood function noticing that the likelihood function is flat for all possible values of **y** given a set of sample values. Hence, from the perspective of the likelihood principle, the model is “empty”; it gives no information about the unknown part of **y.** Moreover from the likelihood principle, since two sampling plans leading to the same sample s has proportional likelihood functions, statistical inference should not depend on the sampling plan. And what else is there from a design point of view?

The only way to still have trust in the design-approach is to disregard the likelihood principle, but since the likelihood principle follows from the principles of sufficiency and conditionality as shown by Birnbaum (1962), then one has to claim that either the sufficiency principle and/or the conditionality principle is not valid. This seems like an impossible task considering that practically no statistician disagrees with these two principles.

So, to sum up, we have the following rather troublesome features of a scientific nature with a pure design-approach to survey sampling:

1) Nonexistence of best estimators no matter what sampling design, sample size and population.
2) A flat likelihood function telling us the data gives us no information about the unknown values in the population. One might say the design-model is a model of "no information" about the unknown part of the population.
3) The sampling plan is irrelevant for doing statistical inference according to the likelihood principle
4) The likelihood principle follows from generally accepted principles of sufficiency and conditionality also in survey sampling

To my mind, there is simply nothing more to discuss. One has to accept that the design approach has a model-basis saying that the data contain no information about the unknown part of the population, and in order to do proper statistical inference one has to model the data versus the unknown quantities as in *any other statistical investigation*. Simply because we have more control of the data collection in survey sampling than in the typical observational study does not mean that we shouldn't do statistical modelling. On the contrary, it should in principle be *easier* in finite population studies based on a controlled sample to do proper statistical modelling than in observational studies.

So as a conclusion on using likelihood considerations on the traditional sampling approach, it reveals the flaws very clearly and tells us what to do. We simply can not avoid following Fisher's modelling and likelihood point of view that revolutionized the science of statistics in the early 1920's. Fisher's fundamental concepts are still very much the focus point of statistical science in all fields of statistics.

It is easy to come up with examples that show real practical shortcomings of the design-approach. For example, regarding variance estimation if one possible sample is the whole population, the estimated sample variance of an estimator would give a meaningless result if the actual sample chosen is the whole population, while the model-based variance is the variance of the prediction error which in this case is zero.

A rather common misunderstanding when it comes to disregarding the sampling design in the inference phase, is that the sampling design is therefore not important. This is, of course, not true. In fact, the opposite is the case. The sampling design is very important for gathering data in the production of official statistics (and for any other finite population study). It is important that we get as informative data as possible for the population at hand making the optimal statistical inference of highest possible quality. This means, typically, that in business surveys to have a high degree of coverage while in household/person statistics we want a representative sample, like a miniature of the population. But once we have made sure we have a good quality sample, the actual plan that was used to select the sample should play no role at the inference stage.

Now, what to do with nonsampling errors like nonresponse is not in principle difficult. There is no way around the fact that we do need to do modelling for these errors. The problem here, of course, is that we do not observe the nonresponse group in the sample. Hence, any modelling here is of a latent type that can be checked for validity only based on what we observe. We have to use the knowledge we have about the units not responding in the actual survey. Of course, closing our eyes and assuming that nonresponse doesn't matter except getting a smaller sample than planned, is also a modelling assumption, and typically of the worst kind.

Once a modelling approach is undertaken, we have the special feature in finite population estimation problems that the unknown quantities are realized values of random variables, so the basic problem has now the feature of being similar to a prediction problem. It is therefore natural to look at a likelihood-based prediction approach here. This leads to predictive likelihood as the basic approach. We shall see what this entails.

3. Likelihood in model-based survey sampling

We now have the following model set-up:

$y_1, y_2, ..., y_N$ are realized values of random variables $Y_1, Y_2, ..., Y_N$.

We have two stochastic elements in the model:

1) Sample $s \sim p(\cdot)$

2) $(Y_1, Y_2, \ldots, Y_N) \sim f_\theta$

In general we shall let $f_\theta(\cdot)$ ($f_\theta(\cdot|\cdot)$) denote the (conditional) probability density or discrete probability function of the enclosed variables. Let us consider the problem of estimating the total t which we can decompose as

$$t = \sum\nolimits_{i \in s} y_i + \sum\nolimits_{i \notin s} y_i .$$

Since the first term is observed, the problem is to estimate $z = \sum\nolimits_{i \notin s} y_i$, the realized value of the random variable

$$Z = \sum\nolimits_{i \notin s} Y_i .$$

Hence, we may say that the problem is to *predict* the value z of Z. This means that the parameter θ labelling the class of distributions for **Y** is a nuisance parameter. Now, the first basic question when it comes to likelihood considerations under a population model is how to define the likelihood function. From a general predictive perspective, if we let $Y_d = y_d$ denote the data in s and Z the unknown variable whose value z we shall predict, Bjørnstad (1996) shows that the likelihood function $l(z,\theta) = f_\theta(y_d, z)$ leads to a likelihood principle that follows from generalized principles of prediction sufficiency and conditionality in the same way as the parametric likelihood function. Hence this is also the likelihood function in the sampling case. The data y_d consists now of s and the observed y-values in s. A likelihood-based method for predicting z is then a partial likelihood $L(z|y_d)$ based on $l(z, \theta)$, by eliminating θ. Typical ways of eliminating θ is by integration (resembling Bayes approach), maximization (resembling the profile likelihood in parametric inference), and conditioning on sufficient statistics. We shall now first, in Section 4, consider predictive likelihoods in general, and in Section 5 predictive likelihood in model-based survey sampling for some specific cases.

4. Predictive likelihood with asymptotic considerations and benchmarks

For a summary and review of predictive likelihood we refer to Bjørnstad (1990, 1998). We shall assume that a chosen predictive likelihood is normalized as a probability distribution in *z*. We shall first consider the problem of asymptotic consistency in predicting sample means, resembling the typical problem of estimating the finite population total in survey sampling. Assume the data consists of *n* observations. Throughout this section we shall let the data be denoted by *y*, i.e., *y* is a realized value of $Y = (X_1,...,X_n)$. We consider the problem of predicting the mean of the unobserved "sample" $Y' = (X_1^{'},...,X_m^{'})$, i.e., $Z = Z_m = \sum_{i=1}^{m} X_i^{'} / m.$

Let now $E_p(Z)$ and $V_p(Z)$ be the (predictive) mean and variance of the normalized predictive likelihood $L(z|y)$. Then $E_p(Z)$ is one possible predictor of *z.* Another important issue in prediction is whether the predictive variance is a correct measure of the prediction uncertainty. Hence, one important aspect of evaluating how a certain predictive likelihood performs as a prediction method is the property of the predictive variance. The main purpose now is to study how $E_p(Z)$ and $V_p(Z)$ should behave asymptotically in *n* and *m.* It is difficult to define benchmarks for the predictive mean and variance for fixed small *m* and *n*. However, for large *m* or large *n* (typical cases in sampling, the first case being typical for sample-based statistics while the second case is typical for register-based statistics) it is possible to derive approximate benchmarks by considering the two asymptotic cases (i) $n \to \infty$ and (ii) $m \to \infty$ separately. If $n \to \infty$, θ is known in the limit. In this case the normalized predictive likelihood is the normalized $l(z,\theta)$, $f_\theta(z \mid y)$. A natural consistency requirement for predictive likelihood is therefore that

$$L(z \mid Y) / f_\theta(z \mid Y) \overset{P}{\longrightarrow} 1 \text{ as } n \to \infty .$$

It is assumed that, conditional on $Y = y, Z_m \overset{P}{\to} \mu$ as $m \to \infty$, where $\mu = g(\theta)$ may depend on *y* if *Y*, *Z* are dependent. When $m \to \infty$, predicting *z* is equivalent to estimating μ in the limit. Let $l(\mu|y)$ denote the chosen normalized likelihood for μ, based on the parametric likelihood function for θ, $lik(\theta|y) = f_\theta(y)$. We denote the mean and variance by $E_l(\mu)$ and $V_l(\mu)$. If $\theta = \mu$, then, of course, $l(\mu|y) \propto f_\mu(y)$. In the general case, when $\mu = g(\theta)$, there are several possible choices for $l(\mu|y)$. It is not possible to avoid a certain degree of arbitrariness. In the 1970's and primo 1980's several articles studied the problem of choosing a marginal parametric likelihood. Two main papers are Kalbfleisch and Sprott (1970) and Barndorff-Nielsen (1983). We shall choose to derive the marginal likelihood in the following way: Normalize the likelihood function for θ to be a probability distribution in θ. Let $l_y(\theta)$ be the normalized likelihood, $l_y(\theta) = lik(\theta \mid y) / \int lik(\theta' \mid y) d\theta'$. Let then $l(\mu|y)$ be the "distribution" of μ, derived from $l_y(\theta)$. Then, e.g., the likelihood expected value of μ is $E_l(\mu) = \int g(\theta) l_y(\theta) d\theta$.

We can summarize these discussions by defining variance consistency and mean consistency as follows:

Definition 1. The predictive likelihood L is variance consistent if the following two properties are satisfied:

1.1. $V_p(Z) / V_\theta(Z \mid Y) \overset{P}{\rightarrow} 1$ as $n \rightarrow \infty$

1.2. $V_p(Z) \rightarrow V_l(\mu)$ as $m \rightarrow \infty$

Definition 2. The predictive likelihood L is mean consistent if the following two properties hold

2.1. $E_p(Z) / E_\theta(Z \mid Y) \overset{P}{\rightarrow} 1$ as $n \rightarrow \infty$

2.2. $E_p(Z) \rightarrow E_l(\mu)$ as $m \rightarrow \infty$

We see that if Z and Y are independent, which is typically the case in model-based sampling, L is variance consistent if

$$V_p(Z) \overset{P}{\rightarrow} V_\theta(Z) \text{ as } n \rightarrow \infty \text{ and } V_p(Z) \rightarrow V_l(\mu) \text{ as } m \rightarrow \infty, \tag{1}$$

and mean consistent if

$$E_p(Z) \overset{P}{\rightarrow} E_\theta(Z) \text{ as } n \rightarrow \infty \text{ and } E_p(Z) \rightarrow E_l(\mu) \text{ as } m \rightarrow \infty. \tag{2}$$

Let us consider four basic predictive likelihoods and some examples. The estimative predictive likelihood L_e is obtained by eliminating θ in the likelihood function using the maximum likelihood estimate (mle) $\hat{\theta}$, i.e, the normalized L_e is given by

$$L_e(z \mid y) = f_{\hat{\theta}}(z \mid y)\,.$$

The profile predictive likelihood L_p, first considered by Mathiasen (1979), is obtained by maximizing the likelihood function with respect to θ for a given z value, i.e.,

$$L_p(z \mid y) = \max\nolimits_\theta l_y(z,\theta) = l_y(z,\hat{\theta}_z)\,.$$

Let $R = r(Y, Z)$ be a minimal sufficient statistic for Y and Z. In cases where sufficiency provides a true reduction in the dimension of the data, Hinkley (1979) suggested essentially the conditional predictive likelihood L_c given by

$$L_c(z \mid y) = f(y,z \mid r(y,z)) = f_\theta(y,z)/f_\theta(r(y,z))\,.$$

L_c is not invariant with respect to choice of minimal sufficient statistics in the continuous case. A canonical-type of conditional predictive likelihood, suggested by Butler (1986), turns out to be invariant to choice of R. It is given by

$$L_I(z \mid y) = L_c(z \mid y) \mid JJ' \mid^{-1/2}$$

where J is the pxq – matrix of partial derivatives of r with respect to (y,z). Here, p is the dimension of r and q is the dimension of (y,z).

A (1- α) predictive interval I_L based on a normalized predictive likelihood L is simply an interval with area (1- α) under L,

$$\int_{I_L} L(z \mid y)dz \; (\textstyle\sum_{I_L} L(z \mid y) \text{ in discrete case}) = 1-\alpha.$$

Example 1. Consider X_i, $X_j^{'}$ independent $N(\mu,\sigma_0^2)$ where σ_0^2 is known and let Z be the mean of the $X_j^{'}$'s . Then L_c, L_I, L_p all give the same predictive likelihood, $L \sim N(\bar{x},(m^{-1}+n^{-1})\sigma_0^2)$, where $\bar{x} = \sum_{i=1}^{n} x_i / n$ is the observed sample mean. Since μ is the only unknown parameter, $l(\mu \mid y) \propto f_\mu(y)$, i.e., $l(\mu \mid y) \sim N(\bar{x},\sigma_0^2 / n)$. Hence, $E_l(\mu) = \bar{x}, V_l(\mu) = \sigma_0^2 / n.$ From (1) and (2) we readily see that mean and variance consistency hold. On the other hand, $L_e \sim N(\bar{x},\sigma_0^2 / m)$, and L_e is not variance consistent as $m \to \infty$, illustrating the well known fact that L_e in general underestimates the prediction uncertainty, by assuming that $\theta = \hat{\theta}$ without taking into consideration the uncertainty in the mle $\hat{\theta}$. We also note that the symmetric predictive interval equals the usual frequentistic prediction interval for Z.

Example 2. Same model as in example 1, except that the variance σ^2 in the normal distribution is now unknown. Then the four predictive likelihoods give different results. Let $\hat{\sigma}^2$ be the mle, and let t_v denote the t-distribution with v degrees of freedom. Define

$$T = \frac{Z - \bar{x}}{\hat{\sigma}\sqrt{\frac{1}{m}+\frac{1}{n}}}.$$

Then L_p is such that $T \sim t_n$. With R = (R_1, R_2) where $R_1 = (n\bar{X} + mZ)/(n+m)$ and $R_2 = \sum_{i=1}^{n}(X_i - R_1)^2 + m(Z - R_1)^2$, L_c is such that $\sqrt{(n-3)/n} \cdot T \sim t_{n-3}$ .The canonical predictive likelihood L_I does not directly lead to a t-distribution. However, L_I based on the transformed $(Y, \sqrt{m}Z)$ is such that $\sqrt{(n-2)/n} \cdot T \sim t_{n-2}$. The estimative L_e is such that $Z \sim N(\bar{x}, \hat{\sigma}^2 / m)$. For all four predictive likelihoods, $E_p(Z) = \bar{x}$. The predictive variances, on the other hand, are all different. We have that the variance of the prediction error, using the sample mean to predict z, equals $V_\theta(Z - \bar{X}) = (\frac{1}{m} + \frac{1}{n})\sigma^2$. Hence, $s_e^2 = (\frac{1}{m} + \frac{1}{n})\hat{\sigma}^2$ is the estimated variance of the prediction error. With the obvious notation we have $V_p^P(Z) = \frac{n}{n-2} s_e^2$, $V_p^c(Z) = \frac{n}{n-5} s_e^2$, $V_p^I(Z) = \frac{n}{n-4} s_e^2$, while $V_p^e(Z) = \hat{\sigma}^2 / m = s_e^2 - \frac{1}{n}\hat{\sigma}^2$. The likelihood for μ is such that $\sqrt{n-2}(\mu - \bar{x}) / \hat{\sigma} \sim t_{n-2}$. Hence, $E_l(\mu) = \bar{x}$ and $V_l(\mu) = \hat{\sigma}^2 /(n-4)$. All predictive likelihoods are mean consistent. Also, $V_p(Z) \overset{P}{\to} \sigma^2 / m = V_\theta(\mathrm{Z})$ as $n \to \infty$ for all four predictive likelihoods. Hence, they are all variance consistent in n. Variance consistency in m holds if

$V_p(Z) \to \hat{\sigma}^2$/n - 4) as $m \to \infty$. Now, $s_e^2 \to \hat{\sigma}^2 / n$ as $m \to \infty$, and as $m \to \infty$, $V_p^p(Z) \to \hat{\sigma}^2 /(n-2), V_p^c(Z) \to \hat{\sigma}^2 /(n-5)$, $V_p^I(Z) \to \hat{\sigma}^2 /(n-4)$ and $V_p^e(Z) \to 0$. Hence, according to this choice of marginal likelihood for μ, L_I is variance consistent, while L_p and L_c are approximately variance consistent. L_c slightly overestimates and L_p slightly underestimates the prediction uncertainty when using $l(\mu|y)$ as benchmark.

5. Predictive likelihood in model-based survey sampling

We shall in this section consider three cases, the first case is a model typically used in business surveys, the second case deals with election surveys and the third case deals with mixtures covering two-stage sampling and missing data with MCAR nonresponse.

5.1. Ratio model

Let us start with a typical model in business surveys, the ratio model. It is usually stratified, but we shall for simplicity consider the pure ratio model. It means that we have an auxiliary variable x available for all units in the population. It is typically a measure of size of the unit, like the number of employees or annual sales of the business. Then the model is given by:

$$Y_i = \beta x_i + \varepsilon_i \text{ for } i = 1,\ldots,N \text{ and the } \varepsilon_i\text{'}s \text{ are independent } N(0, \sigma^2 v(x_i)).$$

Here, $v(x)$ is a known function like $v(x) = x^g$, $0 \leq g \leq 2$. The usual assumption is $g = 1$.
The optimal predictor among all linear model-unbiased predictors for the total is given by

$$\hat{t}_0 = \sum_{i \in s} y_i + \hat{\beta}_0 \sum_{i \notin s} x_i$$

where

$$\hat{\beta}_0 = \frac{\sum_{i \in s} x_i y_i / v(x_i)}{\sum_{i \in s} x_i^2 / v(x_i)}.$$

Hence, the predictor for the unobserved part of the total equals $\hat{z}_0 = \hat{\beta}_0 \sum_{i \notin s} x_i$.

Let $v(\bar{s}) = \sum_{i \notin s} v(x_i)$, $x(\bar{s}) = \sum_{i \notin s} x_i$, and $w_s = \sum_{i \in s} x_i^2 / v(x_i)$. The profile predictive likelihood is such that

$$\frac{Z-\hat{\beta}_0\sum_{i\notin s}x_i}{\hat{\sigma}\sqrt{v(\bar{s})+w_s^{-1}[x(\bar{s})]^2}}\sim t_n-\text{distribution.}$$

We note that the predictive mean is equal to $\hat{z}_0$, the optimal predictor. The predictive variance is given by

$$V_p(Z)=\frac{n}{n-2}\hat{\sigma}^2\left\{v(\bar{s})+w_s^{-1}[x(\bar{s})]^2\right\}.$$

The variance of the prediction error ($Z-\hat{z}_0$) is equal to $\sigma^2\left\{v(\bar{s})+w_s^{-1}[x(\bar{s})]^2\right\}$. Hence, the predictive variance is essentially the estimated variance of the prediction error
Letting R be the mle of (β,σ^2) based on (Y_d, Z), we find that the conditional predictive likelihood L_c is such that

$$\sqrt{\frac{n-3}{n}}\cdot\frac{Z-\hat{\beta}_0\sum_{i\notin s}x_i}{\hat{\sigma}\sqrt{v(\bar{s})+w_s^{-1}[x(\bar{s})]^2}}\text{ has a }t_{n-3}-\text{distribution.}$$

Let $t_k(\alpha/2)$ be the upper $\alpha/2$- quantile of the t_k-distribution. The (1- α) predictive intervals I_p, I_c based on L_p and L_c are given by

$$I_p:\hat{z}_0\pm t_n(\alpha/2)\hat{\sigma}\sqrt{v(\bar{s})+w_s^{-1}[x(\bar{s})]^2}$$

$$I_c:\hat{z}_0\pm t_{n-3}(\alpha/2)\hat{\sigma}\sqrt{\frac{n}{n-3}}\sqrt{v(\bar{s})+w_s^{-1}[x(\bar{s})]^2}$$

while the frequentistic interval with coverage (1- α) equals

$$I_f:\hat{z}_0\pm t_{n-1}(\alpha/2)\sqrt{\frac{n}{n-1}}\sqrt{v(\bar{s})+w_s^{-1}[x(\bar{s})]^2}\ .$$

It follows that L_p generates prediction intervals with coverage slightly less than the nominal level, while L_c leads to slightly wider intervals than the frequentistic one. Some cases are presented in Table 1. One should note that the usual unconditional confidence level is a measure of the method and, from a likelihood perspective, is not in principle a relevant feature of the actual computed prediction interval. From the likelihood perspective it is necessary to look at the conditional coverage given the data and the guarantee of conditional coverage, as considered in Aitchison and Dunsmore (1975). For a discussion of these features on predictive intervals we refer to Bjørnstad (1990, 1996).

Table 1. Confidence levels of predictive intervals based on $L_p(L_c)$

(1-α) \ n	*5*	*10*	*20*	*50*
0.90	*0.854 (0.986)*	*0.880 (0.940)*	*0.890 (0.918)*	*0.896 (0.907)*
0.95	*0.917 (0.996)*	*0.936 (0.975)*	*0.944 (0.962)*	*0.948 (0.955)*

5.2. Election surveys

The problem is to estimate the proportion p in a population that will vote for a certain party A in an upcoming election. We know the proportion q that voted for A in the last election. For each individual in the population we define the following binary variables,

$$y_i = \begin{cases} 1 & \text{if the i'th person will vote for A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_i = \begin{cases} 1 & \text{if the } i'\text{th person voted for A in the last election} \\ 0 & \text{otherwise} \end{cases}$$

We assume the following model: The y_i's are realized values of random variables Y_i's and $Y_1,\ldots,Y_N$ are independent with "transition" probabilities

$$P(Y_i = 1 \mid x_i = 1) = p_{11} \text{ and } P(Y_i = 1 \mid x_i = 0) = p_{01}.$$

A sample *s* of size *n* is selected and the *y*- and *x*- values in *s* are observed. Estimation of *p* is equivalent to prediction of $z = \sum_{i \notin s} y_i$. Let $\bar{s}_1 = \{i \notin s : x_i = 1\}$ and $\bar{s}_0 = \{i \notin s : x_i = 0\}$. Then *Z* = $Z_1 + Z_0$, where

$$Z_1 = \sum_{i \in \bar{s}_1} Y_i = \sum_{i \notin s} x_i Y_i \text{ and } Z_0 = \sum_{i \in \bar{s}_0} Y_i = \sum_{i \notin s} (1 - x_i) Y_i \ .$$

Let $m = N\text{-}n = m_1 + m_0$, where $m_1 = |\bar{s}_1|$ and $m_0 = |\bar{s}_0|$. We see that Z_1, Z_0 are independent, binomially distributed with parameters (m_1, p_{11}) and (m_0, p_{01}) respectively. Let $B_1 = \sum_{i \in s} x_i Y_i$ and $B_0 = \sum_{i \in s} (1 - x_i) Y_i$, and let $n_1 = \sum_{i \in s} x_i$ and $n_0 = \sum_{i \in s} (1 - x_i)$. Then the mle are $\hat{p}_{11} = B_1 / n_1$ and $\hat{p}_{01} = B_0 / n_0$.

Since the distribution of *Z* is not on a closed form we shall derive a joint predictive likelihood for (Z_1, Z_0) based on $f_\theta(y_d, z_1, z_0)$. Based on this joint predictive likelihood we can obtain the predictive mean and variance for *Z*. We shall apply the sufficiency-based conditional L_c . It turns out that

$$L_c(z_1, z_0 \mid y_d) = L_c(z_1 \mid y_d) L_c(z_0 \mid y_d)$$

with

$$L_c(z_i \mid y_d) = \frac{\binom{m_i}{z_i}\binom{n_i}{b_i}}{\binom{m_i + n_i}{z_i + b_i}} \cdot \frac{n_i + 1}{m_i + n_i + 1}, \ 0 \le z_\mathrm{i} \le m_i, \ i = 1{,}0.$$

This means that Z_1, Z_0 are predictively independent and negative hypergeometric. It follows that $E_p(Z) = E_p(Z_1) + E_p(Z_0)$, and $V_p(Z) = V_p(Z_1) + V_p(Z_0)$, where

$$E_p(Z_i) = m_i \frac{b_i + 1}{n_i + 2} \text{ and } V_p(Z_i) = m_i \frac{n_i + m_i + 2}{n_i + 3} \cdot \frac{b_i + 1}{n_i + 2} \cdot \left(1 - \frac{b_i + 1}{n_i + 2}\right).$$

We see that $Z/m \to^{P} \lambda p_{11} + (1-\lambda)p_{01} = \mu$, as $m \to \infty, m_1/m \to \lambda$.

We shall now consider the asymptotic properties of $E_p(Z)$ and $V_p(Z)$. We note that these are the predictive mean and variance of Z based on the convolution

$$L_c^*(z \mid y_d) = \sum_{k=0}^{z} L_c(z_1 = k \mid y_d) L_c(z_0 = z - k \mid y_d).$$

L_c^* is the convolution of two negative hypergeometric distributions and can be computed exact only numerically.

From (1) and (2) the asymptotic consistency requirements are:

Variance consistency

V1: $V_p(Z) \overset{P}{\to} V_\theta(Z)$ as $n_1, n_0 \to \infty$

V2: $V_p(Z/m) = V_p(Z)/m^2 \to V_l(\lambda p_{11} + (1-\lambda)p_{01})$ as $m_1, m_0 \to \infty$,

$\lambda = \lim(m_1/m)$

Expectation consistency

E1: $E_p(Z) \overset{P}{\to} E_\theta(Z)$ as $n_1, n_0 \to \infty$

E2: $E_p(Z/m) \to E_l(\lambda p_{11} + (1-\lambda)p_{01})$ as $m_1, m_0 \to \infty$

In this case there are unique marginal likelihoods for p_{11} and p_{01}, since the likelihood function is given by

$$lik(p_{11}, p_{01} \mid y_d) = p_{11}^{b_1}(1-p_{11})^{n_1-b_1} p_{01}^{b_0}(1-p_{01})^{n_0-b_0} = l_1(p_{11} \mid y_d) l_0(p_{01} \mid y_d)$$

and $l_i(p_{i1} \mid y_d) \sim Beta(b_i+1, n_i-b_i+1)$ for $i = 1,0$. Hence,

$$E_l(\mu) = \lambda E_l(p_{11}) + (1-\lambda)E_l(p_{01})$$
$$V_l(\mu) = \lambda^2 V_l(p_{11}) + (1-\lambda)^2 V_l(p_{01})$$

where $E_l(p_{i1}) = (b_i+1)/(n_i+2)$ and $V_l(p_{i1}) = (b_i+1)(n_i-b_i+1)/\{(n_i+2)^2(n_i+3)\}$.

We readily see that V1,V2 and E1,E2 are fulfilled. So the derived predictive likelihood L_c^* for *Z* is variance and expectation consistent. In this connection we note that the mle based predictor of *Z*, $\hat{Z}_{mle} = m_1\hat{p}_{11} + m_0\hat{p}_{01}$, is not exactly mean consistent, even though is it the uniformly best unbiased linear predictor, i.e., minimizing the variance of the prediction error, as shown by Thomsen (1981).

We shall now study a prediction interval based on $L_c(z_1, z_0 \mid y_d)$, i.e., $L_c^*(z \mid y_d)$. L_c^* is approximately normal when $(n_1, m_1), (n_0, m_0)$ and (b_1, b_0) are large. Computations suggest the normal approximation is valid already when $N = 50$, $n = 20$ and $b_1 + b_0 = 10$. Let $u(\alpha/2)$ be the upper $\alpha/2$-quantile in the N(0,1) – distribution. An approximate (1-α) predictive interval based on L_c^* is now:

$$I_c(Y_s): E_p(Z) \pm u(\alpha/2)\sqrt{V_p(Z)}\,.$$

Here, the notation Y_s stands for the *y* –observations in the sample *s*. The interval I_c should work fairly well, since the actual distribution of *Z* is approximately normal for large m_1, m_0. The confidence level of I_c conditional on selected sample *s*, $P_\theta(Z \in P_c(Y_s))$, can be estimated for various cases by simulation of the population model. Consider 1- α = 0.95, and let *q* be the proportion who voted for A in the last election. For each case of (n, n_1, N, q), 12 combinations of p_{11} and p_{01} are considered: p_{01} = 0.01, 0.10, 0.30 and p_{11} = 0.5, 0.7, 0.8, 0.9. The confidence levels C_c are estimated by simulating, for each case, 10 000 observations of (Y_s, Z_1, Z_0). The smallest and largest confidence levels over these 12 combinations are given in Table 2.

Table 2. Confidence levels for 12 combinations of the parameters

	n	N	q	n$_1$	*Confidence level*
(I)	*10*	*100*	*0.5*	*3, 7*	*0.939 -0.999*
	10	*100*	*0.1*	*1, 3*	*0.933 - 1*
(II)	*100*	*1000*	*0.5*	*40, 60*	*0. 943 – 0.967*
	100	*1000*	*0.1*	*5, 15*	*0.947 – 0.998*
(III)	*1000*	10^4, 10^6	*0.5*	*400, 600*	*0.947 – 0.955*
	1000	10^4, 10^6	*0.1*	*75, 125*	*0.947 – 0.964*

In the most typical real-life cases, i.e. cases (III), when q = 0.5, there are no systematic trends in C_c as functions of (p_{11}, p_{01}). The same holds true when q = 0.1 and p_{01} = 0.1, 0.3. The values of C_c for all these cases lie in the range 0.947 - 0.955. When q = 0.1 and p_{01} = 0.01, C_c increases slightly as p_{11} increases.

For cases (I) and (II), C_c vary, not unexpectedly, quite a bit more. For given p_{01} there is either an increasing trend as p_{11} increases or there is no systematic trend. For cases (II), the high values occur for the most extreme parameter configuration, p_{11} = 0.9, p_{01} = 0.01.

In short we can say: For large samples it seems that I_c is an approximate (1- α) confidence interval, and for small and moderate sample sizes I_c is mainly conservative, i.e., the confidence level is larger than (1- α).

5.3. Prediction of double mixtures

We shall consider prediction of variables of the following form:

$$Z = Z_1 + Z_2 = \sum_{i=1}^{A_m} X_i^{'} + \sum_{i=1}^{B_n} X_i^{''}.$$

Here, A_m may be a random variable be non-decreasing in m and $A_m \to \infty$ in probability as $m \to \infty$. B_n is assumed non-decreasing in n, $B_n \to \infty$ in probability as $n \to \infty$, and is either a function of Y or a constant. This case is designed to cover cases where the "sample" size for the unobserved Z depends also on n, for example when we have nonresponse. Another example of this type of situation with typically large A_m, B_n is two-stage survey sampling with unknown cluster sizes considered by Bjørnstad and Ytterstad (2008).

To simplify the exposition we restrict attention to the case where $Y_s, A_m, B_n, X_i^{'}, X_j^{''}$ are independent. All $X_i^{'}, X_j^{''}$ are assumed independent with the same distribution. Let $\mu = \mu(\theta) = E_\theta(X_i^{'}) = E_\theta(X_j^{''})$ and $\sigma^2 = \sigma^2(\theta) = Var_\theta(X_i^{'}) = Var_\theta(X_j^{''})$.

Let now $L(z_1, z_2 \mid y_d)$ be a predictive likelihood for (z_1, z_2) from which we derive $L(z|y)$, $L(z_1|y_d)$ and $L(z_2|y_d)$. The predictive covariance, $\text{cov}_p(Z_1, Z_2)$ is then the covariance in $L(z_1, z_2 \mid y_d)$ .Clearly, $E_p(Z) = E_p(Z_1) + E_p(Z_2)$ and $V_p(Z) = V_p(Z_1) + V_p(Z_2) + 2\,\text{cov}_p(Z_1, Z_2)$. Even when Z_1, Z_2 are independent we typically have $\text{cov}_p(Z_1, Z_2) \neq 0$, since prediction of Z_1, Z_2 both depend on the same y_d.

Example 3

A typical case is when we have a sample s of size n from a finite population of size N in order to estimate the population total, and we also have nonresponse such that the actual data is from the response sample s_r with size n_r. Let $A_m = m = N - n$, while the $X_j^{''}$'s are the missing values such that $B_n = n - n_r$. Consider the simple case of MCAR nonresponse and $X_1, ..., X_{n_r}, X_1^{'}, ..., X_m^{'}, X_1^{''}, ..., X_{n-n_r}^{''}$ independent with common distribution $N(\mu, \sigma_0^2)$, where

σ_0^2 is known. Let $\bar{x}$ be the observed sample mean in s_r. Then $L_c(z_1, z_2 \mid y_d)$ is bivariate normal with means $((N-n)\bar{x}, (n-n_r)\bar{x})$ and variance-covariance matrix V given by

$$V = \sigma_0^2 \begin{pmatrix} (N-n)(N-n+n_r)/n_r & (N-n)(n-n_r)/n_r \\ (N-n)(n-n_r)/n_r & (n-n_r)n/n_r \end{pmatrix}.$$

Consider the case where $A = A_m$ is stochastic and suppose $f_\theta(z_1|a)$ is easily found while $f_\theta(z_1)$ is not. We then propose a joint predictive likelihood for (Z_1, Z_2, A) of the form

$$L(z_1, z_2, a \mid y_d) = L_a(z_1, z_2 \mid y_d)L(a \mid y_d). \tag{3}$$

where $L_a(z_1, z_2 \mid y_d)$ is based on $f_\theta(y_d, z_1, z_2 \mid a)$. From (3) we obtain the marginal joint predictive likelihood $L(z_1, z_2 \mid y_d)$. Let $E_p(Z_i \mid a)$ and $V_p(Z_i \mid a)$ be the mean and variance of Z_i from $L_a(z_1, z_2 \mid y_d)$. Since $L_a(z_1, z_2 \mid y_d)$ and $L(a| y_d)$ are regular probability distributions we have that

$$E_p(Z_i) = E_p\{E_p(Z_i \mid A)\},$$

$$V_p(Z_i) = E_p\{V_p(Z_i \mid A)\} + V_p\{E_p(Z_i \mid A)\}$$

and $\operatorname{cov}_p(Z_1, Z_2) = E_p\{\operatorname{cov}_p(Z_1, Z_2 \mid A)\} + \operatorname{cov}_p\{E_p(Z_1 \mid A), E_p(Z_2 \mid A)\}$.

Typically $L_a(z_2 \mid y_d) = L(z_2 \mid y_d)$ and then $\operatorname{cov}_p(Z_1, Z_2) = E_p\{\operatorname{cov}_p(Z_1, Z_2 \mid A)\}$.

We observe that $Z_1 / A_m \underset{m}{\overset{P}{\rightarrow}} \mu$ and $Z_2 / B_n \underset{n}{\overset{P}{\rightarrow}} \mu$. When $n \rightarrow \infty$, θ is known in the limit. Hence, prediction of Z_2/B_n should be done with perfection, i.e., $E_p(Z_2 / B_n) \overset{P}{\rightarrow} \mu$ and $V_p(Z_2 / B_n) \overset{P}{\rightarrow} 0$. The predictive likelihood of $Z_1 + B_n^{-1}Z_2$ in the limit should then be $f_\theta(z_1 + B_n^{-1}z_2)$. Hence, Z_1 and Z_2 / B_n are predictively independent in the limit. When $m \rightarrow \infty$, prediction of Z_1/A_m is equivalent in the limit to estimating μ. Let $\bar{Z}_1 = Z_1 / A_m$. Using the same approach as in (3), $L(\bar{z}_1, a \mid y_d) = L_a(\bar{z}_1 \mid y_d)L(a \mid y_d)$ where $L_a(\bar{z}_1 \mid y_d) = aL(z_1 = a\bar{z}_1 \mid y_d)$. It follows that

$E_p(\overline{Z}_1)$ & $V_p(\overline{Z}_1)$ can be obtained by double expectation rules as for Z_1. We can then say $L(z_1|y_d)$ is variance consistent if $V_p(Z_1) \underset{n\to\infty}{\overset{P}{\to}} V_\theta(Z_1)$ and $V_p(\overline{Z}_1) \underset{m\to\infty}{\to} V_l(\mu)$. Similarly, $L(z_1|y_d)$ is mean consistent if $E_p(Z_1) \underset{n\to\infty}{\overset{P}{\to}} E_\theta(Z_1)$ and $E_p(\overline{Z}_1) \underset{m\to\infty}{\to} E_l(\mu)$.

The above considerations lead to the following consistency definitions

Definition 3. $L(z_1, z_2 \mid y_d)$ is variance consistent if the following conditions hold.

(i) As $n \to \infty$: $V_p(Z_2)/B_n^2 \overset{P}{\to} 0$, $V_p(Z_1) \overset{P}{\to} V_\theta(Z_1)$ and $\operatorname{cov}_p(Z_1, Z_2)/B_n \overset{P}{\to} 0$.

(ii) As $m \to \infty$: $V_p(Z_1/A_m) \to V_l(\mu)$ and $\operatorname{cov}_p(Z_1/A_m, Z_2) \to B_n V_l(\mu)$.

Definition 4. $L(z_1, z_2 \mid y)$ is mean consistent if the following conditions hold.

(iii) As $n \to \infty$: $E_p(Z_2)/B_n \overset{P}{\to} \mu$, $E_p(Z_1) \overset{P}{\to} E_\theta(Z_1)$.

(iv) As $m \to \infty$: $E_p(Z_1/A_m) \to E_l(\mu)$.

It is readily seen that L_c in Example 3 is mean and variance consistent.

The final example deals with a pure prediction problem.

Example 4

We want to predict the total number of fatalities from car accidents in a certain area for the next m time periods. The data y are observed values of $Y = (K_i, X_i), i = 1,...,n$ where K_i is the number of accidents in time period i, and X_i is the number of fatalities from d_i accidents in period i. It is assumed that all K_i, X_j are independent, and X_i ~Po($d_i\mu$), K_i ~Po(λ) and d_i is known.. It is assumed that $\lambda >> d_i$. Then A_m is the total number of accidents in the next m time periods, with A_m -1 assumed to be Poisson distributed with mean $m\lambda$. $X_i^{'}$ is the number of fatalities in the i'the accident and Poisson distributed with mean μ. During the data period there are accidents with missing data $X_j^{''}$ on the number of fatalities. We assume MCAR such that $X_j^{''} \sim$ Po(μ). B_n is then the total number of accidents in the data period with missingness on fatalities, such that $B_n = K_n - D_n$ with $K_n = \sum_{i=1}^{n} K_i$ and $D_n = \sum_{i=1}^{n} d_i$, the total number of accidents in the data period.

Let $S_n = \sum_{i=1}^{n} X_i$. Then the maximum likelihood estimates are $\hat{\mu} = S_n / D_n$, $\hat{\lambda} = K_n / n$. Here, the parametric likelihood $lik(\mu,\lambda|y)$ factorizes, so that the marginal likelihood for μ is unique and is given by a gamma-distribution with $E_l(\mu) = \hat{\mu} + D_n^{-1}$, $V_l(\mu) = (s_n + 1)/(D_n)^2$. It follows that a predictive likelihood is variance consistent if

as $n \to \infty$:

$$V_p(Z_2)/(K_n - D_n)^2 \xrightarrow{P} 0,\ V_p(Z_1) \xrightarrow{P} \mu(\mathrm{m}\lambda + 1) + \mathrm{m}\lambda\mu^2 \text{ and } \mathrm{cov}_p(Z_1, Z_2)/(K_n - D_n) \xrightarrow{P} 0.$$

as $m \to \infty$:

$$V_p(Z_1 / A_m) \to (s_n + 1)/(D_n)^2 \text{ and } \mathrm{cov}_p(Z_1 / A_m, Z_2) \to (K_n - D_n)(s_n + 1)/(D_n)^2.$$

Mean consistency requires:

as $n \to \infty$: $E_p(Z_2)/(K_n - D_n) \xrightarrow{P} \mu,\ E_p(Z_1) \xrightarrow{P} \mu(m\lambda + 1)$.

as $m \to \infty$: $E_p(Z_1 / A_m) \to \hat{\mu} + D_n^{-1}$.

We shall derive L from (3) using L_c for each term. Then $L_c(a|y)$ is such that A-1 is $NB(k+1, m/(m+n))$ implying that $E_p(A)$ = 1+ $m(\hat{\lambda} + \frac{1}{n})$ and $V_p(A)$ = $(\hat{\lambda} + \frac{1}{n})m(m+n)/n$. In order to describe $L_a(z_1,z_2|y)$ we need to briefly describe the negative multinomial distribution $NM(n;p_1,\ldots p_k)$, $\sum p_i \le 1$. $W = (W_1,\ldots,W_k) \sim NM(n; p_1,\ldots,p_k)$ if

$$f(w) = \frac{(\sum w_i + n - 1)!}{\prod w_i!(n-1)!} p_1^{w_1} \cdots p_k^{w_k} p_{k+1}^{n},\ p_{k+1} = 1 - \sum_{i=1}^{k} p_i.$$

Each W_i is $NB(n, p_i /(p_i + p_{k+1}))$, $\text{cov}(W_i, W_j) = np_i p_j / p_{k+1}^2$ and $\sum W_i \sim NB(n; \sum_{i=1}^k p_i)$.

We find that $L_{a,c}(Z_1, Z_2 \mid y)$ is $NM(s+1; p_1, p_2)$ where $s = \sum_{i=1}^n x_i$, $p_1 = a/(K_n + a)$, $p_2 =$ $(K_n - D_n)/(K_n + a)$. One can now easily find $E_p(Z)$ and $V_p(Z)$, and it is readily shown that the predictive likelihood is mean and variance consistent.

References

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihoodestimator, Biometrika, 70, 343-365.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one. In "Foundations of Statistical Inference" (editors: V.P. Godambe and D. A. Sprott), 203-242. Toronto: Holt, Rinehart & Winston.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). Journal of the American Statistical Association, 57, 269-306.

Bjørnstad, J. F. (1990). Predictive likelihood: A review (with discussion). Statistical Science, 5, 242-265.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. Journal of the American Statistical Association, 91, 791- 806.

Bjørnstad, J. F. (1998). Predictive likelihood. In "Encyclopedia of Statistical Sciences Update Volume 2" (editors S. Kotz, C.R. Read and D. L. Banks), 539-545. New York: Springer.

Bjørnstad, J. F. and Ytterstad, E. (2008). Two-stage sampling from a prediction point of view when the cluster sizes are unknown. Biometrika, 95, 187-204.

Bolfarine, H. and Zacks, S. (1992). Prediction Theory for Finite Populations. New York: Springer.
Butler, R.W. (1986). Predictive likelihood inference with applications (with discussion).
Journal of the Royal Statistical Society, Series B, 48, 1-38.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. Journal of the Royal Statistical Society, Series B, 17, 269-278.

Godambe, V.P. (1966). A new approach to sampling from finite populations, I, II. Journal of the Royal Statistical Society, Series B, 28, 310-328.

Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling from finite populations I. Annals of Mathematical Statistics, 36, 1707-1722.

Hinkley, D.V. (1979). Predictive likelihood. Annals of Statistics, 7, 718-728. Correction (1980), 8, 694.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Applications of likelihood methods to models involving large numbers of parameters (with discussion). Journal of the Royal Statistical Society, Series B, 32, 175-208.

Mathiasen, P.E. (1979). Prediction functions. Scandinavian Journal of Statistics, 6, 1-21.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 97, 558-625.

Thomsen, I. (1981). The use of Markov chain models in sampling from finite populations. Scandinavian Journal of Statistics, 8, 1-9.