ON THE ESTIMATION OF DISTRIBUTION AND QUANTILE FUNCTIONS FOR SMALL DOMAINS

Milorad S. Kovačević, The University of Iowa
Dept.of Statistics and Actuarial Science,Iowa City, Iowa 52242

Key words: synthetic estimator, model-based estimator, generalized regression estimator, natural estimator, interpolated estimator, smooth estimator, bootstraping

1. INTRODUCTION

The comparison of two or more domains of the population on the basis of their cumulative distribution functions (*cdf*) and related parameters is an interesting alternative to the usual comparison of the respective means (Sedransk & Sedransk, 1979). Special problems arise if domains of interest are small. Although the literature on small domain estimation is extensive, only very few papers (Fay, 1986, 1987) particularly discuss the estimating of the quantile function (*qf*), and none could be found that concern the estimation of the *cdf*. The value of the *cdf* at y for domain d, $F_d(y)$, is defined as the proportion of domain units that are less then or equal to y. The problem of estimating the *cdf* therefore comes out to be that of estimating the corresponding proportion of the small domain. A related problem is that of estimating the *qf*, $Q_d(p)= inf\{y;F(y)\geq p\}$, for the small domain. Nearly all the relevant literature discusses the estimation of the quantile function under the assumption of simple random sampling from an infinite population. Good reviews of the literature, approaches and related results are given in Sedransk & Smith (1987) and Francisco (1987).

This paper focuses on the estimation of the *cdf* and the *qf* of small subpopulations. In section 2, several different estimators of the *cdf* for small domains are derived. Assuming heteroscedastic superpopulation model, M and DM estimators are presented. Section 3 presents a comparison of three types of *qf* estimators. Besides the natural, an interpolated and a smooth estimator are proposed. Bootstrap resampling scheme with probabilities proportional to the weights of sample units is found to be convinient for the variance estimating.

2.ESTIMATION OF THE *CDF* FOR SMALL DOMAINS

2.1 Basic notation and existing methods. We suppose Sarndal's (1984) setup for small domain estimation: The finite population $\mathcal{U}=\{1,...,N\}$ is divided into D nonoverlapping domains $\mathcal{U}_{.d}$ of known sizes $N_{.d}$ $(d=1,...D)$. It is also divided along a second dimension into H nonoverlapping categories (or groups) $\mathcal{U}_{h.}$ of sizes $N_{h.}$ $(h=1,...H)$. Assuming that groups and strata are identical, a probability sample s of size n, a subset of $\mathcal{U}$, is drawn by given stratified sampling design $p(s)$ that determines the inclusion probabilities $\pi_{hk}=\mathcal{P}r(k\epsilon s_h)$ and $\pi_{hk,l}=\mathcal{P}r(k\epsilon s_h, l\epsilon s_h)$ s_h is a sample of predetermined size $n_{h.}$ from the h-th stratum and $\sum_h n_{h.}=n$. Then, s_{hd} is its part with the random size n_{hd} coming from the cell $\mathcal{U}_{hd}$, a part of a domain d in the stratum h of the size N_{hd}. Thus $s_{h.}=\bigcup_d s_{hd}$ and $s_{.d}=\bigcup_h s_{hd}$. In particular, we assume stratified sample design with $\pi_{hk}= n_h/N_h$, $k\epsilon\mathcal{U}_{h.}$, denoted by strs. Let $\mathcal{Y}$ denote the variable of interest with values $Y_1,...,Y_N$ on the population $\mathcal{U}$.

The *cdf* of $\mathcal{Y}$ on $\mathcal{U}$ is defined as $F_{\mathcal{U}}(y)=N^{-1}\sum_{\mathcal{U}} I_{\{Y_k\leq y\}}$, for any real y, where $\sum_{\mathcal{U}}$ means summation over whole population $\mathcal{U}$, and $I_{\{A\}}$ is the indicator function of the event $\mathcal{A}$. For the sake of simplicity we denote $I_{\{Y_k\leq y\}}$ by $I_k(y)$. The estimator of the population *cdf* based on the stratified sample s is $\hat{F}_{\mathcal{U}}(y)=\sum_h W_h\hat{F}_h(y)$, where the W_h are strata weights and the $\hat{F}_h(y)$ are the estimators of strata *cdfs*, for $h=1,...H$, $\hat{F}_h(y)=\hat{N}_{h.}^{-1}\sum_{s_{h.}} I_{hk}(y)/\pi_{hk}$, and $\hat{N}_{h.}=\sum_{s_{h.}} 1/\pi_{hk}$. $\sum_{s_h}$ denotes summation over all units in the sample s_h.

The *cdf* for a small domain d is

$$F_d(y)=N_{.d}^{-1}\sum_{\mathcal{U}_d} I_k(y)=\sum_h A_{hd}F_{hd}(y)$$

$A_{hd}=N_{hd}/N_{.d}$ is the relative size (weight) of the stratum h within the domain d, and $\sum_h A_{hd}=1$, for $d=1,...D$. A set $\{A_{hd},\ h=1,...H\}$ describes the structure of the domain d with the respect to the assumed stratification. $F_{hd}(y)$ is the *cdf* corresponding to the cell hd, $F_{hd}(y)=N_{hd}^{-1}\sum_{\mathcal{U}_{hd}} I_{hk}(y)$.

The *cdf* of a small domain d may be estimated by a **direct estimator**. There are two types of the resulting estimators: separate and combined. In the case of a separate, for each cell hd we make the design-based estimate

$$\hat{F}_{hd}(y)=\hat{N}_{hd}^{-1}\sum_{s_{hd}}\pi_{hk}^{-1}I_{hk}(y)$$

and then using strata weights the final separate form of the direct estimator, $\hat{F}_d^{Ds}(y)=\sum_h W_h\hat{F}_{hd}(y)$ is obtained. It is biased and in the case of strs for given value of y the bias is equal to $\sum_h(W_h-A_{hd})F_{hd}(y)$. An alternative, also biased, is derived from a single combined ratio. A combined form of a direct estimator is

$$\hat{F}_d^D(y)=\hat{N}_{.d}^{-1}\sum_h\hat{N}_{hd}\hat{F}_{hd}(y) \qquad (2.1)$$

where $\hat{N}_{.d}$ and $\hat{N}_{hd}$ are the design-based estimates of a size of the domain d and the cell hd, respectively. It is known (Cochran,1977; pp.167) that when a small sample is available from each cell hd and a ratio estimate is appropriate, the combined estimate is to be recommended. If value of $F_{hd}(y)$ varies substantially among strata, for given y and d, the variance of $\hat{F}_{hd}(y)$ can be very large.

If cells hd are considered as the poststrata, weights W_h in the separate direct estimator should be substituted by the weights of strata in the domain d, A_{hd}. The resulting **poststratified estimator** is

$$\hat{F}_d^{PS}(y)=\sum_h A_{hd}\hat{F}_{hd}(y) \qquad (2.2)$$

In the case of strs (2.2) becomes asimptotically design-unbiased and has a form $\hat{F}_d^{PS}(y)=\sum_h A_{hd}f_{hd}(y)$, where $f_{hd}(y)$ is the empirical distribution function (*edf*) for the sample s_{hd}. The smaller the n_{hd} are, the more $f_{hd}(y)$ varies making $\hat{F}_d^{PS}(y)$ extremely unstable. It should be noted that if at least one n_{hd} is equal to zero, the estimator (2.2) becomes inoperative as well as in the case of the direct separate estimation.

More efficient estimators may be based on borrowing information from wider domains. Borrowing information is done by assuming a certain implicit or explicit model of the actual population structure.

If we adjust the estimated *cdf* for the entire population $\mathfrak{U}$ to the small domain d, assuming that the small domain resembles population at strata levels in sense of the *cdfs*, we have a **synthetic estimator** of the *cdf* $F_d(y)$ as

$$\hat{F}_d^{SY}(y)=\sum_h A_{hd}\hat{F}_h(y) \qquad (2.3)$$

If we suppose strs, the estimator (2.3) takes on the simpler form $\hat{F}_d^{SY}(y)=\sum_h A_{hd}f_h(y)$ where $f_h(y)$ is the *edf* for the sample from the h-th stratum. The design bias of $\hat{F}_d^{SY}(y)$ in this case is $\sum_h A_{hd}[F_h(y)\text{-}F_{hd}(y)]$, where $F_h(y)$ and $F_{hd}(y)$ are the *cdfs* corresponding to the stratum h and cell hd, respectively. The bias becomes zero if $F_h(y)=F_{hd}(y)$ for a given y and all strata. The design variance is often low and if the implicite model assumption is fulfilled the synthetic estimator is a good choice. But, if it does not hold the synthetic estimator may be significantly design biased. Another adventage of the synthetic estimation is a possibility of getting as many as n distinct values of $\hat{F}_d$. In the case of the direct and the poststratified estimation the set of values of $\hat{F}_d$ containes at most $n_{.d}(<n)$ different values.

2.2 Model-based estimation of the *cdf*. Let $\mathfrak{X}$ denote an auxiliary variable, with the values X_k known for all elements of the population. A superpopulation model (ξ) for $\mathfrak{Y}$ could be specified as a regression through the origin with heteroscedastic errors:

$$Y_k=X_k\beta+v_ke_k,\quad k=1,\ldots N \qquad (2.4)$$

where β is an unknown parameter, the $v_k>0$ are known numbers or known up to multipliers that cancel when β is estimated and the e_k, $k=1,\ldots N$, are iid random variables with mean 0 and variance 1.

A **model-based estimator** of the *cdf* for small domain d, under the superpopulation model ξ is :

$$\hat{F}_d^M(y)=N_d^{-1}\left\{\sum_{s_d}I_k(y)+\sum_{u_d\setminus s_d}\hat{I}_k(y)\right\} \qquad (2.5)$$

$\hat{I}_k(y)$ is the predicted value of $I_k(y)$ for the unobserved element of domain d, based on the model ξ and the ob served data.

A suitable predictor $\hat{I}_k(y)$ can be constructed following an idea of Chambers and Dunstan (1986). For any $k\epsilon\mathfrak{U}$ let us consider $e_k=(Y_k-X_k\beta)/v_k$ as a transformation of Y_k, say $e_k=\mathbb{E}_k(Y_k)$. Due to the nature of model ξ, the $\mathbb{E}_k(Y_k)$ are iid random variables. Let us denote by $G_u(y)$ the *cdf* of $\mathbb{E}_k(Y_k)$ on $\mathfrak{U}$. Then, from ξ, for any $k\epsilon\mathfrak{U}$ and given y

$$E_\xi(I_k(y))=G_u(\mathbb{E}_k(y))=N^{-1}\sum_{l\epsilon u}I_{\{e_l\le\mathbb{E}_k(y)\}}$$

So, we can use a sample-based estimator $\hat{G}_u(\mathbb{E}_k(y))$ for estimating $I_k(y)$. That estimator in the case of model estimation is an *edf*, i.e

$$\hat{I}_k(y)=\hat{G}_u(\mathbb{E}_k(y))=n^{-1}\sum_{l\epsilon s}I_{\{e_l\le\mathbb{E}_k(y)\}}$$

Substituting the unknown parameter β by its BLU estimate under model ξ

$$\hat{\beta}=(\sum_{k\epsilon s}X_kY_k/v_k^2)/(\sum_{k\epsilon s}X_k^2/v_k^2) \qquad (2.6)$$

gives the final form of the model-based (M-) estimator

$$\hat{F}_d^M(y)=$$
$$=N_d^{-1}\left\{\sum_{k\epsilon s_d}I_k(y)+n^{-1}\sum_{l\epsilon u_d\setminus s_d}\sum_{k\epsilon s}I_{\{\hat{e}_k\le\frac{y-X_l\hat{\beta}}{v_l}\}}\right\}$$

where, the $\hat{e}_k=(Y_k-X_k\hat{\beta})/v_k$ are the standardized residuals obtained from the ordinary least squares regression of Y_k on X_k, for $k\epsilon s$.

M-estimators of totals and means are model-unbiased (Holt, Smith & Tomberlin 1979). However, in the case of estimating the *cdf* for small domain, the procedure becomes biased, in general. But, if we assume that β is constant and model ξ holds, the model expectation of $\hat{F}_d^M(y)$ is

$$E_\xi\{\hat{F}_d^M(y)\}=N_d^{-1}\sum_{k\epsilon u_d}G_u(\mathbb{E}_k(y))=E_\xi\{F_d(y)\}$$

Thus, in this case the M-estimator is model-unbiased. The M-estimates depend on correctness of the assumed model of the actual population structure and, if ξ is mis specified, the bias of the estimator will increase.

As an illustration, let us take the one-way ANOVA model ξ_o, such that

$$Y_{hk}=\beta_h+\sigma_h e_{hk}, \quad h=1,...H \text{ and } k\epsilon\mathcal{U}_{h.} \quad (2.7)$$

The e_{hk} are identicly distributed with the unknown distribution function $G_u(.)$ but with the mean 0 and variance 1. β_h is estimated by $\hat{\beta}_h=n_{h.}^{-1}\sum_{s_{h.}} Y_{hk}$, and the model--based estimator of the *cdf* for domain $\mathcal{U}_{.d}$ becomes

$$\hat{F}_d^{M_o}(y)=N_{.d}^{-1}\sum_h\Big\{\sum_{k\epsilon s_{hd}} I_{hk}(y) + n_{h.}^{-1}\sum_{l\epsilon u_{hd}\setminus s_{hd}}\sum_{k\epsilon s_{h.}} I_{\{\hat{e}_{hk}\le\sigma_h^{-1}(y-\hat{\beta}_h)\}}\Big\}$$

Using $\hat{e}_{hk}=(Y_{hk}-\hat{\beta}_h)/\sigma_h$ gives this estimator the form

$$\hat{F}_d^{M_o}(y)=\sum_h\Big\{A_{hd}f_h(y)+\frac{n_{hd}}{N_{.d}}[f_{hd}(y)-f_h(y)]\Big\} \quad (2.8)$$

which can be recognized as the synthetic estimator corrected in the direction of its design bias. Design bias of the estimator (2.8) is

$$B_p\{\hat{F}_d^{M_o}(y)\}= =\sum_h A_{hd}[1-\frac{n_{h.}}{N_{h.}}(1-\frac{N_{hd}}{N_{h.}})][F_h(y)-F_{hd}(y)]$$

The synthetic and the estimator based on the model (2.7) behave similarly with the respect to the design bias. Moreover, estimator (2.8) is model-unbiased, since

$$E_\xi\Big\{\hat{F}_d^{M_o}(y)\Big\}=\sum_h A_{hd}G_u(\mathbb{E}_h(y))=E_\xi\Big\{F_d(y)\Big\}$$

where $\mathbb{E}_h(y)=(y-\beta_h)/\sigma_h$.

A model that takes local differences among cells hd into account is the two-way ANOVA model, say ξ_1:

$$Y_{hdk}=\beta_{hd}+\sigma_{hd}e_{hdk}, \; k\epsilon\mathcal{U}_{hd}, \; h=1,...H, \; d=1,...D \quad (2.9)$$

For all h, d and k the e_{hdk} are iid with mean value 0 and the variance 1, with *df* $G_u(.)$. From the general form of the M-estimator (2.5) for model ξ_1, after certain handling, we have model-unbiased estimator

$$\hat{F}_d^{M_1}(y)=\sum_h A_{hd}\, f_{hd}(y) \quad (2.10)$$

This estimator coincides with the assimptoticaly design-unbiased poststratified estimator (2.2) in the case of a strs design. If ξ_1 rather than ξ_o holds but we still use estimator (2.8), model bias is

$$E_{\xi_1}\Big\{\hat{F}_d^{M_o}(y)-F_d(y)\Big\} =\sum_h(A_{hd}-\frac{n_{hd}}{N_{.d}})\cdot[\sum_d\frac{n_{hd}}{n_{h.}}G_u(\mathbb{E}_{hd}(y))-G_u(\mathbb{E}_{hd}(y))]$$

where $\mathbb{E}_{hd}(y)=(y-\beta_{hd})/\sigma_{hd}$.

2.3 **Generalized regression estimator of the *cdf*.** The generalized regression estimator (GRE) of the *cdf* has the form

$$\hat{F}_u^{GR}(y)=N^{-1}\{\sum_{k\epsilon u}\hat{I}_k(y)+C\sum_{k\epsilon s}\delta_k/\pi_k\} \quad (2.11)$$

where $\hat{I}_k(y)$ represents the predicted value of $I_k(y)$ based on the model ξ, assuming that the Y_k are independent ($k\epsilon\mathcal{U}$), and for $k\epsilon s$ $\delta_k=I_k(y)-\hat{I}_k(y)$. The π_k are determined by the design $p(s)$. δ_k may take on the negative values, so a coefficient C has to provide $\hat{F}_u^{GR}(y)$ with the *cdf* properties. In the section 2.5 for the models ξ_o and ξ_1 and the general stratified design we illustrate finding of C

First, let us estimate $I_k(y)$, for $k\epsilon\mathcal{U}$. A predictor $\hat{I}_k(y)$ can be constructed following an idea of Chambers and Dunstan (1986) explained in 2.2. Here we can use a design-based and a model-unbiased estimator $\hat{G}_u(\mathbb{E}_k(y))$ for estimating $I_k(y)$, i.e

$$\hat{I}_k(y)=\hat{G}_u(\mathbb{E}_k(y))=\hat{N}^{-1}\sum_{r\epsilon s}\pi_r^{-1}I_{\{\hat{e}_r\le\hat{\mathbb{E}}_k(y)\}}$$

β could be estimated either using a BLUE, already given by (2.6), say $\hat{\beta}'$, or the π-inverse estimator (Sarndall-1980)

$$\hat{\beta}''=(\sum_{k\epsilon s}X_kY_k/v_k^2\pi_k)/(\sum_{k\epsilon s}X_k^2/v_k^2\pi_k)$$

Therefore,

$$\hat{F}_u^{GR}(y)=N^{-1}\Big\{\sum_{k\epsilon u}\hat{G}_u(\mathbb{E}_l(y))+ +C\sum_{k\epsilon s}\pi_k^{-1}[I_k(y)-\hat{G}_u(\mathbb{E}_k(y))]\Big\} \quad (2.12)$$

The GRE of the *cdf* for the small domain d is

$$\hat{F}_d^{GR}(y)=N_d^{-1}\{\sum_{k\epsilon u_d}\hat{I}_k(y)+C_d\sum_{k\epsilon s_d}\delta_k/\pi_k\}$$

Based on the one-way ANOVA model ξ_o (2.7) for the general stratified design $p(s)$, and using π-inverse estimator $\hat{\beta}_h=(\sum_{s_h}Y_{hk}\pi_{hk}^{-1})/(\sum_{s_h}\pi_{hk}^{-1})$ of β_h, the estimator $\hat{I}_{hk}(y)$ ($k\epsilon\mathcal{U}_h$) is equal to the sample-based $\hat{F}_h(y)$ and

$$\hat{F}_d^{DM_o}(y)= =\sum_h\{A_{hd}\hat{F}_h(y)+C_{hd}\hat{N}_{hd}N_{.d}^{-1}(\hat{F}_{hd}(y)-\hat{F}_h(y))\} \quad (2.13)$$

where $\hat{F}_{hd}(y)$, $\hat{N}_{hd}$, and $\hat{F}_h(y)$ are design-based estimators of corresponding parameters and C_{hd} is the constant which has to ensure that $\hat{F}_d^{DM_o}(y)$ is *cdf*. A super--script *DM* emphasizes the design-model character of the estimator. Estimator (2.13) can be considered as the synthetic estimator corrected for some amount of the estimated design-bias. This estimator is still model-unbiased. If sample design is strs and $C_{hd}=n_{h.}/N_{h.}$, DM-estimator (2.13) becomes just M (2.8). Design bias in that case is

$$B_p\left\{\hat{F}_d^{DM_o}(y)\right\}=$$
$$=\sum A_{hd}(1-C_{hd}+C_{hd}N_{hd}/N_h)\cdot(F_h(y)-F_{hd}(y))\}$$

It vanishes if for each h and given d $C_{hd}=(N_h-N_{hd})/N_h$.

For a two-way ANOVA model (ξ_1) and for the general stratified design $p(s)$, with β_{hd} estimated by $\hat{\beta}_{hd}=(\sum_{s_{hd}}Y_{hk}/\pi_{hk})/\sum_{s_{hd}}1/\pi_{hk}$, the resulting estimator of the $\hat{I}_{hk}(y)$ is $\hat{F}_{hd}(y)$, $k\epsilon\mathcal{U}_{hd}$, and

$$\hat{F}_d^{DM_1}(y)=\sum_h A_{hd}\hat{F}_{hd}(y) \tag{2.14}$$

C_{hd} does not effect the estimator (2.14), which is the same as the poststratified estimator (2.2).

2.4 A general form of the *cdf* estimator. In this section, for the purpose of the unique numeration of sample units in the whole sample $s=\bigcup_h s_h$, we employ the concept of cumulative labels (*cl*). For the k-th unit from the h-th stratum the *cl* j is defined as

$$j=\sum_{l=1}^{h-1} n_{l.}+k \tag{2.15}$$

where $h=1,...H$, and $k=1,...n_h$. The use of *cl* allows us to consider estimators of the *cdf* in the general linear form

$$\hat{F}_d(y)=\sum_{j\epsilon s} w_d(Y_j)\,I_j(y) \tag{2.16}$$

Weights $w_d(Y_j)=w_{dj}$ fulfill the condition $\sum_s w_{dj}=1$. Geometrically, w_{dj} means the height of a jump of the *cdf* estimate at the Y_j, $j\epsilon s$. The respective weights of the *cdf* estimators discussed in previous subsections are given by Table A.1. Consequently, we can interpret $\hat{F}_d(y)$ as the *cdf* of the data $\mathfrak{D}=\{Y_j,j\epsilon s\}$ that puts probability mass w_{dj} on the unit j. In other words, $\hat{F}_d(y)$ can be considered as a reweighted *edf* of the data $\mathfrak{D}$.

w_{dj} is a positive number and $w_{dj}=O(1/n)$, ie. $w_{dj}\rightarrow 0$ for $n\rightarrow\infty$. In the case of a stratified design the assumption about finite number of strata seems to be important.

2.5 Determination of the coefficient C_d for the DM$_o$ estimator. The GRE in the particular case DM_o given by (2.13) can be expressed in the form of (2.16) as

$$\hat{F}_d^{DM_o}(y)=\sum_h\left\{\sum_{k\epsilon s_{hd}}\frac{N_{hd}-C_{hd}\hat{N}_{hd}+C_{hd}\hat{N}_h}{N_{.d}\hat{N}_h\pi_{hk}}\cdot I_{hk}(y)\right.$$
$$\left.+\sum_{k\epsilon s_h\backslash s_{hd}}\frac{N_{hd}-C_{hd}\hat{N}_{hd}}{N_{.d}\hat{N}_h\pi_{hk}}\cdot I_{hk}(y)\right\}$$

The jumps at the $k\epsilon s_{hd}$ are always positive if $C_{hd}\geq 0$. However, for $k\epsilon s_h\backslash s_{hd}$ weights can take on the negative values. So, C_{hd} has to be chosen to satisfy the condition $0\leq C_{hd}\leq N_{hd}/\hat{N}_{hd}$ for $h=1,...H$ and $d=1,...D$. If $C_{hd}=0$ for all h, the estimator (2.13) takes on the synthetic form (2.3). If $C_{hd}=N_{hd}/\hat{N}_{hd}$ for all h, the estima-tor (2.14) becomes DM_1 estimator.

As a reasonable solution for C_{hd} one can find a dampening factor from the "dampened regression estimator" (Hidiroglou,M.A and Sarndal,C.E, 1986), that is $C_{hd}=(N_{hd}/\hat{N}_{hd})^\alpha$ with

$$\alpha=\begin{cases}-1, & \text{if } N_{hd}/\hat{N}_{hd}\geq 1\\ 1+\epsilon^2, & \text{if } N_{hd}/\hat{N}_{hd}<1\end{cases},\ \epsilon \text{ is any small real number.}$$

In the case of strs design we found that if $C_{hd}=N_{h.}/(N_{h.}-N_{hd})$ DM_o estimator is design unbiased, but this value is accaptable if and only if $n_{hd}<n_h(N_{hd}/N_{h.})[1-(N_{hd}/N_{h.})]$.

3. ESTIMATION OF THE *QF* FOR SMALL DOMAINS

The quantile function (*qf*) of the variable $\mathcal{Y}$ is defined as $Q(p)=inf\{y;\ F(y)\geq p\}$, where $0<p<1$, and y is a real number. A corresponding estimator of a *qf* is $\hat{Q}(p)=inf\{y;\ \hat{F}(y)\geq p\}$, where $\hat{F}(y)$ is an estimator of the *cdf*.

In practice, the estimator $\hat{Q}(p)$ is obtained by arranging data $\mathfrak{D}=\{Y_k;\ k\epsilon s\}$ into an ascending sequence $(\mathfrak{D})=\{Y_{(k)};\ k\epsilon s\}$ and cumulating the jumps $w_{(k)}$ until p is reached. In the following, we discuss some estimators of $Q(p)$ in the case of small domain estimation.

3.1 The natural and the interpolated *qf* estimators The natural estimator ${}_1\hat{Q}_d(p)$ is defined as the first $Y_{(j)}$ such that the cumulative sum of the jumps exceeds p:

$${}_1\hat{Q}_d(p)=\min_{j\epsilon s}\{Y_{(j)};\ \sum_{k=1}^{j} w_{d(k)}\geq p\} \tag{3.1}$$

Since ${}_1\hat{Q}_d(p)$ is a step function with the jumps corresponding to the values from the sample s, it is desirable to smooth it. Jumps are specially high in the case of the direct and poststratified estimation of the *cdf*.

The first step towards a smoothing is linear interpolating between the values $Y_{(j-1)}$ and $Y_{(j)}={}_1\hat{Q}_d(p)$, i.e

$${}_2\hat{Q}_d(p)=Y_{(j-1)}+\frac{[p-\hat{F}_d(Y_{(j-1)})][Y_{(j)}-Y_{(j-1)}]}{\hat{F}_d(Y_{(j)})-\hat{F}_d(Y_{(j-1)})} \tag{3.2}$$

This estimator is applicable even if only very few observations come from domain d. Its form makes sure that for different values of p the corresponding quantiles differ, too. This interpolated estimator uses the information just from the two neighboring sample quantiles taking on value of their linear combination, i.e

${}_2\hat{Q}_d(p)=\alpha Y_{(j)}+(1-\alpha)Y_{(j-1)}$, where $\alpha=1-[\hat{F}_d(Y_{(j)})-p]/w_{d(j)}$ and j is such that $0\leq\hat{F}_d(Y_{(j)})-p\leq w_{d(j)}$.

3.2 A smooth estimator of qf. We increase the number of elements in the linear combination (3.2), giving larger weights to observations whose *cdf* values are closer to p. Such a smooth estimator of the *qf* is

$$_3\hat{Q}_d(p)=\sum_{j=1}^{\nu} Y_j\ \mathcal{K}\left(\frac{\hat{F}_d(Y_j)-p}{w_{dj}}\right) \tag{3.3}$$

where ν is the number of observations with positive value of a weight function $w(.)$, and $\mathcal{K}(.)$ is a real function so that $\mathcal{K}(t)\geq 0$, $\mathcal{K}(t)=\mathcal{K}(-t)$ and $\int_{-\infty}^{\infty}\mathcal{K}(t)dt=1$.

This estimator is somehow related to the "smooth nonparametric estimator of the quantile function" mentioned by Parzen(1979)

$$\hat{Q}(p)=\int_0^1 q(t)\ h^{-1}\mathcal{K}\left(\frac{t-p}{h}\right)\ dt$$

$q(t)$ is the sample quantiles function and h is a smoothing window with the property that $h\rightarrow 0$ when $n\rightarrow\infty$. If we take the natural estimator $_1\hat{Q}_d(p)$ as a sample quantile function, Parzen's nonparametric estimator becomes

$$\hat{Q}(p)=h^{-1}\sum_{j=1}^{\nu} Y_{(j)}\int_{\hat{F}_d(Y_{(j-1)})}^{\hat{F}_d(Y_{(j)})}\mathcal{K}\left(\frac{t-p}{h}\right)\ dt$$

where $\hat{F}(Y_{(0)})=0$. A possible approximation of $\hat{Q}(p)$ can be obtained as

$$\hat{Q}(p)=h^{-1}\sum_{j=1}^{\nu} Y_{(j)}w_{d(j)}\ \mathcal{K}\left(\frac{\hat{F}_d(Y_{(j)})-p}{h}\right)$$

In general, the kernel-type estimators are essentially dependent on the choice of "smoothing parametar" h (Parzen, 1979). If not enough smoothing is done, the estimate will be rough, showing features which do not represent the quantile function, but also, if too much smoothing is done, some important features of the *qf* could be smoothed away. So by substituting h with $w_d(.)$ we have the estimator (3.3). Our idea was to adapt the amount of smoothing to the local *qf* of the data. To show formally an advantage in using w_{dj} as the bandwidth is difficult, although it seems reasonable to adapt the amount of smoothing to the local data. Note that the w_{dj} fulfill all the conditions for being smoothing windows, that is $w_{dj}\rightarrow 0$ when $n\rightarrow\infty$, and $w_{dj}>0$.

$_3\hat{Q}_d(p)$ does not directly depend on the order of the observations. Therefore we can rewrite it as

$$_3\hat{Q}_d(p)=\sum_h\sum_{k\epsilon s_h} Y_{hk}\ \mathcal{K}\left(\frac{\hat{F}_d(Y_{hk})-p}{w_{dhk}}\right) \tag{3.4}$$

$\hat{F}_d(.)$ is one of the estimators of the *cdf* for a small domain, considered in Section 2. We have to assume that there are no ties in the sample s.

4. BOOTSTRAP ESTIMATES OF THE VARIANCE

Let us designate by $\mathfrak{D}$ a set of data defined as $\mathfrak{D}=\{(Y_j,\ w_{dj});\ j\epsilon s\}$ and by $\mathfrak{D}^+$ its subset $\mathfrak{D}^+=\{(Y_j,\ w_{dj});\ j\epsilon s\mid w_d(Y_j)>0\}$. Having observed data $\mathfrak{D}^+$ we propose the following:

1. Draw from $\mathfrak{D}^+$ a bootstrap sample $\{Y_j^*,\ w_{dj}^*;\ j\epsilon s^*\}$ with unequal probabilities and with replacement. The size of the sample s^* is the same as the size of $\mathfrak{D}^+$, say ν for convinience. As a set of probabilities we use a set of the weights $\{w_{dj}\}$. The weights are considered as the constants.

2. Then, calculate

$$\hat{F}_d^*(y)=(\nu)^{-1}\sum_{j=1}^{\nu}[w_{dj}^*\ I_j^*(y)]/\ w_{dj}^*=\sum_{j=1}^{\nu} I_j^*(y)/\nu=f_d^*(y)$$

$f_d^*(y)$ is an *edf* of a bootstrap sample s^*.

3. Independently replicate steps 1 and 2 B times and for the given y calculate the corresponding estimates $\hat{F}_d^{*(b)}(y)$, $b=1,...B$.

4. The bootstrap estimator $E_*(\hat{F}_d^*(y))$ of $F_d(y)$ can be approximated by the Monte Carlo approximation

$$\widetilde{F}_d^*(y)=\sum_b\hat{F}_d^{*(b)}(y)/B$$

and the bootstrap variance estimator of the $\hat{F}_d(y)$ is given as

$$var_*(\hat{F}_d(y))=E_*\left\{\hat{F}_d^*(y)-\hat{F}_d(y)\right\}^2$$

with the approximation

$$v\hat{a}r_*(\hat{F}_d(y))=\sum_b\left(\hat{F}_d^{*(b)}(y)-\widetilde{F}_d^*(y)\right)^2/(B-1)$$

It is easy to prove that the bootstrap variance estimator becomes usual unbiased variance estimator:

$$var_*(\hat{F}_d(y))=\frac{1}{\nu}\left\{\sum_1^{\nu}\frac{(w_{dj}I_j(y))^2}{w_{dj}}-(\hat{F}_d(y))^2\right\}$$

$$=\hat{F}_d(y)\ (1-\hat{F}_d(y))/\nu$$

Bootstrap approach can help in the evaluation of suggested *qf* estimators.

Let m_j^* denote a number of times Y_j appears in the bootstrap sample and the corresponding vector m^* as $(m^*)=\{m_{(1)}^*,\dots\ m_{(\nu)}^*\}$. We have proved that $\hat{F}_d^*(y)$ is just the *edf* for the sample s^*, therefore

$$\hat{F}_d^*(Y_{(j)}^*)=[m_{(1)}^*+\dots+m_{(j)}^*]/\nu$$

Let us define a random variable $R=R(\mathfrak{D},F)=\hat{Q}(p)-Q(p)$.

The bootstrap value of R is

$$R^*=R(\mathfrak{D}^*,\hat{F})=\hat{Q}(\mathfrak{D}^*)-Q(\hat{F})=\hat{Q}(\mathfrak{D}^*)-\hat{Q}(\mathfrak{D})=Y_{(j)}^*-\hat{Q}(\mathfrak{D})$$

or the difference between values of an estimator of p-th quantile based on bootstraped data and the same estimate based on actual sample. Now, to derive the *df* of R^* in the case of the natural estimator $_1\hat{Q}_d(p)$ we use a procedure similar to one Efron (1979) used for the median estimation. That is:

For any integer value r, $1\leq r\leq\nu$, and given p, $0<p<1$, assuming that ${}_1\hat{Q}^*_d(p)=Y_{(j)}$

$$\begin{aligned} Prob_*\{\hat{Q}(\mathfrak{D}^*)>Y_{(r)}\} &= Prob_*\{{}_1\hat{Q}^*_d(p)>Y_{(r)}\} \\ &= Prob_*\{m^*_{(1)}+\ldots+m^*_{(r)}\leq j-1\} \\ &= Prob\{\mathfrak{B}(n,\ w^*_{(1)}+\ldots+\ w^*_{(r)}) \leq j-1\} \\ &= \sum_{k=0}^{j-1}\binom{\nu}{k}(w^*_{(1)}+\ldots+w^*_{(r)})^k[1-(w^*_{(1)}+\ldots+\ w^*_{(r)})]^{\nu-k} \end{aligned}$$

where $\mathfrak{B}(n,p)$ means a binomial distributed random variable. So,

$$\begin{aligned} Prob_*\{R^*=Y_{(r)}-Y_{(j)}\} &= Prob\{\mathfrak{B}(\nu,w^*_{(1)}+\ldots+w^*_{(r-1)})\leq j-1\} \\ &\quad - Prob\{\mathfrak{B}(\nu,w^*_{(1)}+\ldots+w^*_{(r)})\leq j-1\} \end{aligned}$$

and for any sample s we can compute

$$E_*\{(\hat{R}^*)^2\}=\sum_{r=1}^{\nu}[Y_{(r)}-\ Y_{(j)}]^2\ Prob_*\{R^*=\ Y_{(r)}-\ Y_{(j)}\}$$

and use this expression as an estimator of $E(\hat{R}^2)=E(\hat{Q}(p)-\ Q(p))^2$, the expected squared error of the *qf* estimator for the specified value of p.

5. SUMMARY AND CONCLUSION

The main objective of this paper was to investigate posibilities of constructing M and DM estimators of the *cdf* for the small domains. The general stratified sample design was considered but the main properties of the resulting estimators were carried out under the strs design. A general form of the *cdf* estimator was created and its convinience for derivating quantile estimators was shown. Big jumps are characteristics of design-based natural *qf* estimators, so we proposed a smooth *qf* estimator with the variable smoothing window. The use of M or DM estimates of a *cdf* in the natural estimator of a *qf* gives good results. We constructed a resampling procedure which yields variance estimates for the *cdf* as well as for the *qf* estimators. In fact, that procedure is bootstraping with the probabilities proportional to height of the jumps.

Acknowledgements. The author wish to thank Prof. Peter Hackl from *The University of Economics in Vienna*, and Prof. Lous Rizzo from *The University of Iowa, Dept. of Statistics and Actuarial Science* for their valuable comments on the first draft and the last version, respectively.

REFERENCES:

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

2. Cochran,W.G.(1977) *"Sampling techniques", 3rd ed.*, John Wiley&Sons, New York

3. Efron,B.(1979) "Bootstrap methods: another look at the jackknife" *The Annals of Statistics, 7,1-26.*

4. Fay,R.E.(1987) "Small domain estimation through components of variance models", *Bulletin of the ISI, 46, 421-434.*

5. Francisco, C.A. (1987) "Estimation of quantiles and the interquartile range in complex surveys" *Unpublished Ph.d. Thesis, Iowa State University, Ames, IA*

6. Hidiroglou,M.A. and Sarndal,C.E.(1986) "Conditional inference for small area estimation" *Proceedings of the American Statistical Association, Section of Survey Research Methods, 147-158*

7. Holt,D., Smith,T.M.F. and Tomberlin,T.J. (1979) "A model-based approach to estimation for small subgroups of a population" *Journal of the American Statistical Association, 74, 405-410*

8. Parzen,E.(1979) "Nonparametric statistical data modeling" (with discussion). *Journal of the American Statistical Association, 74, 105-131*

9. Sarndal,C.E. (1980) "On π-inverse weighting versus best linear unbiased weighting in probability sampling." *Biometrika, 67, 639-650*

10. Sarndal,C.E. (1984) "Design consistent versus model dependent estimators for small domains", *Journal of the American Statistical Association,79, 624-631.*

11. Sedransk,N. and Sedransk,J.(1979) "Distinguishing among distributions using data from complex sample design", *Journal of the American Statistical Association, 74, 740-756.*

12. Sedransk,J. and Smith (1987) "Inference for finite population quantiles". In Rao,J.N.K and Krishniah, P.R., eds. *Survey sampling. Handbook of statistics*, vol 7. North Holland Publishing Company, Inc. Amsterdam

A.1: Weights w_{dj} for different types of the cdf estimators[1]

Estimator	w_{dj}	Relevant observations
DIR	$1/(\hat{N}_{.d}\pi_{hk})$	$k\epsilon s_{hd}$
	0	$k\epsilon s_h \backslash s_{hd}$
PS	$N_{hd}/(N_{.d}\hat{N}_{hd}\pi_{hk})$	$k\epsilon s_{hd}$
	0	$k\epsilon s_h \backslash s_{hd}$
SYN	$N_{hd}/(N_{.d}\pi_{hk})$	$k\epsilon s_h$
M₀	$(N_{hd}-n_{hd}+n_{h.})/(N_{.d}n_h)$	$k\epsilon s_{hd}$
	$(N_{hd}-n_{hd})/(N_{.d}n_{h.})$	$k\epsilon s_{h.} \backslash s_{hd}$
M_1	$N_{hd}/(N_{.d}n_{hd})$	$k\epsilon s_{hd}$
	0	$k\epsilon s_h \backslash s_{hd}$
DM_o	$(N_{hd}+C_{hd}\hat{N}_{h.}-C_{hd}\hat{N}_{hd})/(\hat{N}_hN_{.d}\pi_{hk})$	$k\epsilon s_{hd}$
	$(N_{hd}-C_{hd}\hat{N}_{hd})/(\hat{N}_hN_{.d}\pi_{hk})$	$k\epsilon s_{h.} \backslash s_{hd}$
DM_1	$N_{hd}/(N_{.d}\hat{N}_{hd}\pi_{hk})$	$k\epsilon s_{hd}$
	0	$k\epsilon s_h \backslash s_{hd}$

[1] $h=1,\ldots H$, and relationship of j with k and h is given by the (2.15)