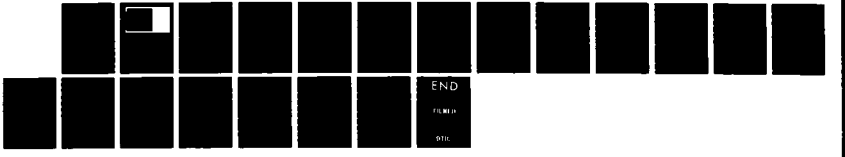
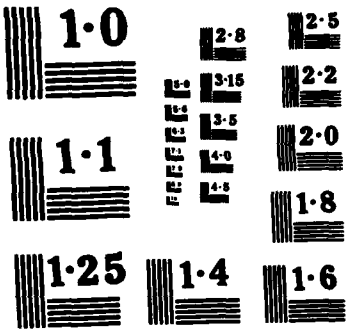
AD-A158 202 CONDITIONING IN A MISSING DATA PROBLEM(U) WISCONSIN 1/1
UNIV-MADISON MATHEMATICS RESEARCH CENTER C J SKINNER
JUL 85 MRC-TSR-2836 DAAG29-80-C-0041

UNCLASSIFIED F/G 12/1 NL





(2)

AD-A158 202

MRC Technical Summary Report #2836

CONDITIONING IN A MISSING DATA PROBLEM

C. J. Skinner

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

July 1985

(Received May 28, 1985)

Approved for public release
Distribution unlimited

Sponsored by

U.S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

CONDITIONING IN A MISSING DATA PROBLEM

C. J. Skinner

Technical Summary Report #2836

July 1985

ABSTRACT

Observations are recorded on variables x and y but a mechanism, which may depend on the observed x values, causes some of the y values to be missing. For three parametric examples, exact or approximate ancillary statistics are constructed. Conditioning on these ancillaries enables the missing data mechanism to be ignored under certain conditions. A correspondence is shown between these conditional procedures and the use of the observed information matrix in measuring the dispersion of the maximum likelihood estimator.

AMS (MOS) Subject Classifications: 62A20, 62D05, 62F25.

Key Words: affine ancillary; ancillary statistic; conditional inference; curved exponential family; ignorability; information; missing data; survey sampling.

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. DMS-8210950, Mod. 1.

SIGNIFICANCE AND EXPLANATION

Many statistical problems can be viewed as missing data problems. Data may be missing for practical reasons, out of the control of the data-collector, or missing by design, such as in sample surveys where, for some variables, data is available for the whole population but, for other variables, data is recorded for the sample and is 'missing' for the remainder of the population.

Rubin (1976) has shown that the mechanism which causes the data to be missing (the sampling design in the survey context) can, for a wide class of situations, be ignored for Bayes or Likelihood inference but not for classical sampling distribution theory inference. This non-ignorability of the missing data mechanism can make classical inference much more complicated and even impossible if the mechanism is unknown.

In this paper we apply ideas of Barndorff-Nielsen (1980), for a class of statistical models called curved exponential families, to construct ancillary statistics for some missing data problems. We show that if classical inference is carried out conditional upon the observed values of these ancillary statistics then the missing data mechanism may be ignored, in certain situations. Some correspondence between these conditional procedures and likelihood methods is established. The approach rests heavily on examples. This is characteristic of the conditioning literature, where specific results are often more enlightening than attempts at generality.

CONDITIONING IN A MISSING DATA PROBLEM

C. J. Skinner

1. INTRODUCTION

In this article we consider the possibility of conditioning, using either exact or approximate ancillary statistics, in a problem of estimation with missing data. The approach is not general but illustrative, using three examples.

Rubin (1976) showed that the mechanism which causes data to be missing may, in a large class of situations, be ignored for Bayesian or Likelihood inference but not for sampling distribution inference. We shall show how conditioning can enable this mechanism to be ignored for a wider class of situations under sampling distribution inference. Essentially we show, as in Efron and Hinkley (1978), that the conditional distribution of the maximum likelihood estimator (MLE) given an approprite ancillary corresponds, at least approximately, to the use of the inverse of the observed Fisher information matrix to measure the dispersion of the MLE. The latter procedure, being purely likelihood based, shares the ignorability properties of Likelihood inference. To provide initial motivation for the form of the conditioning procedure we compare our estimation problem with a prediction problem in survey sampling.

We assume the pairs (y_i, x_i), $i = 1, \ldots, N$ form a random sample from a bivariate distribution $p(y, x; \psi)$ belonging to a family indexed by the (generally vector) parameter ψ. We do not observe the complete data $d_C = (y_1, \ldots, y_N, x_1, \ldots, x_N)$ but only the incomplete data $d_I = (y_{i_1}, \ldots, y_{i_n}, s, x_1, \ldots, x_N)$ where $s = \{i_1, \ldots, i_n\}$ is a subset of size n from $U = \{1, \ldots, N\}$, $n < N$. We assume that s, and hence d_I, is obtained from d_C by a selection mechanism which assigns probabilities $p(s|x_C)$, possibly dependent on $x_C = (x_1, \ldots, x_N)$ but not on ψ, to selecting each of the $\binom{N}{n}$ possible subsets s.

The above set-up occurs, for example, in survey sampling where a sample s is selected from a finite population U, an auxiliary variable x is known for each unit

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. DMS-8210950, Mod. 1.

i in U and is available for use in the selection of s and where y is a variable measured in the survey. The mechanism $p(s|x_C)$ is called the sample design in this context.

Consider two problems:

(A) the prediction of a function of $(y_1,\ldots,y_N)$, specifically $\bar{y} = N^{-1}\sum^{N} y_i$,

(B) the estimation of a parameter of the marginal distribution of y, specifically $\mu_y = E(y)$.

In the sample survey context, (A) is the traditional descriptive use, usually concerned with means and totals, whilst (B) is the analytical use, usually concerned with the estimation of underlying models such as regression models (see e.g. Hartley and Sielken, 1975). Our consideration of only $\bar{y}$ and μ_y may be taken as special cases of this more general use.

Under the sampling distribution approach to inference, it is usual and natural in (A) to make predictive inference about $\bar{y}$ conditional on (s,x_C). Prediction proceeds as in the usual regression context where (x_i,y_i), $i \in s$ are known and x_i, $i \notin s$ are new x values for which we wish to predict y. More formally, note that the conditional distribution of d_C given d_I depends only on the parameter φ which indexes the conditonal distribution of y given x. Suppose we may write $\psi = (\varphi,\lambda)$, where $p(y,x;\psi) = p(y|x;\varphi)p(x;\lambda)$, so that the distribution of the data is

$$p(d_I;\psi) = p(y_{i_1},\ldots,y_{i_n} \mid s,x_C;\varphi)p(s|x_C)p(x_C;\lambda)$$

$$= \Big[\prod_{i\in s} p(y_i|x_i;\varphi)\Big]p(s|x_C)\prod_{i=1}^{N} p(x_i;\lambda) \quad . \tag{1}$$

Then, provided φ and λ are variation independent, it may be argued (Cox and Hinkley, 1974, p. 35; Barndorff-Nielsen, 1978, p. 50) that (s,x_C) is (extended-;S-) ancillary for φ, treating λ as a nuisance parameter, and that inference about φ and hence the prediction of $\bar{y}$ should be made conditional on (s,x_C). (More precisely we condition on a minimal sufficient reduction of (s,x_C), see Section 4.)

For the estimation of μ_y in (B), however, the same argument does not apply. Although (s,x_C) may be ancillary for φ, the parameter of interest μ_y is in general a function not only of φ but also of λ and will not be identifiable in the conditional distribution of d_I given (s,x_C). Hence inference about μ_y conditional on (s,x_C) is generally inappropriate.

Now if N is large, and in the sample survey context N may be very large, the difference between $\bar{y}$ and μ_y may be very small, even though it would appear from the discussion above that sampling distribution inference about these two quantities could proceed quite differently. This apparent 'paradox' provides one source of motivation for seeking an alternative procedure for conditional inference about μ_y.

In Bayesian or Likelihood inference about μ_y the selection mechanism only enters as a multiplicative factor free of ψ in the likelihood in (1) and hence is ignorable. In other words, making the false assumption that s is derived from U by simple random sampling rather than by the true mechanism $p(s|x_C)$, would make no difference to the inference for given d_I. In Rubin's (1976) terminology this is because we have assumed the data is 'missing at random'. The selection mechanism is not, however, ignorable for unconditional sampling distribution inference (nor for inference conditional on s but not on x_C as in Rubin, 1976).

In Section 2 we present the three examples. They are chosen to be of increasing order of complexity. In Example 1 there is an exact ancillary and the conditional distribution for inference about μ_y is exactly normal. In Example 2 there is an exact ancillary but the conditional distribution is only asymptotically normal. In Example 3 there is only an approximate ancillary.

In Section 3 we consider the prediction of $\bar{y}$ conditional on (s,x_C). In Section 4 we consider the estimation of μ_y on the assumption that s is a simple random sample from U. We adopt the approach of Barndorff-Nielsen (1980) to determine appropriate conditioning procedures. These are then compared with Section 3. In Section 5 we compare the conditional procedures with the use of asymptotic maximum likelihood theory using dispersion estimates based on either observed or expected Fisher information matrices. In

Section 6 we consider conditioning under a general selection mechanism $p(s|x_C)$ and discuss in what sense this mechanism is ignorable.

In asymptotic arguments we shall assume that n indexes fixed sequences $N = N_n$ and $p(s|x_C) = p_n(s|x_C)$ and that $n/N_n \to f$, a constant, as $n \to \infty$. Notation will be of three types. The key quantities ψ, φ, λ, μ_y, t, a are defined generally but take different forms in the different examples. The remaining statistics such as $\bar{x}_s$, $\bar{x}$, are defined in terms of d_I and are invariant with respect to the example. The remaining parameters, such as α, β and γ, are example-specific.

2. THE EXAMPLES

Example 1

We assume (y,x) is bivariate normal. The parameter vector is $\psi = (\varphi,\lambda)$ where $\varphi = (\alpha,\beta,\sigma^2_{y\cdot x})$, $\lambda = (\mu_x,\sigma^2_x)$ and $y|x \sim N(\alpha + \beta x, \sigma^2_{y\cdot x})$, $x \sim N(\mu_x,\sigma^2_x)$.

The parameter of interest is $\mu_y = \alpha + \beta\mu_x$. The MLE of ψ is (Anderson, 1957)

$$\hat{\psi} = (\bar{y}_s - b_s\bar{x}_s,\ b_s,\ n^{-1}SS_{y\cdot xs},\ \bar{x},\ N^{-1}SS_x)$$

where
$$\bar{y}_s = n^{-1}\sum_s y_i,\quad \bar{x}_s = n^{-1}\sum_s x_i,\quad \bar{x} = N^{-1}\sum^N x_i\ ,$$

$$b_s = \sum_s y_i(x_i - \bar{x}_s)/SS_{xs},\quad SS_{xs} = \sum_s (x_i - \bar{x}_s)^2\ ,$$

$$SS_{y\cdot xs} = \sum_s (y_i - \hat{\alpha} - \hat{\beta}\, x_i)^2,\quad SS_x = \sum^N (x_i - \bar{x})^2\ .$$

Hence the MLE of μ_y is $\hat{\mu}_y = \hat{\alpha} + \hat{\beta}\,\bar{x} = \bar{y}_s + b_s(\bar{x} - \bar{x}_s)$ which in survey sampling is termed the regression estimator (Cochran, 1977). A minimal sufficient statistic for ψ is $t = (\bar{y}_s,\bar{x}_s,b_s,SS_{xs},SS_{y\cdot xs},\bar{x},SS_x)$. Neither $\hat{\psi}$ nor t depend on the selection mechanism because $p(s|x_C)$ is a multiplicative factor free of ψ in (1). The family of distributions for d_I indexed by ψ is a curved exponential family, labelled (7,5) by Barndorff-Nielsen (1980) since $\dim(t) = 7$, $\dim(\psi) = 5$.

Example 2

We assume that

$$y|x \sim N(\gamma x, \sigma^2 x) \quad , \quad x \sim \text{Gamma}\ (\lambda,k) \quad , \quad k \text{ known}$$

(i. $p(x;\lambda) = \lambda^k x^{k-1} e^{-\lambda k}/\Gamma(k)$). The parameter vector is $\psi = (\varphi,\lambda)$ where $\varphi = (\gamma,\sigma^2)$. The parameter of interest is $\mu_y = \gamma k/\lambda$. The MLE of ψ is

$$\hat{\psi} = (\bar{y}_s/\bar{x}_s \ , \ n^{-1} \sum_s (y_i - \hat{\gamma}\, x_i)^2/x_i \ , \ k/\bar{x}) \ .$$

Hence the MLE of μ_y is $\hat{\mu}_y = \bar{y}_s\bar{x}/\bar{x}_s$, the ratio estimator (Cochran, 1977). A minimal sufficient statistics for ψ is $t = (\bar{y}_s,\bar{x}_s,\sum_s y_i^2/x_i,\bar{x})$. The family of distributions for d_I indexed by ψ is a (4,3) curved exponential family.

Example 3

We assume that y and x are both $0-1$ variables with

$$\Pr(y=1|x=0) = \varphi_0, \ \Pr(y=1|x=1) = \varphi_1, \ \Pr(x=1) = \lambda \ .$$

The parameter vector is $\psi = (\varphi,\lambda)$ where $\varphi = (\varphi_0,\varphi_1)$. The parameter of interest is $\mu_y = \Pr(y=1) = \lambda\varphi_1 + (1-\lambda)\varphi_0$. The MLE of ψ is $\hat{\psi} = (n_{10}/n_{.0}, \ n_{11}/n_{.1}, \ N_{.1}/N)$ where the cell counts in s and U are defined by

$$n_{\alpha\beta} = \sum_s y_i^{\alpha}(1-y_i)^{1-\alpha}\, x_i^{\beta}(1-x_i)^{1-\beta} \qquad \alpha,\beta = 0,1$$

$$N_{\alpha\beta} = \sum_1^N y_i^{\alpha}(1-y_i)^{1-\alpha}\, x_i^{\beta}(1-x_i)^{1-\beta}$$

and the margins are defined by

$$n_{.\beta} = \sum_{\alpha} n_{\alpha\beta}, \ N_{.\beta} = \sum_{\alpha} N_{\alpha\beta} \qquad \beta = 0,1 \ .$$

The MLE of μ_y is thus

$$\hat{\mu}_y = \frac{N_{.0}}{N} \cdot \frac{n_{10}}{n_{.0}} + \frac{N_{.1}}{N} \cdot \frac{n_{11}}{n_{.1}}$$

a poststratification estimator (Cochran, 1977). A minimal sufficient statistic for ψ is $t = (n_{00},n_{10},n_{01},N_{.0})$. This example also defines a (4,3) curved exponential family.

3. PREDICTION OF $\bar{y}$

A natural predictor of $\bar{y}$ given d_I is the regression predictor:

$$N^{-1}\Big[\sum_{i\in s} y_i + \sum_{i\notin s} E(y|x = x_i)\Big|_{\theta=\hat{\theta}}\Big] \quad .$$

This predictor can be shown to be identical to $\hat{\mu}_y$ in each of our three examples. Under conditions obeyed by the examples, this predictor is the minimum variance unbiased predictor of $\bar{y}$ conditional on (s,x_C) (Skinner, 1983). We now evaluate the conditional distribution of $\hat{\mu}_y - \bar{y}$ given (s,x_C) for each of the examples.

Example 1 (continued)

We obtain

$$\hat{\mu}_y - \bar{y}|s,x_C \sim N(0,\ (1 - n/N + na_1^2)\sigma^2_{y\cdot x}/n) \tag{2}$$

where $a_1 = (\bar{x}_s - \bar{x})SS_{xs}^{-1/2}$. Note that it is not necessary to condition on all the x_i values and s but only on a_1. An exact $(1-\alpha)$-level conditional confidence interval for $\bar{y}$ given a_1 is

$$\hat{\mu}_y \pm [(1 - n/N + na_1^2)SS_{y\cdot x}/n(n-2)]^{1/2} t_{n-2}(\alpha/2)$$

where $t_\nu(\alpha)$ is the α^{th} point of Student's t-distribution with ν d.f. Note that this interval does not depend on the selection mechanism. It distinguishes between 'good' samples where a_1 is small and hence $\bar{y}$ may be more precisely predicted and 'bad' samples where a_1 is large and hence $\bar{y}$ is more poorly predicted. Such a distinction is an essential aim of conditioning.

Example 2 (continued)

We obtain

$$\hat{\mu}_y - \bar{y}|s,x_C \sim N(0,\ [\bar{x}(N\bar{x} - n\bar{x}_s)/N\bar{x}_s]\sigma^2/n) \quad . \tag{3}$$

An exact $(1-\alpha)$-level confidence level for $\bar{y}$ conditional on $(\bar{x}_s,\bar{x})$ is

$$\hat{\mu}_y \pm \{[\bar{x}(N\bar{x} - n\bar{x}_s)/N\bar{x}_s]\hat{\sigma}^2/(n-1)\}^{1/2} t_{n-1}(\alpha/2) \quad .$$

In this example a 'good' sample is one in which y is missing for small x values so that $\bar{x}_s$ is large.

Example 3 (continued)

We may write

$$\hat{\mu}_y - \bar{y} = N^{-1}(N_{.0}/n_{.0} - 1)n_{10} + N^{-1}(N_{.1}/n_{.1} - 1)n_{11} - N^{-1}\tilde{n}_{10} - N^{-1}\tilde{n}_{11}$$

where $\tilde{n}_{\alpha\beta} = N_{\alpha\beta} - n_{\alpha\beta}$ $\alpha,\beta = 0,1$.

Conditional on (s,x_C) the quantities $N_{.0}$, $N_{.1}$, $n_{.0}$ and $n_{.1}$ are fixed and the distribution of $\hat{\mu}_y - \bar{y}$ is determined by the independent Binomial distributions of n_{10}, n_{11}, $\tilde{n}_{10}$ and $\tilde{n}_{11}$ which possess parameters $(n_{.0},\varphi_0)$, $(n_{.1},\varphi_1)$, $(N_{.0} - n_{.0},\varphi_0)$ and $(N_{.1} - n_{.1},\varphi_1)$ respectively. An exact conditional confidence interval for $\bar{y}$ appears intractable. Instead, suppose that $0 < \lambda, \varphi_0, \varphi_1 < 1$ and that the sequence of designs is such that $n_{.0}/n$ is almost surely bounded away from 0 and 1 as $n \to \infty$. Then almost surely

$$n^{1/2}(\hat{\mu}_y - \bar{y})|s,x_C \xrightarrow{L} N(0,\ (1 - fa_1)(1-\hat{\lambda})\varphi_0(1-\varphi_0)/a_1 + (1 - fa_2)\hat{\lambda}\,\varphi_1(1-\varphi_1)/a_2) \tag{4}$$

where $f = \lim(n/N)$, $a_1 = n_{.0}N/(nN_{.0})$, $a_2 = n_{.1}N/(nN_{.1})$.

A large sample conditional confidence interval may therefore be obtained substituting $\hat{\varphi}$ for φ.

4. ESTIMATION OF μ_y - MISSING VALUES SELECTED BY SIMPLE RANDOM SAMPLING

In general μ_y is not identified in the conditional distribution of d_I given (s,x_C). For example, the transformation $\mu_y \to \mu_y + \delta$, $\mu_x \to \mu_x + \delta\beta^{-1}$, φ fixed leaves $p(d_I|s,x_C;\psi)$ unaffected in Example 1. Hence conditioning on (s,x_C) as in Section 3 is inappropriate. Instead we seek what Barndorff-Nielsen (1980) term a 'conditionality resolution', that is we seek an exact or approximate ancillary statistic a such that $(\hat{\psi},a)$ is a one-to-one transformation of the minimal sufficient statistic t. The conditional distribution of $\hat{\psi}$ given a is then the appropriate one for inference. For a (k,d) curved exponential family we seek a $(k-d)$-dimensional ancillary. Whilst $\hat{\psi}$ and t are unaffected by the missing data mechanism, the ancillarity of any statistic a

may be affected. Hence, in this section we suppose that s is obtained from U by simple random sampling of fixed size so that $\{x_i; i=1,\ldots,N\}$ are IID whether or not $i \in s$. We consider the general mechanism in Section 5.

Example 1 (continued)

For this (7,5) curved exponential family we seek a two-dimensional ancillary. Since the normal family is a location-scale tranformation family the following function of t:

$$a = (a_1, a_2) = [(\bar{x}_s - \bar{x})/SS_{xs}^{1/2} \quad , \quad SS_{xs}/SS_x]$$

is an exact ancillary, in the sense that its distribution is free of ψ. Note that the tranformation t to $(\hat{\psi}, a)$ is one-one.

The distribution of $\hat{\psi}$ given a is in fact tractable although we shall only be concerned with the distribution of $\hat{\mu}_y$ given a. Corresponding to (2) we have

$$\hat{\mu}_y | s, x_C \sim N(\alpha + \beta\bar{x}, \ (1 + na_1^2)\sigma^2_{y\cdot x}/n) \quad . \tag{5}$$

But $\bar{x}$ is independent of a (since for example a is a function of $x_i - x_1$, $i = 2,\ldots,N$) and so

$$\bar{x} | a \sim N(\mu_x, \sigma_x^2/N) \quad . \tag{6}$$

Hence

$$\hat{\mu}_y | a \sim N(\mu_y, \ (1 + na_1^2)\sigma^2_{y\cdot x}/n + \beta^2\sigma_x^2/N) \quad . \tag{7}$$

This provides an appropriate conditional distribution for inference about μ_y. In fact we only need to condition on a_1 and from (2) this permits us to apply the same level of conditioning in the estimation of μ_y as in the prediction of $\bar{y}$. In the sense of Section 1 we have therefore resolved the 'paradox' of different levels of conditioning for the estimation and prediction problems.

It only appears possible, however, to obtain a large-sample (rather than exact) conditional interval for μ_y, since no pivot based on $\hat{\mu}_y - \mu_y$ seems to exist.

Example 2 (continued)

For this (4,3) curved exponential family we seek a one-dimensional ancillary. Since λ is a scale parameter, an exact ancillary is given by $a = \bar{x}/\bar{x}_g$. The transformation $t \to (\hat{\psi}, a)$ is one-one. It may be shown that a is independent of $\bar{x}$. Hence the exact

distribution of $\hat{\mu}_y$ given a is obtained by integrating $\bar{x}$ out of the joint distribution of $(\hat{\mu}_y, \bar{x})$ defined by

$$\hat{\mu}_y | \bar{x}, a \sim N(\gamma\bar{x}, \bar{x}a\sigma^2/n) \tag{8}$$

$$\bar{x} | a \sim \text{Gamma } (N\lambda, Nk) \quad . \tag{9}$$

As $n \to \infty$

$$n^{1/2}(\hat{\mu}_y - \mu_y) | a \xrightarrow{L} N(0, ka\sigma^2/\lambda + k\gamma^2 f/\lambda^2) \quad . \tag{10}$$

The 'paradox' of different levels of conditioning for $\bar{y}$ and μ_y is only partially resolved here, since in (3) inference about $\bar{y}$ is made conditional not only on a but also on $\bar{x}$, whereas from (8) inference about μ_y given both a and $\bar{x}$ is not possible.

Example 3 (continued)

We again have a (4,3) curved exponential family and seek a 1-dimensional ancillary. In this example no exact ancillary, which is a function of the minimal sufficient statistic, appears to exist. Barndorff-Nielsen (1980) discusses the construction of an approximate ancillary, termed the affine ancillary, for a general curved exponential family. In this 1-dimensional case the affine ancillary is

$$a_{aff} = k(\hat{\psi})[t - \tau(\hat{\psi})]\tau_\perp^T(\hat{\psi}) \tag{11}$$

where $\tau(\psi) = E(t)$, $\tau_\perp(\psi)$ is a 1×4 vector which is orthogonal to the rows of the 3×4 matrix $\partial\tau(\psi)/\partial\psi$ and $k(\psi)$ is a scalar chosen such that

$$\text{var}\{k(\psi)[t - \tau(\psi)]\tau_\perp^T(\psi)\} = 1 \quad .$$

In our example $\tau_\perp(\psi) = (1 \quad 1 \quad 0 \quad -n/N)$, up to a scalar multiple, which implies that

$$a_{aff} = k(\hat{\lambda})(n_{\cdot 0}/n - N_{\cdot 0}/N)$$

where

$$k(\lambda) = [(N-n)\lambda(1-\lambda)/nN]^{-1/2} \quad .$$

Now let

$$T_1 = k(\lambda)(1-\lambda)(a_1 - 1) \quad , \quad T_2 = k(\lambda)\lambda(1-a_2) \quad ,$$

$$a_1 = n_{\cdot 0}N/(nN_{\cdot 0}) \quad , \quad a_2 = n_{\cdot 1}N/(nN_{\cdot 1}) \quad .$$

Then T_1, T_2 and a_{aff} all converge to the same (standard normal) random variable as $n \to \infty$. Hence in the limit as $n \to \infty$, conditioning on a_{aff} is equivalent to the first

order to conditioning on (a_1,a_2). Also in the limit as $n \to \infty$, $N_{.0}/N$ is independent of a_{aff}. Thus

$$n^{1/2}(N_{.0}/N - \lambda)|a_1,a_2 \xrightarrow{L} N[0,\ f\lambda(1-\lambda)] \tag{12}$$

and as in (4), almost surely

$$n^{1/2}[\hat{\mu}_y - (N_{.0}\varphi_0 + N_{.1}\varphi_1)/N]|a_1,a_2,N_{.0}/N \xrightarrow{L}$$
$$N[0,\ \varphi_0(1-\varphi_0)N_{.0}/Na_1 + \varphi_1(1-\varphi_1)(1-N_{.0}/N)/a_1] \ . \tag{13}$$

Hence, almost surely

$$n^{1/2}(\hat{\mu}_y-\mu_y)|a_{aff} \xrightarrow{L} N[0,\ (1-\lambda)\varphi_0(1-\varphi_0)/a_1 + \lambda\varphi_1(1-\varphi_1)/a_2$$
$$+ (\varphi_0-\varphi_1)^2\lambda(1-\lambda)f] \ . \tag{14}$$

Note that the level of conditioning is again less than for the prediction of $\bar{y}$, since in (4) the conditioning is not only on a_1 and a_2 but also on $N_{.0}/N = 1-\hat{\lambda}$.

5. OBSERVED VERSUS EXPECTED FISHER INFORMATION

The asymptotic theory of maximum likelihood estimation for the regular IID case may be extended to the incomplete data structure of d_I (c.f. Hocking and Smith, 1968). According to this theory, two estimates of the (asymptotic) covariance matrix of $\hat{\psi}$ are $j(\hat{\psi})^{-1}$ and $i(\hat{\psi})^{-1}$ where $j(\hat{\psi})$ and $i(\hat{\psi})$ are the observed and expected Fisher information matrices respectively:

$$i(\psi) = E[j(\psi)] \ , \quad j(\psi) = -\partial^2 \log p(d_I;\psi)/\partial\psi\partial\psi^T \ .$$

Efron and Hinkley (1978) show that $j(\hat{\psi})^{-1}$ is often a good approximation to the conditional variance of $\hat{\psi}$ given an appropriate ancillary, a. Barndorff-Nielsen (1980) considers the extension to the multi-parameter case. We now compare $\text{var}(\hat{\mu}_y|a)$, as obtained from Section 4, with the estimates derived from $i(\hat{\psi})$ and $j(\hat{\psi})$, which are

$$v_{obs}(\hat{\mu}_y) = g'(\hat{\psi})^T j(\hat{\psi})^{-1} g'(\hat{\psi})$$
$$v_{exp}(\hat{\mu}_y) = g'(\hat{\psi})^T i(\hat{\psi})^{-1} g'(\hat{\psi})$$
$$\text{where } \mu_y = g(\psi) \ , \quad g'(\psi) = \partial g(\psi)/\partial\psi \ .$$

Example 1 (continued)

We obtain

$$v_{obs}(\hat{\mu}_y) = (1 + na_1^2)\hat{\sigma}^2_{y\cdot x}/n + \hat{\beta}^2\hat{\sigma}^2_x/N$$

$$v_{exp}(\hat{\mu}_y) = \hat{\sigma}^2_{y\cdot x}/n + \hat{\beta}^2\hat{\sigma}^2_x/N$$

and comparing with (7), v_{obs} is identical to $\text{var}(\hat{\mu}_y|a)$ evaluated at $\psi = \hat{\psi}$.

Example 2 (continued)

We obtain

$$v_{obs}(\hat{\mu}_y) = ka\,\hat{\sigma}^2/\hat{\lambda}n + k\hat{\gamma}^2/\hat{\lambda}^2N$$

$$v_{exp}(\hat{\mu}_y) = k\sigma^2/\hat{\lambda}n + k\hat{\gamma}^2/\hat{\lambda}^2N$$

and it follows immediately from (8) and (9) that v_{obs} is identical to $\text{var}(\hat{\mu}_y|a)$ evaluated at $\psi = \hat{\psi}$.

Example 3 (continued)

We obtain

$$v_{obs}(\hat{\mu}_y) = (1-\hat{\lambda})\hat{\varphi}_0(1-\hat{\varphi}_0)/na_1 + \hat{\lambda}\hat{\varphi}_1(1-\hat{\varphi}_1)/na_2 + (\hat{\varphi}_1-\hat{\varphi}_0)^2\hat{\lambda}(1-\hat{\lambda})/N$$

$$v_{exp}(\hat{\mu}_y) = (1-\hat{\lambda})\hat{\varphi}_0(1-\hat{\varphi}_0)/n + \hat{\lambda}\hat{\varphi}_1(1-\hat{\varphi}_1)/n + (\hat{\varphi}_1-\hat{\varphi}_0)^2\hat{\lambda}(1-\hat{\lambda})/N$$

and comparing with (14), nv_{obs} is identical to $\lim \text{var}[n^{1/2}(\hat{\mu}_y-\mu_y)|a_{aff}]$ with ψ and f replaced by $\hat{\psi}$ and n/N respectively.

6. ESTIMATION OF μ_y-GENERAL MISSING DATA MECHANISM

The mechanism $p(s|x_C)$ does not affect $\hat{\psi}$ or t but may affect the ancillarity of the statistics a discussed in Section 4. We shall first show that the exact ancillaries given for Example 1 and 2 remain exactly ancillary for a broad class of mechanisms and that the distribution $\hat{\mu}_y|a$ is also invariant. We then consider the more general problem.

Example 1 (continued)

Condition C1: the selection mechanism depends only on the configuration $z = [(x_3-x_1)/(x_2-x_1), (x_4-x_1)/(x_2-x_1),\ldots,(x_N-x_1)/(x_2-x_1)]$ of x_C.

In other words, under C1, $p(s|x_C)$ is invariant with respect to location or scale changes in x. This condition holds for a variety of sample designs, for example in stratified random sampling when strata are determined by quantiles of x_C and in truncated sampling where the point(s) of truncation are quantiles of x_C.

Under C1, a remains ancillary. For a is a function of z and s. The distributions $p(s|z) = p(s|x_C)$ and $p(z)$ are free of ψ. Hence the distribution of a is free of ψ. Furthermore $\bar{x}$ is independent of z and is conditionally independent of s given z. Hence $\bar{x}$ is independent of (s,z) and therefore of a. Thus (5) and (6) still apply and the distribution $\hat{\mu}_y|a$ is again given by (7).

Example 2 (continued)

Condition C2: the selection mechanism depends only on the ratios $w = (x_2/x_1, x_3/x_1,\ldots,x_N/x_1)$.

In other words, under C2, $p(s|x_C)$ is invariant with respect to scale changes of x. This condition holds, for example, with probability proportional to size designs, where x is a size measure, which are often used in conjunction with the ratio estimator.

Under C2, a remains ancillary by a similar argument to that in Example 1. Also it may be shown that $\bar{x}$ is independent of (s,w) and therefore of a and hence that the expressions (8), (9) and (10) remain valid.

We now turn to the general situation where the statistics a defined in Section 4 need no longer be ancillary, even approximately. We begin by attempting to construct ancillaries which may now depend on the mechanism $p(s|x_C)$. We could consider the affine ancillary but a slightly modified approach simplifies the distribution theory for $\hat{\mu}_y|a$. Let $\hat{\lambda}_s$ be the MLE of λ were only x_i, $i \in s$, to be observed, so for our three examples $\hat{\lambda}_s = (\bar{x}_s, n^{-1}SS_{xs})$, $k/\bar{x}_s$ and $n_{.1}/n$ respectively. Let $T(\hat{\lambda}) = E(\hat{\lambda}_s|\hat{\lambda})$. In each of our examples $\hat{\lambda}$ is sufficient for λ, and hence $T(\hat{\lambda})$, although critically dependent on

$p(s|x_C)$, is free of λ. We assume that the sequence of designs is such that $T(\cdot)$ is continuous at λ and that

$$n^{1/2}[\hat{\lambda}_s - T(\hat{\lambda})] \xrightarrow{L} N[0, k(\lambda)]$$

where $k(\lambda)$ is a finite positive-definite matrix. This seems a fairly weak condition although, for example, it would exclude the stratified design in Example 3 where $n_{\cdot 1}$ is set at a fixed fraction of $N_{\cdot 1}$ so that $k(\lambda) = 0$. We adopt as an approximate ancillary

$$a = n^{1/2}[\hat{\lambda}_s - T(\hat{\lambda})]k(\hat{\lambda})^{-1/2} \quad . \tag{15}$$

Note that, when the missing values are randomly selected, a reduces to the ancillary in Section 4 for Example 3 and is asymptotically equivalent to the ancillaries for Example 1 and 2. Since $\text{cov}(\hat{\lambda}_s - T(\hat{\lambda}), \hat{\lambda}) = 0$, $n^{1/2}(\hat{\lambda}-\lambda)$ will be asymptotically independent of a. Now the conditional distribution of $\hat{\mu}_y$ given (s,x_C) is unaffected by selection and in each of our examples we may write (almost surely)

$$n^{1/2}[\hat{\mu}_y - f_0(\hat{\lambda},\varphi)]|s,x_C \xrightarrow{L} N[0, f_1(\hat{\lambda},\hat{\lambda}_s,\varphi)] \tag{16}$$

where f_0 and f_1 are certain functions, continuous at $\hat{\lambda} = \lambda$, such that $f_0(\lambda,\varphi) = \mu_y$. Now define f_2 such that $f_1(\hat{\lambda},\hat{\lambda}_s,\varphi) = f_2(\hat{\lambda},a,\varphi)$. Then

$$n^{1/2}[\hat{\mu}_y - f_0(\hat{\lambda},\varphi)]|a,\hat{\lambda} \xrightarrow{L} N[0, f_2(\hat{\lambda},a,\varphi)]$$

$$n^{1/2}(\hat{\lambda}-\lambda)|a \xrightarrow{L} N[0, f_3(\lambda)]$$

so that $n^{1/2}(\hat{\mu}_y-\mu_y)|a \xrightarrow{L} N[0, f_2(\lambda,a,\varphi) + f_4(\psi)]$ where

$$f_4(\psi) = \lim_{n\to\infty} n \text{ var}[f_0(\hat{\lambda},\varphi)] \quad .$$

The estimated asymptotic variance of $n^{1/2}(\hat{\mu}_y-\mu_y)$ given a is thus

$$\hat{V} = f_2(\hat{\lambda},a,\hat{\varphi}) + f_4(\hat{\psi}) \quad .$$

But from (16) $f_4(\cdot)$ and $f_2(\hat{\lambda},a,\hat{\varphi}) = f_1(\hat{\lambda},\hat{\lambda}_s,\hat{\varphi})$ are unaffected by the missing data mechanism and hence this mechanism is ignorable for conditional inference. In other words, if we supposed incorrectly that the missing values were randomly selected so that x_i, $i \in s$ is distributed identically to x_i, $i \notin s$ we would obtain the same $\hat{V}$.

For a given mechanism the quantity $T(\hat{\lambda})$ in (15) and hence the ancillary a may be very complicated to compute. In practice, however, this is unnecessary. We showed in

Section 4 that $n^{-1}\hat{V}$ was equal to $v_{obs}(\hat{\mu}_y)$ (with n/N replaced by f) for our examples. Hence for inferential purposes we only require the straightforward computation of $\hat{\mu}_y$ and $v_{obs}(\hat{\mu}_y)$ which do not depend on $p(s|x_C)$.

Note in contrast that the estimation of the unconditional variance of $\sqrt{n}(\hat{\mu}_y - \mu_y)$ or the evaluation of $v_{exp}(\hat{\mu}_y)$ does depend on $p(s|x_C)$ and can be quite intractable. This provides a further practical advantage of conditioning.

7. CONCLUSION

We have indicated how exact or approximate conditioning arguments may be applied in three examples of a missing data problem. Conditioning is attractive here for several reasons: (1) it can permit the mechanism which causes the data to be missing to be ignored, (2) it can lead to more tractable procedures, (3) it makes inference more 'data-dependent' (Fisher's original motivation).

The results of this article are specific to the examples chosen although some possible generalization is suggested in Section 6 for models where t is a $1-1$ transformation of $(\hat{\psi},\hat{\lambda}_s)$. In this case the asymptotic maximum likelihood approach using the observed information matrix corresponds, under certain conditions, to conditioning on the ancillary defined in (15).

Acknowledgements

I am grateful to D. R. Cox for fundamental suggestions.

REFERENCES

ANDERSON, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Am. Statist. Assoc. 52, 200-3.

BARNDORFF-NIELSEN, O. (1978). Information and Exponential Families. Chichester: Wiley.

BARNDORFF-NIELSEN, O. (1980). Conditionality resolutions. Biometrika 67, 293-310.

COCHRAN, W. G. (1977). Sampling Techniques (3rd Edition). New York: Wiley.

COX, D. R. & HINKLEY, D. V. (1974). Theoretical Statistics. London: Chapman and Hall.

EFRON, B. & HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika 65, 457-82.

HARTLEY, H. O. & SIELKEN, R. L. (1975). A "super-population viewpoint" for finite population sampling. Biometrics 31, 411-22.

HOCKING, R. R. & SMITH, W. B. (1968). Estimation of parameters in the multivariate normal distribution with missing observations. J. Am. Statist. Assoc. 63, 159-173.

RUBIN, D. B. (1976). Inference and missing data. Biometrika 63, 581-92.

SKINNER, C. L. (1983). Multivariate prediction from selected samples. Biometrika 70, 289-92.

CJS/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2836	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Conditioning in a Missing Data Problem		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) C. J. Skinner		8. CONTRACT OR GRANT NUMBER(s) DMS-8210950, Mod. 1 DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53705		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS See Item 18 below		12. REPORT DATE July 1985 13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

affine ancillary; ancillary statistic; conditional inference; curved exponential family; ignorability; information; missing data; survey sampling

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Observations are recorded on variables x and y but a mechanism, which may depend on the observed x values, causes some of the y values to be missing. For three parametric examples, exact or approximate ancillary statistics are constructed. Conditioning on these ancillaries enables the missing data mechanism to be ignored under certain conditions. A correspondence is shown between these conditional procedures and the use of the observed information matrix in measuring the dispersion of the maximum likelihood estimator.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

9-85

DTIC