CLASSICAL AND PREDICTION APPROACHES TO ESTIMATING DISTRIBUTION FUNCTIONS FROM SURVEY DATA

Lynn Kuo, University of Connecticut
U-120, Storrs, CT 06268

KEY WORDS: auxiliary information, design-based approach, finite population distribution function, model based approach, nonparametric regression.

Abstract

Both classical and superpopulation approaches to estimating finite population distribution functions are considered. For the superpopulation approach, nonparametric regression methodology is applied to predict the finite population distribution when auxiliary information is available. Some comparisons are made for the estimators by Monte Carlo methods.

1. Introduction

Estimation of the finite population distribution function from survey data with a design-based approach has received some attention recently. Sedransk and Sedransk (1979) illustrate the usefulness of the sample cumulative distribution functions (CDFs) for stratified designs in making comparisons among subpopulations. Cohen and Kuo (1985a and b) study the properties of the sample CDF from a decision theoretical point of view. They show the sample CDF is admissible for estimating the population distribution function for a class of loss functions with any fixed size sample design. For each of the loss functions, they show that the simple random sampling combined with a step function estimator is the minimax strategy. Francisco and Fuller (1986) study the large sample properties of the sample CDF from stratified cluster samples. Kuk (1988) evaluates the mean squared errors of the Horvitz and Thompson estimator for the distribution function and other related estimators.

Model-based approach to estimating a distribution function has also been studied. Binder (1982) proposes a nonparametric Bayesian approach to estimating the finite population distribution function for simple random sampling and stratified designs. Chamber and Dunstan (1986) propose an estimator when auxiliary information is available. The variable of interest Y is assumed to be related to the auxiliary variable X by a regression function through origin with heteroscedastic errors.

This paper focuses on the finite population distribution function when auxiliary information is available. The regression assumption used by Chamber and Dunstan is relaxed. Let us assume that the finite population consists of N ordered pairs (X_i,Y_i) generated from a bivariate distribution P. The finite population joint distribution function is defined by

$$F(s,t) = \frac{1}{N} \sum_{i=1}^{N} I[X_i \le s,\ Y_i \le t].$$

The finite population distribution function (of the Y variable) is defined by

$$F_Y(t) = \frac{1}{N} \sum_{i=1}^{N} I[Y_i \le t].$$

Let us assume we observe all the X variables $\{X_i\}_{i=1}^{N}$ and a sample of size n of the Y variables by a design. The objective is to estimate $F(s,t)$ and $F_Y(t)$ given the sample. Several nonparametric regression estimators are proposed. These regression-type estimators include the naive weighting, the kernel, and the nearest neighbor estimators.

The method proposed here makes use of a nonparametric superpopulation model. Consequently it is a nonparametric model based approach. The finite population $F(s,t)$ is generated by a bivariate distribution P. The superpopulation P is usually the object of inference in nonparametric density estimation. Stone (1977) and Silverman (1985) provide more detailed discussion in this area, where a sample of size n, (X_i,Y_i), $i = 1,\cdots,n$, is chosen from the bivariate distribution P. Since we observe all the X variables in the finite population, we can assume that the $\{X_i\}$, $i = n+1,\cdots,N$ are random variables chosen from the X marginal distribution P_X. In addition, we observe the ordered pairs $\{X_i,Y_i\}$, $i = 1,\cdots,n$, from the bivariate distribution P. Cohen and Kuo (1988) derive the nonparametric generalized maximum likelihood estimator, nonparametric Bayesian estimator and histogram estimator of P. The predictor of F is studied in this paper by means of the nonparametric regression method.

This method has at least three advantages. (1) It is nonparametric. Therefore, it alleviates survey statisticians of the burden of selecting a parametric model for P. (2) It incorporates the information from the auxiliary variable X by means of the superpopulation P. (3) It adapts the amount of smoothing to the local density dP.

Our primary interest is to predict the finite population distribution function F and marginal distribution F_Y. Other parameters of interest (for example, the population total $Y = \sum_{i=1}^{N} Y_i$ or the ratio $R = \sum_{i=1}^{N} Y_i / \sum_{i=1}^{N} X_i$) can also be predicted using the predictor $\hat{F}_Y(t)$ and $\hat{F}(s,t)$. These predictors are also studied in this paper.

Nonparametric predictors of the distributions F, F_Y and their functionals Y and R are given in Section 2. Monte Carlo results are given in Section 3.

2. Nonparametric Regression Estimators

Let us recall the data consist of n completely observed ordered pairs (X_i,Y_i), $i = 1,\cdots,n$, and additional X_i values $i = n+1,\cdots,N$. Two predictors of $F(s,t)$ can be obtained.

$$\tilde{F}(s,t) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \le s,\ Y_i \le t], \text{ and } (2.1)$$

$$\hat{F}(s,t) = \frac{1}{N}\Big[\sum_{i=1}^{n} I(X_i \le s,\ Y_i \le t) + \sum_{i=n+1}^{N}\sum_{j=1}^{n} W_{ij} I(X_i \le s,\ Y_j \le t)\Big]$$

$$= \frac{n}{N}\tilde{F}(s,t) + (1 - \frac{n}{N})\tilde{G}(s,t), \quad (2.2)$$

where

$$\tilde{G}(s,t) = \frac{1}{(N-n)}\sum_{i=n+1}^{N}\sum_{j=1}^{n} W_{ij} I(X_i \le s,\ Y_j \le t).$$

The weights W_{ij} can be evaluated from one of the following expressions.

(a) The naive estimator:

$$W_{ij} = \frac{I(|X_i - X_j| < \epsilon)}{\sum_{j=1}^{n} I(|X_i - X_j| < \epsilon)}.$$

(b) The kernel estimator:

$$W_{ij} = \frac{K((X_i - X_j)/h}{\sum_{j=1}^{n} K((X_i - X_j)/h)},$$

where the kernel function K satisfies $\int_{-\infty}^{\infty} K(x)dx = 1$. The usual choices of K will be symmetric probability densities, for example, the normal density, uniform density over (-1,1), etc.

(c) The nearest neighbor k estimator:

$$W_{ij} = \begin{cases} \frac{1}{k}, & \text{if } X_j,\ j = 1,\cdots,n, \text{ is one of the k nearest neighbors to } X_i \\ 0, & \text{otherwise.} \end{cases}$$

Let 0 be the set of all completely observed order pairs; $0 = \bigcup_{j=1}^{n}\{(X_j,Y_j)\}$. Let M be the additional observed X variables with unobserved Y values. The predictor $\hat{F}$ with any weighting scheme borrows the Y variables in the completely observed ordered pairs to impute for the unobserved Y values.

For a fixed $X_i \in M$, the naive estimator assigns weight $1/(N\ell)$ to (X_i,Y_{j1}), (X_i,Y_{j2}), $\cdots,(X_i,Y_{j\ell})$, where $Y_{j1},\cdots,Y_{j\ell}$ is the subset of the Y values in the set 0 with the corresponding X values located within ϵ distance from X_i. The number ℓ is the total number of ordered pairs in the set 0 which are located in the strip of width 2ϵ centered at X_i. This estimator encounters the possibility that there is no data in a strip. It also gives a somewhat ragged character to the estimator $d\hat{P}$.

For a fixed $X_i \in M$, the kernel estimator assigns the weights to the points (X_i,Y_i), $j = 1,\cdots,n$ according to the kernel function where Y_j is the Y coordinate of the j^{th} point in the set 0. In the literature on nonparametric density estimation, the kernel estimator has been studied extensively. It has a slight drawback when applied to data from long-tailed distributions. Because the window width is fixed, there is a tendency for spurious noise to appear in the tail of the distribution.

For a fixed $X_i \in M$, the nearest neighbor k estimator assign weight $1/(Nk)$ for $(X_i,Y_{j1}),\cdots(X_i,Y_{jk})$, where $Y_{j1},\cdots,Y_{jk}$ is the subset of the Y values in the set 0 with the corresponding X values to be the closest k values to X_i. While the naive estimator is based on the number of observations falling in a strip centered at X_i, the nearest neighbor estimator is inversely proportional to the size of the strip needed to obtain k observations. The problem of undersmoothing the tail of the distribution should be reduced.

Two predictors for the marginal distribution F_Y can be derived from equations (2.1) and (2.2),

$$\tilde{F}_Y(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le t) \quad \text{and}$$

$$\hat{F}_Y(t) = \frac{n}{N}\tilde{F}_Y(t) + (1 - \frac{n}{N})\sum_{i=n+1}^{N}\sum_{j=1}^{n}\frac{W_{ij}}{N-n} I(Y_j \le t). \quad (2.3)$$

Similarly, two predictors of the population total Y can be obtained from the $\tilde{F}$ and $\hat{F}$ estimators,

$$\tilde{Y} = \sum_{i=1}^{n} NY_i/n \quad \text{and}$$

$$\hat{Y} = \sum_{i=1}^{n} Y_i + \sum_{i=n+1}^{N}\sum_{j=1}^{n} W_{ij}Y_j. \quad (2.4)$$

The corresponding ratio estimators are $\tilde{R} = \tilde{Y}/X$ and

$$\hat{R} = \hat{Y}/X, \quad (2.5)$$

where $X = \sum_{i=1}^{N} X_i$.

3. Monte Carlo Studies

The classical design based estimators for $F(s,t)$ and $F_Y(t)$ are

$$\hat{F}_d(s,t) = \sum_{i\epsilon S} I(x_i \le s,\ y_i \le t)/\pi_i \Big/ \sum_{i\epsilon S} \frac{1}{\pi_i}, \quad (3.1)$$

$$\hat{F}_{Y,d}(t) = \sum_{i\epsilon S} I(y_i \le t)/\pi_i \Big/ \sum_{i\epsilon S} \frac{1}{\pi_i}, \quad (3.2)$$

where S is the index set of the sample, π_i is the probability that the unit i is included in the sample. When unit i is included in the sample, the observed (X_i, Y_i) ordered pair is included in the set 0 defined in Section 2.

The predictors exhibed in Section 2 will be compared to the above estimators by Monte Carlo studies.

Three finite populations of size N=300 are constructed. The first finite population consists of a random sample of 300 points selected from the superpopulation suggested by Hansen, Madow and Tepping (1983). The variable X in the superpopulation has a gamma distribution $\Gamma(2.5)$ with density $f(x) = 0.04\ x \exp(-x/5)$, and the variable Y, conditional on X, has a gamma distribution $\Gamma(c,b)$, where $c = .04\ x^{-3/2}\ (8+5x)^2$ and $b = 1.25\ x^{3/2}\ (8+5x)^{-1}$. The second finite population consists of a random sample of size 300 generated from the models $Y_i = h(X_i) + \nu(X_i)e_i$, where X_i's are generated independently from the beta distribution $\beta e\ (3,1)$ and Y_i, conditional on X_i, has a normal distribution $N(h(X_i),\ \nu^2(X_i))$. The functions $h(x)$ and $\nu(x)$ are chosen to be $h(x) = 12(x^3 - 1.5x^2 + .59x) - .045$ and $\nu(x) = .5x$. The third population is constructed exactly as the second one except $h(x) = 2.5x^2 - 2.5x + 0.8$.

Three different types of sample of size 30 are selected by (i) simple random sampling, (ii) stratified sampling with optimal Neyman allocation based on X, and (iii) stratified sampling with proportional allocation. For sampling plans (ii) and (iii), two strata are constructed for each population, so the sums of the measure of size (X variable) in each stratum are approximately equal. An example of the three types of samples for the three populations is given in Figures 1–3.

Three estimators are computed. The first estimator is the design one given in (3.1) or (3.2); the second one is the nearest neighbor (NN) estimator in (2.2) with k=3 in part (c) for W_{ij}; the third one is the kernel estimator in (2.2) with standard normal density as the kernel and $h = 1.06(30)^{-1/5}$ in part (b) for W_{ij}. (B.W. Silverman (1985) discusses the optimal choice of h). We repeat the sampling 5000 times. The mean squared error (designed based) of each of the estimators for F(s,t) is estimated by

$$MSE(i) = \sum_{\ell=1}^{5000} (\hat{F}^{\ell,i}(s,t) - F(s,t))^2/5000, \quad \text{where}$$

$\hat{F}^{\ell,i}(s,t)$, i = 1,2,3, is the design, NN, and kernel estimator respectively in the ℓ^{th} iteration. For other parameters of interest, such as $F_Y(t)$, $R = \sum_{i=1}^{N} Y_i / \sum_{i=1}^{N} X_i$, and $Y = \sum_{i=1}^{N} Y_i$, the MSE(2) and MSE(3) are defined similarly using (2.3), (2.5) and (2.4) for $\hat{F}_Y(t)$, $\hat{R}$, and $\hat{Y}$.

Tables 1–3 list the $\sqrt{MSE}$ evaluation of the various estimators for the three different populations and three different sampling plans. It can be seen from these tables that the NN estimator has smaller $\sqrt{MSE}$ than the design estimator in almost all cases. The kernel estimator generally performs well. However in a few cases, there is no improvement. Perhaps, better choice of h is needed for those cases.

Acknowledgements

The author wishes to thank Dr. Micahel P. Cohen for helpful discussion and comments and Mr. Tai Ming Lee for assistance with the Monte Carlo studies.

References

Binder, D.A. (1982). Nonparametric Bayesian models for samples from finite populations. J. Roy. Statist. Soc. Ser. B, 44, 388–393.

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

Cohen, M.P. and Kuo, L. (1985a). Admissibility of the empirical distribution functions. Annals of Statistics, 11, 262–271.

Cohen, M.P. and Kuo, L. (1985b). Minimax sampling strategies for estimating a finite population distribution function. Statistics and Decisions, 3, 205–224.

Cohen, M.P. and Kuo, L. (1988). Estimating a bivariate distribution function with partially missing data. University of Connecticut, Statistics Department, Technical Report No. 88–11.

Francisco, C.A. and Fuller, W.A. (1986). Estimation of the distribution function with a complex survey. Proc. Sec. Survey Res. Methods, Amer. Statist. Assoc., Washington, D.C., 37–45.

Hansen, M., Madow, W.G., and Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample survey. Journal of American Statistical Association, 78, 776–793.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. Biometrika, 75, 97–104.

Sedransk, N., and Sedransk, J. (1979). Distinguishing among distributions using data from complex sample designs. Journal of American Statistical Association, 74, 754–760.

Silverman, B.W. (1985). Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

Stone, C.J. (1977). Consistent nonparametric regression. Ann. Statist., 5, 595–645.

Figure 1. The Gamma–Gamma Population and an Example of its Samples.

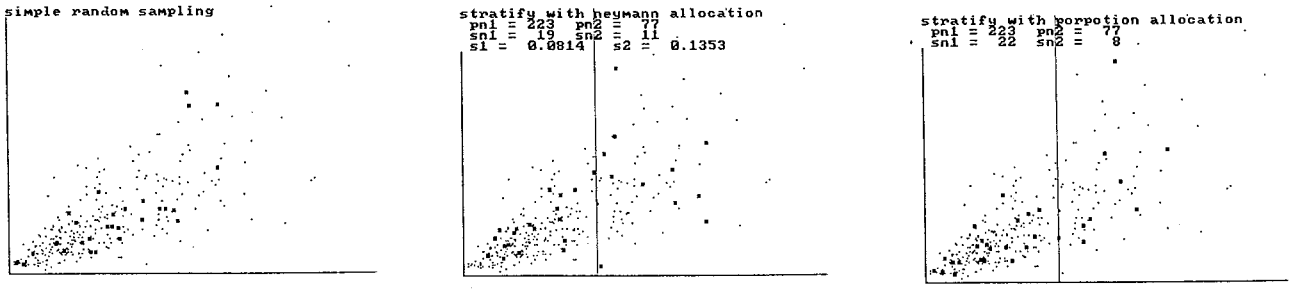


Figure 2. The Cubic Population and an Example of its Samples.

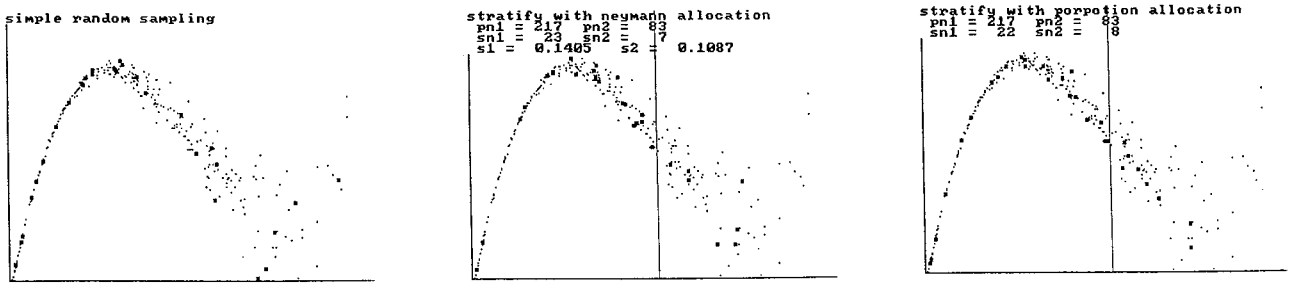
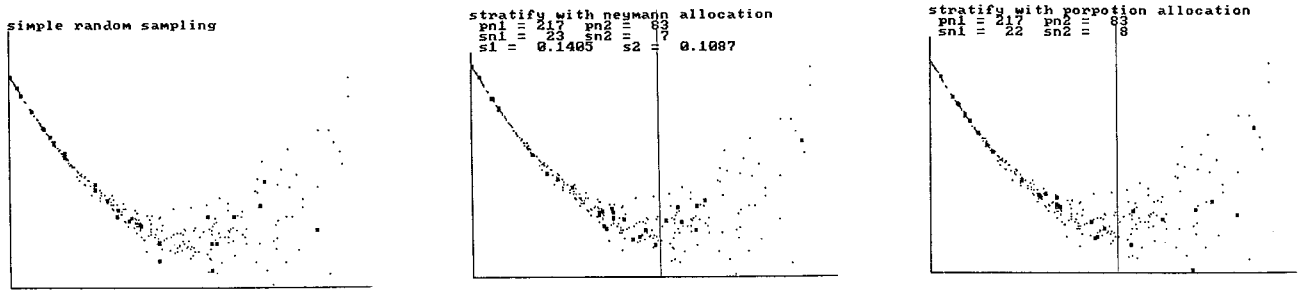


Figure 3. The Quadratic Population and an Example of its Samples.



Note 1: The sample consists of 30 completely observed ordered pair marked by ■(set 0) and the X variable of the remaining points (set M).

Note 2: The middle line is the stratum boundary.

Table 1. Monte Carlo Evaluation of the Root Mean Squared Errors of the Three Estimators for the Gamma–Gamma Population

Sampling Plan	Parameters of Interest: F(5, 1.25) = .163			F(10, 2.5) = .453			F(20, 5) = .817		
	Estimator								
	1	2	3	1	2	3	1	2	3
i	.065	.041	.059	.085	.057	.058	.067	.046	.043
ii	.068	.045	.061	.080	.061	.063	.045	.045	.045
iii	.061	.041	.046	.072	.058	.059	.050	.046	.043

	$F_Y(1.25) = .257$			$F_Y(2.5) = .55$			$F_Y(5) = .827$		
	1	2	3	1	2	3	1	2	3
i	.076	.067	.069	.086	.077	.077	.063	.056	.054
ii	.078	.072	.072	.079	.077	.078	.047	.050	.049
iii	.073	.068	.072	.076	.076	.077	.051	.053	.056

	R = .296			Y = 919.136		
	1	2	3	1	2	3
i	.042	.030	.031	126.5	91.7	92.8
ii	.032	.028	.028	96.3	83.9	83.9
iii	.034	.029	.027	102.4	88.8	87.6

Note: The sampling plan i is the simple random sampling plan without replacement;
the sampling plan ii is the stratified sampling with Neyman allocation;
the sampling plan iii is the stratified sampling with proportional allocation.
The estimator 1 is the design based estimator;
the estimator 2 is the nearest neighbor regression estimation;
the estimator 3 is the kernel regression estimator.

Table 2. Monte Carlo Evaluation of the Root Mean Squared Errors of the Three Estimators for the Cubic Population

Sampling Plan	Parameters of Interest: F(.25, .25) = .093			F(.5, .5) = .223			F(.75, .75) = .717		
	Estimator								
	1	2	3	1	2	3	1	2	3
i	.051	.036	.032	.071	.044	.102	.076	.065	.068
ii	.047	.033	.029	.067	.042	.094	.077	.064	.068
iii	.048	.035	.030	.068	.043	.104	.078	.064	.066

	$F_Y(.25) = .23$			$F_Y(.5) = .44$			$F_Y(.75) = .773$		
	1	2	3	1	2	3	1	2	3
i	.073	.063	.071	.084	.058	.085	.071	.066	.072
ii	.067	.060	.065	.074	.055	.077	.070	.065	.073
iii	.065	.060	.065	.073	.055	.074	.069	.065	.070

	R = 1.495			Y = 149.9		
	1	2	3	1	2	3
i	.151	.104	.150	15.1	10.5	15.0
ii	.136	.102	.140	13.6	10.2	14.0
iii	.133	.102	.132	13.3	10.2	13.2

Table 3. Monte Carlo Evaluation of the Root Mean Squared Errors of the Three Estimators for the Quadratic Population

Parameters of Interest

Sampling Plan	F(.5, .5) = .423 Estimator 1	2	3	F(.75, .75) = .85 1	2	3
i	.085	.046	.064	.061	.041	.042
ii	.071	.044	.062	.061	.039	.041
iii	.073	.045	.063	.056	.039	.041

	$F_Y(.25) = .277$ 1	2	3	$F_Y(.5) = .647$ 1	2	3	$F_Y(.75) = .923$ 1	2	3
i	.077	.072	.078	.083	.065	.082	.046	.045	.045
ii	.080	.076	.078	.083	.066	.081	.047	.044	.045
iii	.076	.071	.077	.081	.062	.081	.046	.042	.045

	R = 1.071 1	2	3	Y = 115.5 1	2	3
i	.126	.106	.122	13.5	11.4	13.1
ii	.131	.115	.118	14.2	12.4	12.7
iii	.123	.104	.121	13.3	11.3	13.0