Statistica Sinica **9**(1999), 407-424

DENSITY ESTIMATION FROM COMPLEX SURVEYS

D. R. Bellhouse and J. E. Stafford

University of Western Ontario

Abstract: Three classes of kernel density estimates are proposed which are appropriate in the analysis of complex survey data. The three classes of estimates pertain to use of the whole data file, to use of binned data and to smoothing binned data. In each class a model-based asymptotic integrated mean square error is obtained under the complex sampling design. The parallel design-based asymptotic integrated mean square errors are obtained for the binned data and the smoothed binned data only. Quantile estimates from the smoothed binned data are proposed. The methodology is applied to data from the Ontario Health Survey of 1990.

Key words and phrases: Histograms, integrated mean square error, kernel density estimation, quantile estimation, smoothing.

1. Introduction

A population from which a measurement y can be taken has density function $f(y)$. The use of kernel density estimation to obtain $\hat{f}(y)$, an estimate of $f(y)$, is investigated in the context of data collection from sample surveys. We assume that $f(y)$, the density function of interest, is defined on one of two types of superpopulations. The first is an infinite superpopulation in which the observations are independent and identically distributed. The second is a superpopulation defined as a limiting sequence of finite populations. The purpose of this paper is two-fold. The major purpose is to examine the effect of the complex design on the asymptotic integrated mean square error of the density estimate. In particular, we show that there is no effect on the bias after customary probability weighting, but there is an effect, related to the design effect, on the variance. Further, the finite population approach allows us to examine properties of density estimates using correlated data in large samples. A second purpose is to show how this methodology may be applied to complex surveys.

As outlined in Jones (1989), three steps to kernel density estimation may be considered: prebinning, smoothing and postbinning. Prebinning refers to binning the data before smoothing. This may occur naturally in survey questionnaires when a nominally continuous random variable is coded in grouped form. Postbinning refers to binning after smoothing has been carried out. With the appropriate choice of the kernel to smooth the raw data, the usual histogram can be expressed as a postbinned estimate. The binning, pre or post, may be presented in piecewise constant, or in piecewise linear forms. Jones (1989) calls these discretized and interpolated forms respectively. The bar graph or histogram is an example of a piecewise constant presentation and the frequency polygon is an example of a piecewise linear presentation. We investigate kernel density estimation from complex surveys for three situations: the complete data file is used, a histogram of the data is obtained (postbinning) and the histogram is smoothed (prebinning).

Our results are finite population based, in particular a finite population of size N is considered with measurements denoted by $y_1, \ldots, y_N$. Inferences are based on n sample measurements $y_j, j \in s$, where s denotes the set of sample units. Although a single subscript j is used to denote the unit, the sampling design $P(s)$ may be complex in the sense that there may be a combination of stratification, several stages of sampling and clustering.

Previously the approach that has been used was through the distribution function. The starting point to this approach is the finite population empirical distribution function defined by

$$F_N(y) = \frac{\#y_j\text{'s} \leq y}{N}.$$

Given any sampling design, an unbiased or consistent estimator for $F_N(y)$ may be obtained by standard design-based techniques since $F_N(y)$ is a finite population proportion. This approach has been used, for example, by Chambers and Dunstan (1986) and later by Kuk (1988). Chambers and Dunstan (1986) also introduced covariates through a regression model and took a model-based approach to estimation of $F_N(y)$ using the prediction approach introduced by Brewer (1963) and Royall (1970). Various techniques for the use of covariates were further studied by Kuo (1988), Rao, Kovar and Mantel (1990), Chambers, Dorfman and Wehrly (1993), Dorfman (1993) and Dorfman and Hall (1993). In the estimation techniques presented here covariates are not used. Further, estimation of the density function is carried out directly through smoothing rather than a strict finite population approach, either model-based or design-based, through sampling estimation of the finite population distribution function.

The rationale for considering prebinned and postbinned estimates through kernel density estimation arises from the practicalities of survey methodology. Large samples can make subsequent computations intensive, for example in the estimation of percentiles with standard errors. Therefore it is not uncommon, even at the collection stage, for data to be grouped, or binned, in an effort to simplify coding and computation. Binning may result in a histogram which, depending on the size of the bins, does not resemble the original density estimate. Subsequent smoothing over these bins may recover some of the lost structure, while retaining the computational simplicity of a histogram.

The binning of continuous data provides a cell means approach to data analysis similar to, for example, Rao and Scott (1981) in contingency table analysis, Roberts, Rao and Kumar (1987) in logistic regression and Bellhouse and Rao (1994) in the analysis of domain means.

Two approaches may be taken for the estimation of $f(y)$; model-based and design-based. Associated with these approaches are two ways of handling the asymptotics. This has been noted, for example, by Fuller (1975). We denote these as Superpopulations 1 and 2. The first is used in the model-based approach and the second in the design-based approach.

Superpopulation 1. The N finite population units are a sample of independent and identically distributed units from some infinite superpopulation with density function $f(y)$.

Superpopulation 2. There is a nested sequence of finite populations of size N_l, $l = 1, 2, 3, \ldots$, such that $N_l \to \infty$ as $l \to \infty$ and $F_{N_l}(y) \to F(y)$ as $l \to \infty$, where $F(y)$ is a smooth function. For simplicity we write $F_N(y) \to F(y)$ as $N \to \infty$.

Under both the model-based and design-based approaches, the estimand is the hypothetical density function $f(y)$. Under Superpopulation 2 the density function of interest is $f(y) = \partial F(y)/\partial y$. This is in contrast to the approach of Chambers and Dunstan (1986), and others, where the estimand is the finite population empirical distribution function $F_N(y)$. We make the following assumption on $f(y)$:

Assumption 1.1. The superpopulation density $f(y)$ is a continuous function on the real line with finite third derivatives thus assuring that a Taylor series expansion to second order has a vanishing remainder.

In view of the two approaches the expectation operator E can be used in two ways. Under the model-based approach the operator E is a composite expectation $E_m E_p$, where E_p is the expectation with respect to the sampling design and E_m is the expectation with respect to the superpopulation model with density function $f(y)$. In the design-based approach $E = E_p$. Once the operator E_p is applied then the limit as $N \to \infty$ is taken on the sequence of finite populations. We will first obtain results under the model-based approach and then show the analogous design-based results.

The measure of variation of the estimate $\hat{f}(y)$ is the integrated mean square error or

$$\int E\{\hat{f}(y) - f(y)\}^2 dy. \tag{1.1}$$

See, for example, Scott ((1992), Section 2.3). Expression (1.1) may be rewritten as

$$IV\{\hat{f}(y)\} + ISB(\hat{f}). \tag{1.2}$$

Scott ((1992), p.131) has labeled the integrated variance as $IV\{\hat{f}(y)\} = \int \text{Var}\,\{\hat{f}(y)\}dy$ and the integrated square bias as $ISB(\hat{f}) = \int [E\{\hat{f}(y)\} - f(y)]^2 dy$. When (1.1) or (1.2) are taken to some order of approximation, then (1.1) is called *the asymptotic integrated mean square error* (AIMSE) and (1.2) becomes $AIMSE = AIV + AISB$. On taking the model-based approach of Superpopulation 1, a standard calculation gives

$$E_m E_p\{\hat{f}(y) - f(y)\}^2 = E_m \mathrm{Var}_p\{\hat{f}(y)\} + \mathrm{Var}_m E_p\{\hat{f}(y)\} + [E_m E_p\{\hat{f}(y)\} - f(y)]^2, \tag{1.3}$$

where Var_m and Var_p denote variance with respect to the model and design respectively. Note that in (1.3), $E_p\{(\hat{f}(y)\}$ is a finite population quantity. Cumulants of such quantities are generally $O(N^{-1})$ and hence negligible for large populations. Consequently, the measure of variation effectively simplifies to

$$\int E_m \mathrm{Var}_p\{\hat{f}(y)\}dy + \int [E_m E_p\{\hat{f}(y)\} - f(y)]^2 dy. \tag{1.4}$$

The expression in (1.4) corresponds to the integrated variance plus the integrated squared bias as in (1.2). Although we can develop results for both (1.3) and (1.4), it is generally the latter that is of interest. In the design-based approach (1.1) may be written as (1.2) with E replaced by E_p and Var replaced by Var_p.

In Section 2 we describe three kernel density estimators that are appropriate to complex surveys. The asymptotic integrated mean square errors of these estimators are derived in Section 3 and the effect of the complex sampling design on this measure is examined. Proofs of the results are given in the Appendix. Data from a large-scale survey with a complex design, the 1990 Ontario Health Survey, are used in Section 4 to illustrate the application of kernel density estimation to survey data. In the final section, ideas for future work, in particular quantile estimation from kernel density estimates, are discussed.

2. Discretization and Smoothing

In order to develop kernel density estimates in finite population sampling, it is useful first to obtain such an estimate based on the entire finite population as an estimate of a density defined on a larger superpopulation. Denote the standard kernel density estimate based on the entire population by $f_s(y)$. This is given by

$$f_s(y) = \frac{1}{Nh_S} \sum_{j=1}^{N} K_S(\frac{y - y_j}{h_S}), \tag{2.1}$$

where K_S denotes the choice of kernel with window width h_S. For a sample survey $f_S(y)$ is then estimated by

$$\hat{f}_S(y) = \frac{1}{h_S} \sum_{j \in S} w_{jS} K_S(\frac{y - y_j}{h_S}), \tag{2.2}$$

where w_{js} are the sample weights determined from the complex design.

Assume that the population can be discretized by dividing the range of y into bins defined by the boundaries $x_0, \ldots, x_k$ where $x_0 \leq \min\{y_j\}$ and $x_k \geq \max\{y_j\}$. The ith bin is denoted by $B_i = [x_{i-1}, x_i)$ with midpoint m_i and common length $b = x_i - x_{i-1}$. When this is done the kernel density estimate may be postbinned. For any point y we denote the midpoint of the bin containing y by $m(y)$. The postbinned estimate based on the finite population measurements is given by $f_{S}(m(y))/\delta$ with the sample version being $\hat{f}_{S}(m(y))/\hat{\delta}$. The density estimate needs to be scaled by $\delta = b\sum_{i=1}^{k} f_{S}(m_i)$, or the sample equivalent, in order to be a true density. For convenience we have assumed that x_0 and x_k are finite. As $b \to 0$ we can allow $x_0 \to -\infty$ and $x_k \to \infty$ if necessary.

The histogram estimate of the density function can now be obtained in one of two ways. Let the proportion of observations in the finite population falling into B_i be p_i. The sample estimate is $\hat{p}_i$. The histogram estimate of $f(y)$ is

$$f_{H}(y) = p_i/b \qquad \text{for } y \in B_i \tag{2.3}$$

for the finite population and

$$\hat{f}_{H}(y) = \hat{p}_i/b \text{ for } y \in B_i \tag{2.4}$$

for the sample. Note that the finite population asymptotics in Superpopulation 2 yield $f_{H}(y) \to f(y)$ as $N \to \infty$ and $b \to 0$. Alternately the histogram estimate may be obtained as special cases of $f_{S}(m(y))/\delta$ or $\hat{f}_{S}(m(y))/\hat{\delta}$ as appropriate. In (2.1) and (2.2) replace K_{S} by a naive kernel K_H which is the uniform density on the interval $(-1/2, 1/2)$ and set $h_{S} = b$. With these choices, $\delta = \hat{\delta} = 1$ and the histogram based on sampled data is

$$\hat{f}_{H}(y) = \frac{1}{b}\sum_{j\in S} w_{sj} K_{H}\Big(\frac{m(y) - y_j}{b}\Big). \tag{2.5}$$

The expression for $f_{H}(y)$ as a special case of $f_{S}(m(y))/\delta$ may be similarly obtained.

The effect of binning is to reduce the original data to a collection of evenly spaced midpoints and counts. Treating this as the only data available, a weighted kernel density estimate $\hat{f}_{B}(y)$ based on the sample data can be obtained as

$$\hat{f}_{B}(y) = \frac{1}{h_{B}}\sum_{i=1}^{k} \hat{p}_i K_{B}(\frac{y - m_i}{h_{B}}) \tag{2.6}$$

using (2.4), where K_{B} is the kernel and h_{B} the bandwidth. Equivalently

$$\hat{f}_{B}(y) = \frac{b}{h_{B}}\sum_{i=1}^{k} \hat{f}_{H}(m_i) K_{B}(\frac{y - m_i}{h_{B}}) \tag{2.7}$$

using (2.5). The expression for the finite population $f_B(y)$ is obtained on replacing $\hat{p}_i$ by p_i in (2.6) or on replacing $\hat{f}_H(m_i)$ by $f_H(m_i)$ in (2.7). In the same manner that a standard density estimate weights a sample point by its distance from y, $\hat{f}_B(y)$ weights the weight, $\hat{f}_H(m_i)$, for the ith bin by the distance this bin is from y.

The choice of the kernel K_T ($T = S$, H or B) is in the hands of the data analyst. We make the standard assumptions about the kernel K_T:

Assumption 2.1. K_T is a symmetric function with $\int K_T(t)dt = 1$, $\int tK_T(t)dt = 0$ and $0 < \int t^2K_T(t)dt < \infty$ and $\int K_T^2(t)dt < \infty$.

See, for example, Silverman (1986, p.38).

3. Determination of the Asymptotic Integrated Mean Square Error

Assuming independence and identical distributions for the observations, expressions for the asymptotic integrated mean square error (AIMSE) of $f_S(y)$, $f_H(y)$ and $f_B(y)$ are given by Scott (1992) and Jones (1989). In the current context the situation for $\hat{f}_S(y)$, $\hat{f}_H(y)$ and $\hat{f}_B(y)$ is complicated by the sampling structure and the possible lack of independence. We first obtain the asymptotic integrated square bias for each of $\hat{f}_S(y)$, $\hat{f}_H(y)$ and $\hat{f}_B(y)$ under the model-based and design-based frameworks (Lemmas 3.1 and 3.2). Then the asymptotic integrated variances are obtained, again under the model-based and design-based frameworks (Lemmas 3.3, 3.4 and 3.5). The resulting integrated mean square errors (Theorems 3.1 and 3.2) are obtained according to (1.2) and (1.4). For notational convenience in what follows, the subscript T refers collectively to S, H or B. Further, we set $R(\phi) = \int \phi(t)^2dt$ and $\sigma_T^2 = \int t^2K_T(t)dt$, and denote $\partial f(t)/\partial t$ by f' and $\partial^2 f(t)/\partial t^2$ by f''. For particular cases of $R(\phi)$ we make the following assumption:

Assumption 3.1. For the superpopulation density function $f(y)$, assume that $R(f) < \infty$, $R(f') < \infty$ and $R(f'') < \infty$.

We make the following assumption on the density estimators:

Assumption 3.2. The density estimator $\hat{f}_T(y)$ is asymptotically unbiased for $f_T(y)(T = S, H \text{ or } B)$ in the sense of Särndal, Swensson and Wretman ((1992), pp.166-167).

Under Assumption 3.2, when the weights w_{js} are chosen to give asymptotically unbiased estimates, we have as $n \to \infty$,

$$E_p(\hat{f}_T(y)) = f_T(y) \tag{3.1}$$

under Superpopulation 2, and

$$E_mE_p(\hat{f}_T(y)) = E_m(f_T(y)) \tag{3.2}$$

under Superpopulation 1. The expressions in (ii) and (iii) of Lemma 3.1 are given by Jones (1989) while (i) is given by Silverman (1986). Consequently, there is no effect of the complex sampling design on the bias.

Lemma 3.1. *Under Superpopulation* 1*, with the y_j's, $j = 1, \ldots, N$, having identical distributions, and under Assumptions* 1.1*,* 2.1*,* 3.1 *and* 3.2*,*
(i) $B(\hat{f}_S(y)) = \sigma_S^4 h_S^4 R(f'')/4 + O(h_S^6)$ *as* $h_S \to 0$*;*
(ii) $B(\hat{f}_H(y)) = b^2 R(f')/12 - b^4 R(f'')/360 + O(b^6)$ *as* $b \to 0$*;*
(iii) $B(\hat{f}_B(y)) = (\sigma_B^2 h_B^2 + b^2/12)^2 R(f'')/4 + O(h_B^6 + b^6)$ *as* $h_B \to 0$ *and* $b \to 0$.

Proof. See the Appendix.

Lemma 3.2. *Under Superpopulation* 2 *and under Assumptions* 1.1*,* 2.1*,* 3.1 *and* 3.2 *with $n/N \to \pi$ as $n, N \to \infty$, where π is a constant in the interval* $(0, 1)$*,*
(i) $B(\hat{f}_H(y)) = b^2 R(f')/12 - b^4 R(f'')/360 + O(b^6)$ *as* $b \to 0$ *and*
(ii) $B(\hat{f}_B(y)) = (\sigma_B^2 h_B^2 + b^2/12)^2 R(f'')/4 + O(h_B^6 + b^6)$ *as* $h_B \to 0$ *and* $b \to 0$.

Proof. See the Appendix.

Comment 3.1. Note that finite population asymptotics apply only to the situation in which a histogram has been used so that there is no result parallel to (i) in Lemma 3.1 for the standard kernel estimator. Also note that once again there is no effect of the complex sampling design on the bias.

In general terms each density estimate can be written as a weighted sample sum

$$\hat{f}_T(y) = \sum_{j \in S} w_{js} q_T(y_j), \tag{3.3}$$

where $q_T(y_j)$ is a function of the data y_j depending upon which of $T = S$, H or B is used. For example, from (2.2) we have that $q_S(y_j) = K_S((y - y_j)/h_S)/h_S$. We obtain the following for the design variance of $\hat{f}_T(y)$ averaged over the model in the first superpopulation.

Lemma 3.3. *Under Superpopulation* 1*,*

$$E_m V_p\{\hat{f}_T(y)\} = -\sum_{j \neq l = 1}^{N} w_{jl} E_m(q_T(y_j)^2\} + \sum_{j \neq l = 1}^{N} w_{jl} E_m\{q_T(y_j) q_T(y_l)\},$$

where $w_{jl} = E_p(w_{sj} - 1/N)(w_{sl} - 1/N)$*.*

Proof. See the Appendix.

We can use Lemma 3.3 to derive the asymptotic variances of each of the kernel density estimators under the first superpopulation. In particular, we have

Lemma 3.4. *Under Superpopulation* 1 *and under Assumptions* 1.1*,* 2.1*,* 3.1 *and* 3.2*, as* $n \to \infty$*,*

$$AIV(\hat{f}_S(y)) = \Big[\frac{R(K_S)}{nh_S} - \frac{R(f)}{n}\Big]\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big] + O(n^{-1}),$$

$$AIV(\hat{f}_H(y)) = \Big[\frac{1}{nb} - \frac{R(f)}{n}\Big]\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big] + O(n^{-1}), \text{ and}$$

$$AIV(\hat{f}_B(y)) = \Big[\frac{R(K_B)}{nh_B} - \frac{R(f)}{n}\Big]\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big] + O(n^{-1}).$$

Proof. See the Appendix.

Comment 3.2. Under a simple random sampling without replacement sampling design, the expression $-n\sum w_{jl}$ reduces to $[1-(n/N)]$, which is 1 for large N. In this case each of the asymptotic integrated variances reduces to the appropriate expressions in Silverman (1986) or Jones (1989). Consequently the expression $-n\sum w_{jl}$ may be interpreted as a design effect so that the effect of the complex design is to inflate the variance by the value of the design effect.

Comment 3.3. Under Superpopulation 1 it is assumed that the observations are independent and identically distributed. When the y's are not independent then the effect on the asymptotic integrated variances may be negligible. See the Appendix for further details.

In the design-based framework, denote the variance-covariance matrix of $(\hat{p}_1, \ldots, \hat{p}_k)$ by V and its estimate based on the complex design by $\hat{V}$. The (i,j)th element of V is v_{ij} and of $\hat{V}$ is $\hat{v}_{ij}$. The design effect for the ith bin is given by $d_i = nv_{ii}/[p_i(1-p_i)]$ so that the estimated design effect is $\hat{d}_i = n\hat{v}_{ii}[\hat{p}_i(1-\hat{p}_i)]$. The covariance design effect is given by $d_{ij} = -nv_{ij}/[p_ip_j]$. We let $\bar{d} = \sum_{i=1}^{k} d_i/k$ and $\bar{d}' = \sum_{i=1}^{k}\sum_{i=1}^{k} d_{ij}/k^2$, where $d_{ii} = d_i$. Using this notation we have

Lemma 3.5. *Under Superpopulation* 2 *and under Assumptions* 1.1*,* 2.1*,* 3.1 *and* 3.2*, then as* $n, N \to \infty$ *with* $n/N \to \pi$*, where* π *is a constant in the interval* $(0,1)$*,*

$$AIV(\hat{f}_H(y)) = \bar{d}\Big[\frac{1}{nb} - \frac{R(f)}{n}\Big] + O(n^{-1}), \text{ and}$$

$$AIV(\hat{f}_B(y)) = \bar{d}\frac{R(K_B)}{nh_B} - \bar{d}'\frac{R(f)}{n} - (\bar{d} - \bar{d}')\frac{bR(f)R(K_B)}{n} + O(n^{-1}).$$

Proof. See the Appendix.

Comment 3.5. Note that in the case of the smoothed histogram the effect of the complex design is not as straightforward as in all other cases, except in the special case when $\bar{d} = \bar{d}'$. Note also that the average design effect $\bar{d}$ is analogous to the use of $-n\sum w_{jl}$ in Lemma 3.4.

Comment 3.6. As in Scott (Section 3.2.1) the differentiability condition of $f(y)$, Assumption 1.1, can be changed to Lipshutz continuity of $f(y)$ within each bin for the derivation of $AIV(\hat{f}_{H}(y))$.

Dropping the higher order terms, for the first superpopulation we can summarize the results of Lemmas 3.1, 3.3 and 3.4 as

Theorem 3.1. *Under Superpopulation* 1*, and under Assumptions* 1.1*,* 2.1*,* 3.1 *and* 3.2*, then as $n \to \infty$ and $h_{S}, h_{B}, b \to 0$ as appropriate,*

$$
\begin{aligned}
AIMSE(\hat{f}_{S}(y)) &= \Big[\frac{R(K_{S})}{nh_{S}} - \frac{R(f)}{n}\Big]\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big] + \frac{1}{4}\sigma_{S}^{4}h_{S}^{4}R(f''),\\
AIMSE(\hat{f}_{H}(y)) &= \Big[\frac{1}{nb} - \frac{R(f)}{n}\Big]\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big] + \frac{b^2}{12}R(f') - \frac{b^4}{360}R(f''), \text{ and}\\
AIMSE(\hat{f}_{B}(y)) &= \Big[\frac{R(K_{B})}{nh_{B}} - \frac{R(f)}{n}\Big]\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big] + \frac{1}{4}(\sigma_{B}^{2}h_{B}^{2} + \frac{b^2}{12})^2R(f'').
\end{aligned}
$$

Again dropping the higher order terms, for the second superpopulation we can summarize the results of Lemmas 3.2 and 3.5 as

Theorem 3.2. *Under Superpopulation* 2 *and under Assumptions* 1.1*,* 2.1*,* 3.1 *and* 3.2*, then as $n, N \to \infty$ with $n/N \to \pi$, where π is a constant in the interval $(0,1)$, and $h_{B}, b \to 0$ as appropriate,*

$$
\begin{aligned}
AIMSE(\hat{f}_{H}(y)) &= \bar{d}\Big[\frac{1}{nb} - \frac{R(f)}{n}\Big] + \frac{b^2}{12}R(f') - \frac{b^4}{360}R(f''), \text{ and}\\
AIMSE(\hat{f}_{B}(y)) &= \bar{d}\frac{R(K_{B})}{nh_{B}} - \bar{d}'\frac{R(f)}{n} - (\bar{d}-\bar{d}')\frac{bR(f)R(K_{B})}{n}\\
&\quad + \frac{1}{4}(\sigma_{B}^{2}h_{B}^{2} + \frac{b^2}{12})^2R(f'').
\end{aligned}
$$

4. Example: The Ontario Health Survey

We illustrate the techniques of density estimation from large-scale surveys with data from a Canadian survey known as the Ontario Health Survey (see Ontario Ministry of Health (1992)). A stratified two-stage cluster sample of Ontarians was carried out in 1990 to measure the health status of the population and to collect data relating to the risk factors of major causes of morbidity and mortality in the Province of Ontario. The survey was designed to be compatible with the Canada Health Survey carried out in 1978-1979. A total sample size of 61,239 people was obtained from 43 public health units across the Province of Ontario. The public health unit was the basic stratum with an additional division of the health unit into rural and urban strata so that there were a total of 86 strata. The first stage units within a stratum were enumeration areas. These enumeration areas, taken from the 1986 Census of Canada, are the smallest geographical units from which census counts can be obtained automatically. An average of 46 enumeration areas was chosen within each stratum. Within an enumeration area dwellings were selected, approximately 15 from an urban enumeration area and 20 from a rural enumeration area. Information was collected on members of the household within the dwelling.

Several health characteristics were measured. We will focus on two continuous variables from the survey, Body Mass Index (BMI) and Desired Body Mass Index (DBMI). The BMI is a measure of weight status and is calculated from the weight in kilograms divided by the square of the height in metres. The DBMI is the same measure with actual weight replaced by desired weight. The index is not applicable to adolescents, adults over 65 years of age, pregnant or breastfeeding women. The measures vary between 7.0 and 45.0. A value of the BMI less than 20.0 is often associated with health problems such as eating disorders. An index value above 27.0 is associated with health problems such as hypertension and coronary heart disease. A total of 44,457 responses were obtained for the BMI and 41,939 for the DBMI. When the data were binned, the design effects for the proportion of observations falling in each bin were usually in the range of 2.0 to 3.9, except when the estimated proportion was small ($< 0.2\%$), in which case the design effect was less than 2.0. Had the sampling design been more heavily clustered, the probable effect would be an increase in the design effects so that there would be an increase in the integrated variance with the bias remaining the same.

For both the BMI and DBMI we constructed density estimates in *S-Plus* using both $\hat{f}_H(y)$ and $\hat{f}_B(y)$. The implementation of the estimation procedure was straightforward, the only complication being the determination of the appropriate window width h_B. Standard techniques involve differentiating the AIMSE and solving for h_B. This leads to a choice of h_B that is proportional to $1/\sqrt[5]{n}$. However, given the large sample size typically encountered in large-scale surveys, this criterion is clearly inappropriate as the resulting window width is extremely small.

Consequently, we prefer to use a criterion appearing in Jones (1989) that compares the AIMSE of a kernel density estimate involving binning to the ideal estimate where no binning occurs. This criterion, R_a, is reported in Jones (1989) for the prebinned kernel density estimates. The full developments given there are only briefly outlined here and are as follows. Clearly the window size and the bin size are related. For example, for a smoothed histogram larger bins would lead to larger window sizes. Hence Jones (1989) sets $b = ah$ and makes this substitution in the expression for AIMSE, denoting the result as $AIMSE_a$. He notes that $AIMSE_0$ is the expression for the AIMSE of the ideal kernel density estimate. Upon further substitution given in Jones (1989), the expression $R_a = AIMSE_a/AIMSE_0$ may be simplified to give (4.3) of Jones (1989). Pursuing the same strategy in our case leads to an expression equivalent to (4.3) in Jones (1989), but only if one ignores the term $R(f)/n$ in the AIMSE, as Jones has; otherwise R_a depends on the design effect. We conveniently adopt the former approach and hence values of a near 1.25 are reasonable (Jones (1989)).

The bin sizes and bin widths for both the BMI and DBMI measures were determined after an initial examination of the data. It was decided in both cases to use 30 bins. For the BMI the bin width was set at 1.22 with the lowest bin boundary at 8.9 and the highest boundary at 45.5. In this case h_B, the window width, was set at 0.976. The associated values for the DBMI were a bin width of 1.27, lower and upper boundaries of 7.0 and 45.1 respectively, and a window width of 1.016. Figures 1 and 2 show the histogram, frequency polygon and prebinned kernel density estimates for the BMI and DBMI datasets respectively. The height of the ith bar in the histogram is proportional to $\hat{p}_i$, the survey estimate of the proportion of population observations falling in bin i. From Figures 1 and 2 it is clear that the prebinned kernel density estimates has a more appealing appearance that the histogram or frequency polygon. However, in this case smoothing the histogram has tended to erode the peak of the histogram while the tails become fatter.

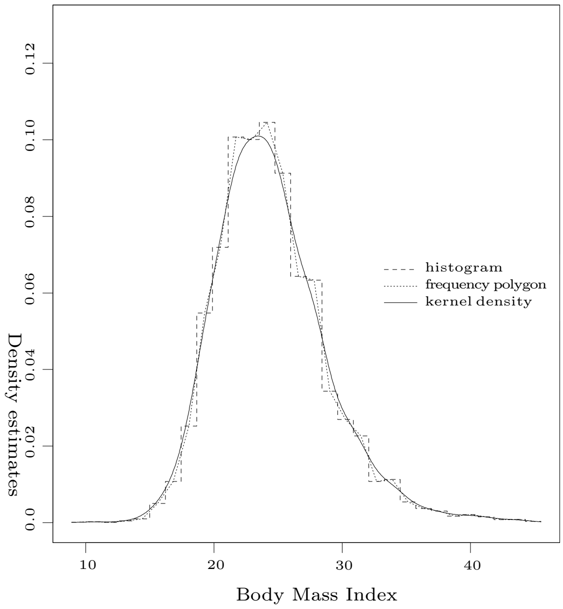

Figure 1. Density estimates for the Body Mass Index. There include a histogram, a frequency polygon and a kernel density estimate.

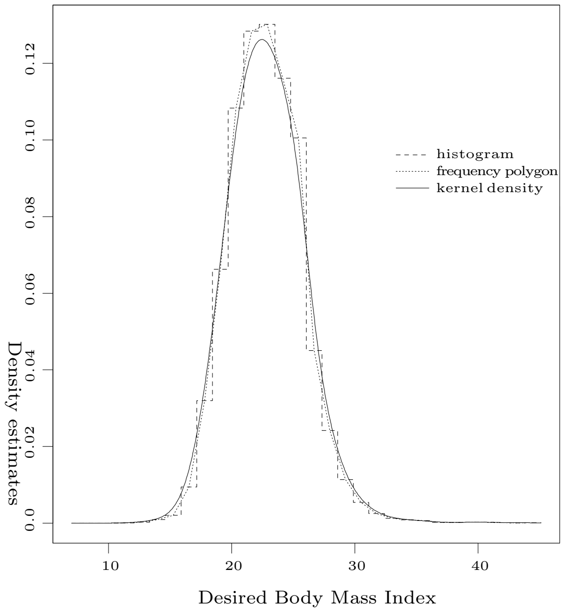

Figure 2. Density estimates for the Desired Body Mass Index. There include a histogram, a frequency polygon and a kernel density estimate.

We also report quantile estimates from $\hat{f}_H(y)$ and $\hat{f}_B(y)$ for both the BMI and DBMI datasets. These are given in Table 1. Quantile estimates from the histogram were obtained in the usual way through linear interpolation. Quantile estimates based on $\hat{f}_B(y)$ required solving the equation

$$g(\hat{q}_\alpha) = \int_{-\infty}^{\hat{q}_\alpha} \hat{f}_B(t)dt - \alpha = 0 \tag{4.1}$$

through the use of the following Newton-Raphson recipe.

1. Set the initial value $\hat{q}_0$ to $\hat{q}_H$, the quantile based on the histogram.
2. Compute the Newton-Raphson step as $\Delta_i = -g(\hat{q}_{i-1})/g'(\hat{q}_{i-1})$ where g' is simply $\hat{f}_B$.
3. Steps 1 and 2 were repeated until $\Delta_i < 10^{-4}$, giving the final estimate as $\hat{q}_\alpha = \hat{q}_0 + \sum_{j=1}^{i} \Delta_j$.

Implementation involves using the kernel density estimate to evaluate the denominator of Δ_i and the computing of tail areas for a gaussian kernel to compute the numerator. Both of these are readily available, making the implementation so convenient that no other method was considered. Convergence was generally obtained after no more than four iterations and often after only two. This is not surprising given that the quantiles of the histogram serve as excellent initial values, the density estimate is a nice smooth function and g is monotone. Bin and window widths were the same as those used for the estimation of $\hat{f}_H(y)$ and $\hat{f}_B(y)$. The quantile estimates given reflect the bias pattern noted for $\hat{f}_B(y)$. The question now remains whether this smoothed histogram used for quantile estimation provides any advantage over existing quantile estimation procedures. This is addressed below.

Table 1. Estimated quantile values from $\hat{f}_H(y)$ and $\hat{f}_B(y)$

	Quantile Value	Quantile Value	Quantile Value	Quantile Value
Quantile	$\hat{f}_H(y)$ - BMI	$\hat{f}_B(y)$ - BMI	$\hat{f}_H(y)$ - DBMI	$\hat{f}_B(y)$ - DBMI
0.05	18.56	18.26	18.21	17.92
0.15	20.31	20.21	19.78	19.59
0.25	21.53	21.46	20.70	20.63
0.35	22.52	22.51	21.52	21.49
0.45	23.52	23.50	22.30	22.30
0.55	24.48	24.51	23.07	23.09
0.65	25.53	25.61	23.88	23.92
0.75	26.90	26.94	24.74	24.82
0.85	28.52	28.73	25.73	25.92
0.95	32.17	32.42	27.76	27.96

5. Future Work

An immediate benefit of the availability of a histogram or kernel density estimate is the ability to estimate quantiles and this issue is briefly addressed. As mentioned earlier, if the complete data file is available, and often this is not the case, the histogram and prebinned kernel density estimate still have the advantage that quantile estimates can be more easily computed. In addition, quantiles based on the kernel density estimate are effectively smoothed versions of quantiles based on the histogram. The effect of smoothing is to reduce variance and increase bias. Whether this results in a decrease in the mean square error needs to be studied formally and is beyond the scope of the paper. However, we do present a small preliminary simulation study. The purpose of the study is to estimate the mean square error of the quantile estimates.

In the study a finite population of $N = 1,000$ was taken from a standard normal superpopulation. A simple random sample of $n = 100$ was obtained from the population. Quantiles were obtained from both $\hat{f}_H(y)$ and $\hat{f}_B(y)$. Given the stable nature of the Newton-Raphson algorithm described earlier, we felt comfortable with setting the number of iterations at two. For each quantile 100 sample estimates were obtained from which one estimate of the mean square error was obtained. This estimate was decomposed to assess the relative contributions of variance and squared bias. The entire process was repeated 100 times and the average variance, squared bias and mean square errors were calculated along with the standard error of simulation. The results are reported in Table 2. The whole simulation study was repeated using a chi-square superpopulation with five degrees of freedom. These results are reported in Table 3. The notable trend in Tables 2 and 3 is a drop in the mean square error for quantiles $\hat{q}_\alpha$ based on $\hat{f}_B(y)$. This is evidently due to a relatively larger drop in the variance of $\hat{q}_\alpha$ compared with a smaller increase in squared bias. This is, in part, not surprising since smoothing, as noted earlier, has just this effect. Histograms are binned kernel density estimates and the quantiles are obtained from a smoothed version of the histogram. This trend is simply inherited by $\hat{q}_\alpha$. However, the fact that variance reduction would dominate as it has can only be explained through further investigation, which is our intention in forthcoming work.

Table 2. Simulated mean square errors of quantile estimates using $\hat{f}_H(y)$ and $\hat{f}_B(y)$ for a Normal superpopulation

	$\hat{f}_H(y)$			$\hat{f}_H(y)$		
Quantile	Variance $\times 10^{-5}$	$(\text{Bias})^2$ $\times 10^{-5}$	MSE $\times 10^{-5}$	Variance $\times 10^{-5}$	$(\text{Bias})^2$ $\times 10^{-5}$	MSE $\times 10^{-5}$
0.15	2159(28)*	24(3.0)	2173(28)	1921(25)	55(6.4)	1977(26)
0.25	1713(23)	17(2.5)	1730(23)	1553(20)	35(4.4)	1589(20)
0.35	1550(22)	15(2.1)	1565(22)	1414(20)	18(2.4)	1432(20)
0.45	1513(23)	13(1.5)	1526(22)	1375(21)	13(1.5)	1388(21)
0.55	1478(22)	18(2.4)	1496(22)	1343(21)	18(2.3)	1361(21)
0.65	1531(24)	13(1.5)	1544(24)	1384(22)	15(1.7)	1400(22)
0.75	1713(22)	20(3.1)	1733(22)	1547(20)	33(5.1)	1580(21)
0.85	2134(28)	25(3.5)	2159(28)	1908(27)	73(7.3)	1981(27)

* Simulation standard errors are in parentheses.

Table 3. Simulated mean square errors of quantile estimates using $\hat{f}_H(y)$ and $\hat{f}_B(y)$ for a Chi-square superpopulation

	$\hat{f}_H(y)$			$\hat{f}_H(y)$		
Quantile	Variance $\times 10^{-4}$	$(\text{Bias})^2$ $\times 10^{-4}$	MSE $\times 10^{-4}$	Variance $\times 10^{-4}$	$(\text{Bias})^2$ $\times 10^{-4}$	MSE $\times 10^{-4}$
0.15	578(8)*	6(1.1)	584(8)	506(6)	48(3.1)	554(7)
0.25	724(13)	10(1.1)	734(13)	639(12)	6(0.7)	645(12)
0.35	896(12)	10(1.4)	906(12)	791(11)	10(1.4)	801(11)
0.45	1176(30)	15(2.1)	1192(30)	1036(30)	30(3.1)	1066(30)
0.55	1393(29)	18(2.8)	1411(30)	1223(29)	44(4.8)	1267(30)
0.65	1855(41)	31(4.4)	1886(41)	1657(39)	76(7.8)	1734(40)
0.75	2672(70)	27(3.9)	2699(70)	2410(69)	77(8.2)	2486(69)
0.85	4427(112)	45(6.5)	4471(112)	3990(110)	107(11.6)	4097(110)

* Simulation standard errors are in parentheses.

Appendix: Proofs of Lemmas

Proof of Lemma 3.1. For any S, H or B, the techniques of Silverman ((1986), Section 3.3) or Jones ((1989), Appendix) may be applied to the right hand side of (3.2). We illustrate this with result (i). From (1.4) and (3.2) and since the y's are identically distributed,

$$B(\hat{f}_S(y)) = \int [\int K(y-t)/h_S)f(t)dt/h_S - f(y)]^2 dy.$$

On making the substitution $t = y - uh_S$, and on expanding $f(y - uh_S)$ in a Taylor series about y and simplifying, the result is obtained.

Proof of Lemma 3.2. For H or B, the technique of Scott ((1992), Section 3.2.2) may be applied to the right hand side of (3.1). We illustrate this with result (i). From (2.3) and (3.1) the bias

$$B(\hat{f}_H(y)) = \sum_{i=1}^{k} \int_{B_i} [p_i/b - f(y)]^2 dy.$$

Now p_i is approximated by $\int_{B_i} f(t)dt$ as $N \to \infty$. On expanding $f(t)$ in a Taylor series about y and simplifying, the result is obtained.

Proof of Lemma 3.3. On applying a result obtained by Rao (1979) (see, for example, Rao (1988), p.429) to (3.3), we obtain

$$V_p\{\hat{f}_T(y)\} = \sum_{j<l=1}^{N} w_{jl}(q_T(y_j) - q_T(y_l))^2,$$

where $w_{jl} = E_p(w_{sj} - 1/N)(w_{sl} - 1/N)$. Application of the operator E_m to $V_p\{\hat{f}_T(y)\}$ and expansion of the square yields the required result.

Proof of Lemma 3.4. We prove only the result for (iii). The proof of (i) is similar, and (ii) is a special case of (i) in which K_S is the uniform density on the interval $(-1/2, 1/2)$ and $h_S = b$ so that $R(K_S) = 1$.

Part (iii). For $\hat{f}_B(y)$, the smoothed histogram or kernel estimate based on binned data, we note from (2.5) and (2.7) that

$$q_B(y_j) = \frac{1}{h_B} \sum_{i=1}^{k} K_H(\frac{m_i - y_i}{b}) K_B(\frac{y - m_i}{h_B}).$$

Now $q_B(y_j)^2$ will involve a double sum with indices i and i' so that $E_m(q_B(y_j)^2)$ involves the evaluation of the term

$$\frac{1}{b^2} \int K_H(\frac{m_i - t}{b}) K_H(\frac{m_{i'} - t}{b}) f(t)dt. \tag{A.1}$$

Since $K_{_H}$ is the uniform density on $(-1/2, 1/2)$, (A.1) reduces to 0 for $i \neq i'$ and to

$$\frac{1}{b^2}\int K^2_{_H}(\frac{m-i-t}{b})f(t)dt \tag{A.2}$$

for $i = i'$. In (A.2), $f(t)$ may be expanded in a Taylor series around m_i. On retaining the leading term only, (A.2) reduces to $f(m_i)/b$. By similar arguments and by assuming that the joint density function $g_{jl}(t,u) = f(t)f(u)$,

$$E_m\{q_{_B}(y_j)q_{_B}(y_l) = \frac{1}{b^2}\int K_{_H}(\frac{m_i-t}{b})K_{_H}(\frac{m_i-u}{b})g_{jl}(t,u)dtdu = f(m_i)f(m_{i'})$$

for $i \neq i'$. Consequently $E_mV_p(\hat{f}_{_B}(y))$ can be expressed as

$$\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big]\Big[\sum_{i=1}^{k} bf(m_i)K^2_{_B}(\frac{y-m_i}{h_{_B}}) \\ -\sum_{i\neq i'=1}^{k} b^2 f(m_i)f(m_{i'}K_{_B}(\frac{y-m_i}{h_{_B}})K_{_B}(\frac{y-m_{i'}}{h_{_B}})\Big]\Big/nh^2_{_B}. \tag{A.3}$$

On noting that $\sum_{i=1}^k bf(m_i)z(m_i) \cong \int f(t)z(t)dt$, where $z(m_i)$ is $K_{_B}$ or $K^2_{_B}$, and on expanding $f(t)$ in a Taylor series around y, then (A.3) integrated over y is approximately

$$\Big[\frac{R(K_{_B})}{nh_{_B}} - \frac{R(f)}{n}\Big]\Big[-n\sum_{j\neq l=1}^{N} w_{jl}\Big].$$

Proof of Lemma 3.5. Part (i). For the histogram estimate

$$\text{Var}_{\,p}(\hat{f}_{_H}(y)) = \frac{d_ip_i(1-p_i)}{nb^2} \tag{A.4}$$

for $y \in B_i$. The integrated variance is obtained on summing (A.4) over i and multiplying the result by b. Write $d_i = \bar{d} + (d_i - \bar{d})$. If the variation in the d_i's is small then the integrated variance is approximately $\bar{d}\sum_{i=1}^k p_i(1-p_i)/(nb)$. As $N \to \infty$ and $b \to 0$ then, on using the results of Scott ((1992), Section 3.2.2) in a fashion similar to Lemma 3.2, the asymptotic integrated variance is given by

$$\bar{d}(\frac{1}{nb} - \frac{R(f)}{n}).$$

Part (ii). For the smoothed histogram the design variance is given by

$$V_p(\hat{f}_{_B}(y)) = \frac{1}{nh^2_{_B}}\Big[\sum_{i=1}^{k} d_ip_i(1-p_i)K^2_{_B}(\frac{y-m_i}{h_{_B}}) \\ -\sum_{i\neq j=1}^{k} d_{ij}p_ip_jK_{_B}(\frac{y-m_i}{h_{_B}})K_{_B}(\frac{y-m_j}{h_{_B}})\Big]. \tag{A.5}$$

As in the case of the histogram estimator we write $d_i = \bar{d} + (d_i - \bar{d})$. Also $d_{ij} = \bar{d}' + (d_{ij} - \bar{d}')$. If the variation in the d_i's is small, as well as the variation in the d_{ij}'s, then (A.5) is approximately

$$\frac{1}{nh_B^2}\Big[\bar{d}\sum_{i=1}^{k} p_i K_B^2(\frac{y-m_i}{h_B}) - \bar{d}' \sum_{i,j=1}^{k} p_i p_j K_B(\frac{y-m_i}{h_B})K(\frac{y-m_j}{h_B}) - (\bar{d}-\bar{d}')\sum_{i=1}^{k} p_i^2 K_B^2(\frac{y-m_i}{h_B})\Big].$$

On noting that, for example, $\sum_{i=1}^k p_i K_B^2\{(y-m_i)/h_B\}$ is an approximation to $\int K_B^2\{(y-t)/h_B\}dt$, the techniques of Scott ((1992), p.130) may be applied to obtain the approximate integrated variance. This yields

$$\bar{d}\frac{R(K_B)}{nh_B} - \bar{d}'\frac{R(f)}{n} - (\bar{d}-\bar{d}')\frac{bR(f)R(K_B)}{n}.$$

Further Detail on Comment 3.3

In a multistage design it may be assumed that $g_{jl}(t,u) = f(t)f(u)$ when j and l are in different primaries, so that the effect of the lack of independence of units within primaries may be small. Further, although the term $R(f)/n$ in Lemma 3.4 appears in Scott (1992), it is considered negligible by Jones (1989). This is the only component of variance calculation in which the assumption of independence is necessary. Finally, when $g_{jl}(t,u)$ is bivariate normal, the resulting term in the integrated variance is proportional to $R(f)/n$ with the constant of proportionality depending on the correlation. This leads to different design effect corrections to $R(K_S)$ and $R(f)$ appearing in (3.11). A similar result occurs in the design-based approach to the smoothed histogram. In this approach independence is not assumed.

Acknowledgements

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors are grateful to a referee who provided many very helpful comments on an earlier draft.

References

Bellhouse, D. R. and Rao, J. N. K. (1994). Analysis of domain means in complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 29-34. American Statistical Association, Alexandria, Virginia.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Austral. J. Statist.* **5**, 93-105.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597-604.

Chambers, R. L., Dorfman, A. H. and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *J. Amer. Statist. Assoc.* **88**, 268-277.

Dorfman, A. H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Austral. J. Statist.* **35**, 29-41.

Dorfman, A. H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Ann. Statisti.* **21**, 1452-1475.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya (C)* **37**, 117-132.

Jones, M. C. (1989). Discretized and interpolated kernel density estimates. *J. Amer. Statist. Assoc.* **84**, 733-741.

Kuk, A. Y. C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika* **75**, 97-103.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. *Proceedings of the Section on Survey Research Methods*, 280-285. American Statistical Association, Alexandria, Virginia.

Ontario Ministry of Health (1992). Ontario Health Survey: User's Guide. Volumes I and II. Queen's Printer for Ontario.

Rao, J. N. K. (1979). On deriving mean square errors and their non-negative unbiased estimators in finite population sampling. *J. Indian Statist. Assoc.* **17**, 125-136.

Rao, J. N. K. (1988). Variance estimation in sample surveys. In *Handbook of Statistics*, Vol. 6 (Edited by P. R. Krishnaiah and C. R. Rao), 427-447. North-Holland, Amsterdam.

Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365-375.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *J. Amer. Statist. Assoc.* **76**, 221-230.

Roberts, G., Rao, J. N. K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**, 1-12.

Royall, R. M. (1970). On finite population sampling under certain linear regression models. *Biometrika* **57**, 377-387.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer Verlag, New York.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley, New York.

Scott, D. W. and Sheather, S. J. (1985). Kernel density estimation with binned data. *Comm. Statist. Theory Methods* **14**, 1353-1359.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7, Canada.

E-mail: bellhouse@stats.uwo.ca

Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7, Canada.

E-mail: stafford@fisher.stats.uwo.ca

(Received February 1997; accepted January 1999)