International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 3 (Special Issue), 2004, pp 259–268

Double Sampling for Stratification with Application to Fisheries Surveys

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

W. C. Liu
Statistics Canada
Ottawa, Ontario
Canada K1A 0T6

[Received April 15, 2004; Accepted August 14, 2004]

Abstract

Under double sampling for stratification scheme, a density estimate is obtained for the distribution of the measurements taken on the second phase. Estimates of quantiles and their standard errors are derived from this estimated density. Further, estimates and standard errors are obtained for the mean of the first phase measurements conditional on certain values of the second phase measurements. The results are applied to a double sampling scheme for length and age measurements on fish taken from trawl surveys in fisheries. At the first phase the lengths of the fish are determined and the fish are stratified by length. For the second phase independent subsamples are selected within each stratum and the ages of the subsampled fish are determined. We consider estimation of the age distribution of the fish and estimation of the average length of the fish at each age.

Keywords and Phrases: Double sampling, Kernel density estimation, Mode estimation, Quantile estimation, Trawl surveys of fish populations.

AMS Classification: Primary 62D05.

1 Introduction and Notation

Double or two-phase sampling is a technique used when the cost to measure a variable y is relatively high either in terms of time or money. A second variable x is less costly to measure. The general idea is to obtain an initial or first phase sample and make the less costly measurement. Then a subsample of the initial sample is chosen and the more costly measurement is made on this smaller sample. The information collected from the first phase is used to improve the efficiency of the estimation scheme. Cochran (1977, Ch. 12) has given the theory for estimating the finite population mean of the y-variable using a regression estimator based on the sample means of the x's taken at the first and second phases and on the sample mean of the y's taken at the second phase. Rao (1973) has stratified on the x-variable and then has taken a stratified sample at the second phase, again to estimate the finite population mean of the y-variable.

In this paper, double sampling for stratification has been used in trawl surveys for fisheries. This work was motivated by a problem in the estimation of the mean and standard error in the length of fish at certain ages as well as estimation of the age distribution of the fish taken from a trawl survey in an off-shore fishery on the east coast of Canada. A first phase sample was chosen and the lengths (x) of the fish were quickly determined. The data at this phase are binned on the lengths. Within each bin or stratum a second phase subsample of the fish was selected and the age in years (y) of the fish was determined. The determination of age is more time consuming than the determination of length and hence more costly. Estimation of the mean age of fish in a catch based on double sampling was initiated by Ketchen (1949). A review of the methodology in this field is given in Gulland and Rosenberg (1992) who give variance estimates similar to Rao (1973) for estimation of mean age. The theory for stratification of the trawl survey comprising several catches has been given in Cotter (1998).

Initially, our work is similar to Rao's and hence Gulland and Rosenberg's. Assume that the x-variable is binned and that the bin values are given by x_i for $i = 1, \cdots, I$. A random sample of n is taken and the data are binned according to the I bins with n_i units observed in the i^{th} bin. These bins are taken as strata so that at the second phase a simple random sample without replacement of m_i from the n_i is chosen independently for each $i = 1, \cdots, I$ and the measurement x is obtained. This measurement at the second phase can also be binned with bin values given by y_j for $j = 1, \cdots, J$. Rao (1973) proceeds without binning on y to estimate the finite population mean of the y-variable. Here we wish to estimate the distribution of y using binned data and to estimate the conditional mean of the x-values given a bin value for y. For the purpose of moment calculations we make an assumption that appears in Rao (1973): the sample sizes m_i at the second phase are determined by the relationship $m_i = v_i n_i$, where the fractions v_i have been fixed in advance of sampling.

Let π_{ij} be the population proportion that would be observed in both the i^{th} bin on the x's and the j^{th} bin on the y's. The marginal totals π_{i+} and π_{+j} are the population proportions for the i^{th} bin on the x's and the j^{th} bin on the y's respectively. The finite

population mean of the y's at each distinct x-value is given by

$$\bar{y}_i = \sum_{j\,=\,1}^{J} \pi_{ij} y_j, \tag{1}$$

and the overall population mean of the y's is

$$\bar{y} = \sum_{j=1}^{J} \pi_{+j} y_j. \tag{2}$$

The mean in (2) is the estimand in Rao (1973).

Consequently, based on the binning, the distribution for y is given by π_{+j} for $j = 1, \cdots, J$. It may also be of interest to estimate$f(y)$, the underlying density function for y rather than the discretized version given by π_{+j}. The density function $f(y)$ may be obtained on assuming an infinite population and on letting the bin size approach 0. Details of the possible superpopulation assumptions that can be made to justify $f(y)$ are found in Bellhouse and Stafford (1999). In particular, we assume that there is a nested sequence of finite populations such that the "empirical" cumulative distribution functions on the finite populations in the sequence converge to a smooth function $F(y)$. The density function $f(y) = dF(y)/dy$. The other quantity of interest is the conditional mean of the x's given y_j, which may be expressed as

$$\bar{x}_j = \sum_{i\,=\,1}^{I} x_i \pi_{ij}/\pi_{+j}. \tag{3}$$

Here we pursue estimation of $f(y)$ and $\bar{x}_j$, as well as related quantities, rather than $\bar{y}$.

The data that are available are the counts from the first phase showing n_i for the i^{th} bin on the x's, $i = 1, \cdots, I$ and the counts from the second phase showing m_{ij} for the j^{th} bin on the y's at the subsampled bin i on the x's. The second phase subsample size $m_i = \sum_j m_{ij}$. Assuming an infinite population or a large finite population and that the first phase sample is chosen randomly, then the proportions π_{ij} are probabilities from a multinomial distribution.

2 Distribution of Binned Observations Taken at the Second Phase

Based on the multinomial assumption at the first phase and simple random sampling without replacement in each bin at the second phase, an estimate of π_{ij} and hence π_{+j} may be obtained. Had all n_i been measured for x in the i^{th} bin, the count would have been n_{ij} and an estimate of π_{ij} would be n_{ij}/n. From sampling theory, conditional

on the first phase sample the estimate of n_{ij} is $\hat{n}_{ij} = n_i m_{ij}/m_i$ so that ultimately $\hat{\pi}_{ij} = n_{ij}/n$. The estimate of the distribution of observations in the j^{th} bin for y

$$\hat{\pi}_{+j} = \sum_{i\,=\,1}^{I} \hat{\pi}_{ij} \tag{4}$$

is obtained. The variance of (4) may be derived using conditional expectations and variances. In particular, $V(\hat{\pi}_{+j}) = E_m V_d(\hat{\pi}_{+j}) + V_m E_d(\hat{\pi}_{+j})$, where E_m and V_m denote the expectation and variance with respect to the multinomial sampling model associated with the first phase of sampling and where E_d and V_d denote the expectation and variance with respect to the sampling design within the strata at the second phase. We find

$$V_m E_d(\hat{\pi}_{+j}) = \frac{\pi_{+j}(1-\pi_{+j})}{n}. \tag{5}$$

On using Rao's assumption in two-phase sampling that the v_i have been fixed in advance of sampling and that sampling at the second phase is independent between strata, we find

$$E_m V_d(\hat{\pi}_{+j}) = \sum_{i\,=\,1}^{I} \frac{\pi_{ij}(\pi_{i+} - \pi_{ij})(1-\nu_i)}{(n\pi_{i+} - 1)\nu_i} \tag{6}$$

to a first order approximation. A consistent estimator for $V(\hat{\pi}_{+j})$ is found from (5) and (6) with each π replaced by the appropriate $\hat{\pi}$. We obtain

$$\hat{V}(\hat{\pi}_{+j}) = \frac{\hat{\pi}_{+j}(1-\hat{\pi}_{+j})}{n} + \sum_{i\,=\,1}^{I} \frac{\hat{\pi}_{ij}(\hat{\pi}_{i+} - \hat{\pi}_{ij})(1-\nu_i)}{(n\hat{\pi}_{i+} - 1)\nu_i},$$

where $\hat{\pi}_{i+} = n_i/n$ and where $\hat{\pi}_{+j}$ and $\hat{\pi}_{ij}$ have been previously defined.

3 Kernel Smoothing and Related Derived Quantities

In the results that follow it is useful to have the estimated variance of $\hat{\bar{y}}$, where $\hat{\bar{y}}$ is the estimate of the population mean with the estimate given by (2) with the appropriate π replaced by $\hat{\pi}$. Adapting the estimate from Rao (1973, eq. 3) to the notation used here and letting $N \to \infty$, the estimated variance is given by

$$\sum_{i\,=\,1}^{I} \frac{n_i(n_i-1)}{n(n-1)m_i}\, s_i^2 + \frac{1}{n(n-1)} \sum_{i\,=\,1}^{I} n_i(\hat{\bar{y}}_i - \hat{\bar{y}})^2, \tag{7}$$

where

$$s_i^2 = \sum_{j\,=\,1}^{J} m_{ij}(y_j - \hat{\bar{y}}_i)^2 \tag{8}$$

and where $\hat{\hat{y}}_i$ and $\hat{\hat{y}}$ are (1) and (2) respectively with each π replaced by the appropriate $\hat{\pi}$.

The underlying density function $f(y)$ for y may be estimated through kernel smoothing of the histogram defined by the $\hat{\pi}_{+j}$ in (4) for $j = 1, \cdots, J$. Following Bellhouse and Stafford (1999) the kernel density estimate based on any histogram is the sum over the bins of the estimated proportion of observations in the bin times the kernel for a given bandwidth. In other words, the kernel estimate is a weighted sum of the bin proportions with the weights given by the kernel function. This reduces to the usual kernel density estimate given in Silverman (1986) when the bin contains a single observation so that the estimated proportion is one over the sample size. For the data at hand in the fisheries survey, the estimated bin proportion is $\hat{\pi}_{+j}$ and the bin observation is y_j for the j^{th} bin. This results in the kernel density estimator

$$\hat{f}(y) = \frac{1}{h} \sum_{j\,=\,1}^{J} \hat{\pi}_{+j} K(\frac{y - y_j}{h}). \tag{9}$$

In (9) $K(t)$ is the kernel evaluated at the point t and h is the bandwidth or smoothing parameter to be chosen. The kernel $K(t)$ can be any function such that $K(t) \geq 0$ for all t, $\int tK(t)dt = 0$ and $\int t^2K(t)dt$ is finite. Typically, $K(t)$ is chosen to be a probability density function with mean value 0 and finite higher moments so that the above conditions on $K(t)$ are satisfied. At any point y the estimated variance of $\hat{f}(y)$, $\hat{V}(\hat{f}(y))$ can be obtained from (7) and (8) with y_j in these expressions replaced by $K((y - y_j)/h)/h$. In the data analysis that follows we use a standard normal kernel, i.e. $K(t) = \exp(-t^2/2)/\sqrt{2\pi}$. Note that as the bandwidth h in (9) increases, more bins are given nonnegligible weight so that increasing the bandwidth gives greater weight to bins that are farther from the point y at which the estimate is desired. Similarly, as h decreases, the estimate at y depends mostly on bins very close to y.

Denote the estimated α^{th} quantile from $\hat{f}(y)$ by $\hat{q}_\alpha$. One approach to estimation of q_α for finite populations is through Woodruff's (1952) procedure. This procedure involves finding the survey-based estimate of the finite population cumulative distribution function, say $F_N(y)$ estimated by $\hat{F}_N(y)$ and obtaining the α-quantile from this estimated cdf. In particular, the Woodruff estimator $\hat{q}^w_\alpha$ is the smallest value of y in $\hat{F}_N(y)$ such that the value of $\hat{F}_N(y)$ is at least α. Rao, Kovar and Mantel (1990), for example, have taken this approach when considering covariates in the estimation procedure. Further, they provide a formula for an estimate of the variance of $\hat{q}_\alpha$ (Rao *et al.*, 1990, eq. 17) due originally to C.A. Francisco and W.A. Fuller. The variance estimate may be calculated as follows. Let $z_{\gamma/2}$ be that number from the standard normal distribution such that $\gamma/2$ of the curve lies to the right of this value. For the α-quantile estimate and given γ, calculate the interval given by $\alpha \pm z_{\gamma/2}\sqrt{\hat{V}(\hat{F}_N(\hat{q}^w_\alpha)}$. Let $L(\gamma)$ be the length of this interval. Then the variance estimate of $\hat{q}^w_\alpha$ is given by $\{L(\gamma)/z_{\gamma/2}\}^2/4$.

We use a different approach to variance estimation since our estimation procedure is based on kernel smoothing rather than the Woodruff method. Since the kernel estimate obtained here is a smooth function, we use the Newton-Raphson iterative method from Bellhouse and Stafford (1999) to obtain $\hat{q}_\alpha$. In this approach, the estimated quantile $\hat{q}_\alpha$ is obtained as the solution to the equation

$$g(\hat{q}_\alpha) = \int_{-\infty}^{\hat{q}_\alpha} \hat{f}(y)dy - \alpha = 0.$$

The solution is achieved by iterating the equation

$$\hat{q}_\alpha^i = \hat{q}_\alpha^{i-1} - g(\hat{q}_\alpha^{i-1})/g'(\hat{q}_\alpha^{i-1}) \tag{10}$$

for $i = 1, 2, 3, \cdots$ until convergence is achieved. The function $g'(\hat{q}_\alpha^i)$ is the kernel density estimate at the i^{th} iterative step. For a standard normal kernel, $g(\hat{q}_\alpha^i)$ has the same functional form as $g'(\hat{q}_\alpha^i)$ with the density replaced by the standard normal cumulative distribution function. An initial estimate $\hat{q}_\alpha^0$ can be obtained through the associated histogram. A variance estimate of $\hat{q}_\alpha$ can be obtained through Serfling's (1980) approach. In this approach we start with the identity $F(\hat{q}_\alpha)$, i.e. the theoretical cumulative distribution function of the estimated α-quantile yields an estimate of α. Consequently, $V(F(\hat{q}_\alpha)) = V(\hat{\alpha})$. On using a Taylor expansion to approximate $V(F(\hat{q}_\alpha))$, we have $V(F(\hat{q}_\alpha)) \cong V(\hat{q}_\alpha)\{dF(q_\alpha)/d\alpha\}^2$. On noting that $dF(q_\alpha)/d\alpha = f(q_\alpha)$ and rearranging terms, we have

$$V(\hat{q}_\alpha) \cong V(\hat{\alpha})/[f(q_\alpha)]^2$$

so that an estimate of variance is given by

$$\hat{V}(\hat{q}_\alpha) = \hat{V}(\hat{F}(\hat{q}_\alpha))/[\hat{f}(\hat{q}_\alpha)]^2. \tag{11}$$

The term $\hat{V}(\hat{F}(\hat{q}_\alpha))$, which is the estimated variance of the proportion less than or equal to $\hat{q}_\alpha$, may be obtained as a special case from (7) and (8) in which y_j is replaced by an indicator variable taking the value 1 for $y_j \leq \hat{q}_\alpha$ and 0 otherwise.

On denoting a mode of the distribution by y_m, the corresponding estimated mode $\hat{y}_m$ is by definition a solution to $\hat{f}'(y) = 0$. Alternatively, $\hat{y}_m$ may be found graphically or by numerical trial and error. The definition of the mode through the solution to the estimating equation $\hat{f}'(y) = 0$ leads to variance estimates of $\hat{y}_m$ using Binder's (1983) approach. In Binder's approach, suppose we have a parameter vector θ with a vector of estimating equations $\mathbf{U}(\theta) = \mathbf{0}$, the solution of which yields $\hat{\theta}$, where the entries of $\mathbf{U}(\theta)$ are themselves survey estimates. On letting $\mathbf{J}(\theta) = \partial\mathbf{U}(\theta)/\partial\theta$, then the estimated variance-covariance matrix for $\hat{\theta}$ is given by $[\mathbf{J}(\hat{\theta})]^{-1}\hat{\mathbf{V}}(\mathbf{U}(\hat{\theta}))[\mathbf{J}(\hat{\theta})]^{-1}$, where $\mathbf{V}(\mathbf{U}(\hat{\theta}))$ is the variance-covariance matrix of the survey estimate $\mathbf{U}(\theta)$ evaluated at $\hat{\theta}$. In the current situation we have a single parameter y_m ($\equiv \theta$, now a 1×1 vector)

and hence the estimating equation is of the form $U(y_m) = 0$. For a standard normal kernel

$$U(y_m) = \sum_{j=1}^{J} \hat{\pi}_{+j}(y_m - y_j)\exp\{-(y_m - y_j)^2/(2h^2)\}, \tag{12}$$

so that

$$J(y_m) = \sum_{j=1}^{J} \hat{\pi}_{+\mathrm{j}}\left\{\, 1 - (y_m - y_j)^2/h^2\right\} \;\exp\left\{\; -(y_m - y_j)^2/(2h^2)\right\}. \tag{13}$$

Note from the right hand side of (12) that $U(y_m)$ is a survey estimate of the finite population expression $\sum_{j=1}^{J} \pi_{+j}(y_m - y_j)\exp\{-(y_m - y_j)^2/(2h^2)\}$. From (12) and (13), the estimated variance of $\hat{y}_m$ is given by

$$\frac{\hat{V}(\bar{y}')}{\left[\sum\limits_{j=1}^{J} \hat{\pi}_{+\mathrm{j}}\left\{\, 1 - (\hat{y}_m - y_j)^2/h^2\right\} \;\exp\left\{\; -(\hat{y}_m - y_j)^2/(2h^2)\right\}\right]^2}, \tag{14}$$

where

$$\bar{y}' = \sum_{j=1}^{J} \hat{\pi}_{+j}(\hat{y}_m - y_j)\exp\{-(\hat{y}_m - y_j)^2/(2h^2)\}.$$

The numerator in (14) can be evaluated from (7) and (8) where y'_j replaces y_j and

$$y'_j = (\hat{y}_m - y_j)\exp\{-(\hat{y}_m - y_j)^2/(2h^2)\}.$$

4 Estimation of the Second Phase Mean Conditional on First Phase Results

The estimator of the conditional mean defined in (3) is

$$\hat{\bar{x}}_j = \sum_{i=1}^{I} x_i\hat{\pi}_{ij}/\hat{\pi}_{+j} = \sum_{i=1}^{I} x_i\hat{\pi}_{ij}/\sum_{i=1}^{I}\hat{\pi}_{ij}. \tag{15}$$

Given the first phase sample, this is a ratio of sums of independent variables since samples were chosen independently from each stratum on the second phase. On using a standard Taylor series expansion for the variance of a ratio, then after some algebra the variance of $\hat{\bar{x}}_j$ conditional on y_j taken under the expectations E_m and E_d can be expressed as

$$\frac{1}{\pi_{+j}^2}\sum_{i=1}^{I}(x_i - \bar{x}_j)^2\left(\frac{\pi_{ij}}{n} + \frac{\pi_{ij}(\pi_{i+} - \pi_{ij})(1-\nu_i)}{\nu_i(n\pi_{i+} - 1)}\right) \tag{16}$$

to a first order approximation. A consistent estimator of the variance of $\hat{\bar{x}}_j$ can be found from (15) with each π replaced by the appropriate $\hat{\pi}$. When $m_i = n_i$ there is no sampling variation within a stratum, so that the second term in the round brackets in (15) is 0 whenever $\nu_i = 1$. Consequently, when $\nu_i = 1$ this term will also be set to 0 in the variance estimate.

5 Example

The data we examine were obtained in 1987 by the Fisheries and Oceans Canada during their annual fall multi-species bottom trawl survey. At total of 9936 Yellowtail flounder were initially sampled and their lengths in centimeters were determined. The lengths were binned using one centimeter as the bin width. Within each bin a subsample of fish was obtained by simple random sampling without replacement and the ages in years of the subsampled fish were determined. The data may be found at the web site for the Statistical Society of Canada.

The estimate of the lengths of the fish at each age, given by (14), and its standard error, obtained through (15), were calculated for these data. The results appear in Table 1. It may be noted that there is a near perfect linear trend in lengths with age.

Table 1: Estimated Lengths Conditional on Age

Age in Years (y_j)	Length in cm at Age j, $\hat{\bar{x}}_j$	Standard Error $\sqrt{\hat{V}(\hat{\bar{x}}_j)}$	Estimated Distribution for Age $\hat{\pi}_{+j}$
0	2.58	0.10	0.003950
1	6.04	0.08	0.081457
2	10.66	0.15	0.247526
3	14.52	0.24	0.219273
4	20.48	0.13	0.199192
5	24.54	0.37	0.042416
6	30.12	0.13	0.038824
7	34.95	0.11	0.083389
8	39.10	0.13	0.081961
9	44.26	0.20	0.001711
10	47.50	0.00	0.000201

Figure 1 shows the histogram estimate for age obtained from (4). From this graph it may be seen that the age distribution is bi-modal. The bin width $b = y_{j+\,1} - y_j$ for the histogram in Figure 1 has value $b = 1$. For binning and kernel smoothing of independent and identically distributed observations, Jones (1989) has suggested the relationship $b = 1.25h$ between the bin size and the smoothing parameter h.

Bellhouse and Stafford (1999) have shown that this relationship should be maintained when moving from the case of independence to complex surveys. Since the bin size has been fixed at 1, we choose the smoothing parameter $h = 0.8$. The smoothed histogram, shown in Figure 2, was obtained from (9). It maintains the bi-modality shown by the histogram estimate. On using the methodology in (10) with variance estimate given by (11) we can calculate the quartiles of the distribution shown in Figure 2. The quartiles and their standard errors are shown in Table 2.

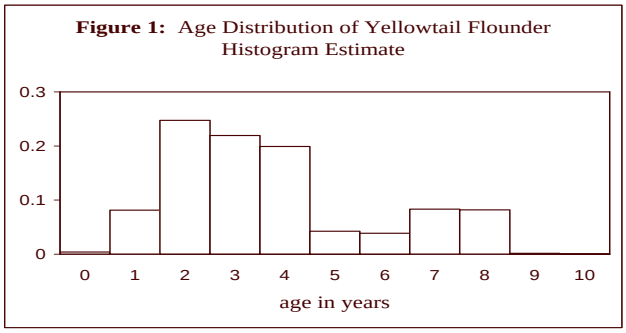

Figure 1: Age Distribution of Yellowtail Flounder
Histogram Estimate

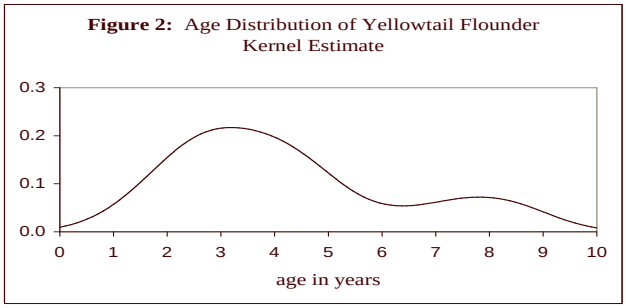

Figure 2: Age Distribution of Yellowtail Flounder
Kernel Estimate

Table 2: Estimated Quartiles for Age

Quartile	Estimate	Standard Error
First	2.63	0.289
Second	3.80	0.133
Third	5.44	0.221

Since the distribution shown in Figures 2 is bi-modal, it is of interest to estimate the locations of the two modes and their standard errors. These may be obtained

through (12) and (13) respectively. The estimates of the modes of this distribution are 3.19 and 7.82 with standard errors 0.059 and 0.041 respectively.

Acknowledgements

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The data are currently available on the Statistical Society of Canada's website at http://www.ssc.ca/documents/case_studies/2000/mixtures_e.html.

References

[1] Bellhouse, D.R. and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, **9**, 407 – 424.

[2] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279 – 292.

[3] Cochran, W.G. (1977). *Sampling Techniques*, 3^{rd} Edition. New York: Wiley.

[4] Cotter, A.J.R. (1998). Method for estimating variability due to sampling of catches on a trawl survey. *Canadian Journal of Fisheries and Aquatic Sciences*, **55**, 1607 – 1617.

[5] Gulland, J.A. and Rosenberg, A.A. (1992). A review of length-based approaches to assessing fish stocks. *FAO Fisheries Technical Paper 323*. Rome: FAO.

[6] Jones, M.C. (1989). Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, **84**, 733 – 741.

[7] Ketchen, K.S. (1949). *Transactions of the American Fisheries Society*, **79**, 205 – 212.

[8] Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, **60**, 125 – 133.

[9] 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.

[10] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

[11] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

[12] Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association* , **47**, 635 – 646.