STATISTICS IN TRANSITION-new series, October 2011
Vol. 12, No. 2, pp. 331—344

FINITE MIXTURES MODEL APPROACH TO SENSITIVE QUESTIONS IN SURVEYS

Shcherbina Artem[1], **Maiboroda Rostyslav**[2]

ABSTRACT

Observations from mixtures of different subpopulations are common in biological and sociological studies. We consider the case, when the observations are taken from a set of groups containing subjects, which belong to different subpopulations. Proportion of each subpopulation in a group is known and can vary from group to group. Our aim is to estimate the means of an observed variable for subjects, which belong to each subpopulation. In this paper we consider the case, when subpopulations are defined by answers on so called "sensitive questions". We consider some parametric and nonparametric estimates of the subpopulation means, such as weighted means, maximum likelihood and weighted least squares estimates. Finite sample properties of these estimates are analyzed. Mean square errors of the estimates are compared on simulated data. Some asymptotic results are also given.

Key words: anonymous survey; sensitive questions; maximum likelihood; weighted mean; weighted least squares.

1. Introduction

Problems of anonymous survey data analysis arise in many sociologic studies. E.g. anonymous surveys are usually used to avoid inadequate answers on so called "sensitive questions" (see Kerkvliet, 1994; Ong & Weiss, 2000). For a recent review of some well-known techniques to treat sensitive questions statistics the reader is addressed to E. Coutts & B. Jann (2011) and references herein. Statistical inference by voting data is another familiar example. In this paper we discuss the case when the results of an anonymous survey are used for inference together with some non-anonymous information on its' respondents.

[1] National Taras Shevchenko University of Kyiv, Ukraine.
E-mail: artshcherbina@gmail.com
[2] National Taras Shevchenko University of Kyiv, Ukraine. E-mail: mre@univ.kiev.ua

Let us consider the motivating example. Suppose that a survey was held in which first year university students were asked the question "Have you ever cheated on a school exam? (Yes/No)" Since this question is sensitive, the survey was held anonymously in different academic groups. As a result, the set of numbers of cheaters (N_{i1}) and non-cheaters (N_{i2}) was obtained for the groups $i = 1,\ldots, K$. The researcher would like to analyze the influence of cheating on the students' school marks. Say, it may be interesting to estimate and compare the mean marks in math for cheaters (μ_1) and non-cheaters (μ_2). Here marks of the students can be obtained from University journals. Thus we don't need to include such questions in the survey.

The simplest way to estimate μ_l is to use the following regression model for the mean marks T_i over the i-th group:

$$T_i \approx p_{i1}\mu_1 + p_{i2}\mu_2 ,$$

where $p_{il} = N_{il} /(N_{i1} + N_{i2})$ is the proportion of cheaters ($l = 1$) or non-cheaters ($l = 2$) in the i-th group. Ordinary least squares estimates (OLSE) for μ_l are

$$\hat{\mu}_l = \frac{1}{K}\sum_{i=1}^{K} a_{il} T_i \qquad (1)$$

where a_{il} are the minimax weights described in Section 3.1. These estimates are unbiased and consistent if p_{i1} is not constant for all $i = 1,...,K$.

But using this approach we drop a good deal of information on the structure of our data which is a mixture of two subpopulations (of cheaters and non-cheaters). Taking this into account one can construct more accurate estimates for μ_{il}. E.g., if some parametric model is imposed on the distributions of the marks of cheaters and non-cheaters, then maximum likelihood estimates are available. Such parametric approach is natural if the considered marks may attain only a small number of fixed values. The simplest case is a binary mark (success=1/failure=0) whose distribution is determined by the probability of success. The ML estimates of such probabilities μ_l for both subpopulations are given in Section 3.3. Note that these estimates can be consistent even when p_{i1} is constant. The ML estimates are asymptotically efficient under suitable assumptions (cf. Borovkov, 1998, Shcherbina 2011a).

On the other hand, the ML estimates can be inadequate if no good parametric model is known for the observed variable distribution. Therefore we developed three new non-parametric estimates for μ_l :

- weighted means of the form (1) with adaptive choice of the weights a_{il} ;
- estimates based on an approximate likelihood (weighted least squares) involving both linear and quadratic statistics from the data;

- estimates in which the ML approach is utilized for the CDF estimation over subpopulations of cheaters and non-cheaters. The estimates for μ_l then are derived as integrals by the CDFs' estimates.

Of course, the proposed estimates may be used not only for the cheating effects analysis, but also for statistical inference by any anonymous survey data.

The estimates are based on the theory of finite mixture models with varying mixing proportions (see Maiboroda, 1996; Maiboroda & Sugakova, 2008).

The rest of the paper is organized as follows. In Section 2 a formal description of the model is presented. In Section 3 we introduce the estimates and discuss their asymptotic properties. In Section 4 performance of the estimates is compared on simulated data. Concluding remarks are placed in Section 5.

2. Model description

Assume that the considered population U consists of two subpopulations U_1 or U_2. Some sample is taken from U at random. This sample is divided into K groups of subjects (sub-samples). Numbers N_{i1} and N_{i2} of the subjects from U_1 and U_2 in group i are known. Then $N_i = N_{i1} + N_{i2}$ is the total number of subjects in the i-th group. Let X_{ij} be the observed variable X of subject j from the group i. The variables X_{ij} are modelled as generated by a probability model, namely as independent random variables with distribution F_k for all subjects from the subpopulation U_k, $k = 1,2$.

To analyse such data a finite mixture model (FMM) may be used. In our case it is of the following form:

$$\mathrm{P}(X_{ij} \in A) = p_{i1}F_1(A) + p_{i2}F_2(A),$$

where A is any measurable subset of the observations space, p_{il} are the probabilities to observe a unit from the l-th subpopulation for subjects from the i-th group (the mixing probability, the concentration of the i-th component in the mixture):

$$p_{il} = \frac{N_{il}}{N_i}, \quad i = 1,2,\ldots,K, \quad l = 1,2.$$

For recent results on FMMs see McLachan and Pell (2000).

Finite mixtures with varying concentrations were considered in Maiboroda (1996), Maiboroda & Sugakova, (2008) for independent observations.

The observations discussed in this paper are dependent since the numbers N_{ij} of subjects belonging to each subpopulation are known. The best results in the

estimation will be achieved if the concentrations p_{il} are widely distributed on $[0,1]$.

Let $\mu=(\mu_1,\mu_2)$ be the mean values and $\sigma^2=(\sigma_1^2,\sigma_2^2)$ be the variances of the distributions F_1 and F_2.

3. Estimation

We want to estimate the parameter μ by observed characteristics X_{ij} and subpopulations sizes N_{i1} and N_{i2} in groups $i=1,...,K$.

3.1. Weighted means

Consider the mean value of X in i-th group $T_i=\frac{1}{N_i}\sum_{j=1}^{N_i}X_{ij}$. The expectation of T_i is

$$\mathrm{E}\,T_i=\mathrm{E}\frac{1}{N_i}\sum_{j=1}^{N_i}X_{ij}=\frac{1}{N_i}(N_{i1}\mu_1+N_{i2}\mu_2)=p_{i1}\mu_1+p_{i2}\mu_2,\quad i=1,2,\ldots,K.$$

Consider the following estimate of the mean value of the l-th subpopulation:

$$\hat{\mu}(a_l)=\frac{1}{K}\sum_{i=1}^{K}a_{il}T_i,$$

where $a_l=(a_{1l},a_{2l},\ldots,a_{Kl})$ is some coefficients vector.

It is readily seen, that the estimate $\hat{\mu}(a_l)$ for μ_l is unbiased if the following conditions hold:

$$\frac{1}{K}\sum_{i=1}^{K}a_{il}p_{im}=I_{\{m=l\}},\quad m=1,2. \tag{2}$$

By straightforward calculations we get

Proposition 1. The variance of the estimate $\hat{\mu}(a_l)$ is

$$\mathrm{D}\,\hat{\mu}(a_l)=\frac{1}{K^2}\sum_{i=1}^{K}a_{il}^2d_i,\quad d_i=\frac{1}{N_i}\left(p_{i1}\sigma_1^2+p_{i2}\sigma_2^2\right),\quad i=1,2,\ldots,K.$$

The best coefficients a_l minimize the variance of $\hat{\mu}(a_l)$ under (2). Coefficients a_l that satisfy (2) and minimize the sum $\sum_{i=1}^{K}a_{il}^2\hat{d}_i$ are

$$a_{i1}(\hat{d})=\frac{\left(r_0(\hat{d})-r_1(\hat{d})\right)p_{i1}+r_2(\hat{d})-r_1(\hat{d})}{\hat{d}_i\left(r_0(\hat{d})r_2(\hat{d})-r_1^2(\hat{d})\right)},\quad a_{i2}(\hat{d})=\frac{r_2(\hat{d})-r_1(\hat{d})p_{i1}}{\hat{d}_i\left(r_0(\hat{d})r_2(\hat{d})-r_1^2(\hat{d})\right)},$$

where $r_j(\hat{d})$ are weighted empirical moments of subpopulations proportions:

$$r_j(\hat{d}) = \frac{1}{K}\sum_{i=1}^{K}\frac{p_{i1}^j}{\hat{d}_i}, \quad j = 0,1,2.$$

The equality $r_0(\hat{d})r_2(\hat{d}) - r_1^2(\hat{d}) = 0$ is possible only when all proportions p_i are equal. In this case the weighted means cannot be used for estimation of the mean value.

Since the values d_i depend on the unknown parameter σ^2, we have to replace the vector $d = (d_1, d_2, \ldots, d_K)$ by some estimate $\hat{d} = \left(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_K\right)$. Taking $\hat{d}$ as the vector of units $I_K = (1,1,\ldots,1)$, we obtain minimax weights $a_l\left(I_K\right)$ introduced in Maiboroda (1996). On general theory of minimax estimation see section 2.21 in Borovkov (1998).

Although minimax weights do not minimize the variance of $\hat{\mu}(a_l)$, they are not too bad. The following theorem establishes conditions of consistency and asymptotic normality of estimates with minimax coefficients.

Theorem 1. Let there exist $C > 0$, such that $r_{0,K}\left(I_K\right)r_{2,K}\left(I_K\right) - r_{1,K}^2\left(I_K\right) > C$ for all K. Then the estimate $\hat{\mu}_l\left(a_l\left(I_K\right)\right)$ is consistent and distributions of the normalized estimate

$$\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}\left(a_l\left(I_K\right)\right)}}\left(\hat{\mu}\left(a_l\left(I_K\right)\right) - \mu_l\right)$$

converge weakly to the standard normal distribution as $K \to \infty$.

The proof of this theorem is based on the Central Limit Theorem.

As we shall see now, the variances σ_l^2 in the formula for $D\hat{\mu}(a_i)$ can be estimated by the same way as μ_l, so the asymptotic normality can be used for asymptotic confidence intervals construction. The same is true for other estimates considered below.

To improve the weighted means performance we may use the adaptive approach. To do this we use the weighed means with minimax weights as pilot estimates for parameter σ^2. E.g. the estimate for the l-th subpopulation is

$$\hat{\sigma}_{l,K}^2 = \frac{1}{K}\sum_{i=1}^{K} a_{il}\left(I_K\right)\frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}^2 - \hat{\mu}^2\left(a_l\left(I_K\right)\right).$$

Then we can estimate the vector d by $\hat{d}_K = \left(\hat{d}_{1,K}, \hat{d}_{2,K}, \ldots, \hat{d}_{K,K}\right)$ where $\hat{d}_{i,K} = \left(p_{i1}\hat{\sigma}_1^2 + p_{i2}\hat{\sigma}_2^2\right)\big/N_i$ and use $r_j\left(\hat{d}_K\right)$ to derive coefficients $a_l\left(\hat{d}_K\right)$. The

adaptive estimate for μ_l is now $\hat{\mu}\left(a_l\left(\hat{d}_K\right)\right)$. On efficiency of adaptive estimates in mixture models see Maiboroda (1999).

A valuable property of the adaptive estimate is its asymptotic normality with the same asymptotic variance as for the estimate with best coefficients $a_l\left(d_K\right)$.

Theorem 2. Under the conditions of Theorem 1, the estimate $\hat{\mu}\left(a_l\left(\hat{d}_K\right)\right)$ is consistent and distributions of the normalized estimate

$$\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}(a_l(d_K))}}\left(\hat{\mu}\left(a_l\left(\hat{d}_K\right)\right)-\mu_l\right)$$

converge weakly to the standard normal distribution as $K\to\infty$.

The proof is based on the following two observations:

(i) $\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}(a_1(d_K))}}(\hat{\mu}(a_1(d_K))-\mu_1)$ converge weakly to the standard normal distribution as $K\to\infty$ by the Central Limit Theorem.

(ii) $\frac{1}{\sqrt{\mathrm{D}\,\hat{\mu}(a_1(d_K))}}(\hat{\mu}(a_1(d_K))-\hat{\mu}(a_1(\hat{d}_K))\to 0$ in probability. It can be shown by applying lemma 3.2.1 from Maiboroda & Sugakova, (2008).

Complete proofs of Theorems 1 and 2 can be found in Shcherbina (2011).

Hence, the adaptive weights allow one to build estimates with the same asymptotic quality, as with the best coefficients.

But such weights should be used carefully for small samples. Estimation of the subsamples' variances can introduce additional variability. That is why the simple minimax coefficients sometimes perform better.

3.2. Weighted least squares

Consider the statistics $S_i=\left(\sum_{j=1}^{N_i}X_{ij},\sum_{j\neq k}X_{ij}X_{ik}\right)$. They are independent random vectors for $i=1,...,K$ with the following mathematical expectations and covariance matrices:

$$f_i(\mu)=\mathrm{E}\,S_i=f\left(N_{i1},N_{i2},\mu\right),$$
$$\Sigma_i=\mathrm{Cov}\,S_i=\Sigma\left(N_{i1},N_{i2},\mu,\sigma^2\right).$$

The elements of the function f are the following:

$$f_1\left(n_1,n_2,\mu\right)=n_1\mu_1+n_2\mu_2,\quad f_2\left(n_1,n_2,\mu\right)=n_1\left(n_1-1\right)\mu_1^2+2n_1n_2\mu_1\mu_2+n_2\left(n_2-1\right)\mu_2^2.$$

The matrix Σ is symmetric with the following elements:

$$\Sigma_{11}\left(n_1,n_2,\mu,\sigma^2\right)=n_1\sigma_1^2+n_2\sigma_2^2,$$
$$\Sigma_{12}\left(n_1,n_2,\mu,\sigma^2\right)=2n_1n_2\mu_1\sigma_1^2+2n_1n_2\mu_2\sigma_2^2+2n_1\left(n_1-1\right)\mu_1\sigma_1^2+2n_2\left(n_2-1\right)\mu_2\sigma_2^2,$$
$$\Sigma_{22}\left(n_1,n_2,\mu,\sigma^2\right)=4n_1\left(n_1-1\right)^2\mu_1^2\sigma_1^2+4n_2\left(n_2-1\right)\mu_2^2\sigma_2^2+4n_1^2n_2\mu_1^2\sigma_2^2+4n_1n_2^2\mu_2^2\sigma_1^2$$
$$+8n_1n_2\mu_1\mu_2\left(\left(n_1-1\right)\sigma_1^2+\left(n_2-1\right)\sigma_2^2\right)+2\left(n_1\sigma_1^2+n_2\sigma_2^2\right)^2-2n_1\sigma_1^4-2n_2\sigma_2^4.$$

The matrix Σ depends on μ and σ^2. Here we estimate them with weighed means estimates as described above, compute matrices $\Sigma_i=\Sigma\left(N_{i1},N_{i2},\hat{\mu}^2\left(a_l\left(I_K\right)\right),\hat{\sigma}^2_{l,K}\right)$ and then use them as fixed. Consider the generalized least squares criterion

$$\sum_{i=1}^{K}\left(S_i-f_i\left(\mu\right)\right)\Sigma_i^{-1}\left(S_i-f_i\left(\mu\right)\right)^T\to\min. \quad (3)$$

Denote the solution of (3) by $\hat{\mu}_{LS}$. This estimator can perform better than linear and adaptive estimates since it uses second order sums.

Note that this LS criterion is approximately equal to the log-likelihood when the sizes N_i of the groups are large due to the asymptotic normality of S_i as $N_i\to\infty$.

In practice it can be difficult to minimize (3). By differentiating (3) with respect to μ we get the following estimating equations (see Heyde, 1997):

$$\sum_{i=1}^{K}\frac{\partial f_i}{\partial\mu}\left(\mu\right)A_i\left(S_i-f_i\left(\mu\right)\right)^T=0.$$

They can be solved by a Newton-type algorithm (see Small & Wang, 2003).

3.3. Maximum likelihood for parametric case

Let the characteristic X be binary, i.e. attains only two values 1 (success) or 0 (failure) with probabilities of success μ_1 and μ_2 for subjects from the first and second subpopulations. The statistic $Z_i=\left(N_{i1},N_{i2},\sum_{j=1}^{N_i}X_{ij}\right)$ is then sufficient for estimating the parameter μ by the observations from i-th group. The density function of Z_i is

$$f_i(x,q)=\mathrm{P}\left(\sum_{j=1}^{N_i}X_{ij}=x\,|\,\mu=q\right)=\sum_{k=0\vee\left(x-N_{i2}\right)}^{x\wedge N_{i1}}\binom{N_{i1}}{k}q_1^k(1-q_1)^{N_{i1}-k}\binom{N_{i2}}{x-k}q_2^{x-k}(1-q_2)^{N_{i2}-x+k}$$

Then, the maximum likelihood estimate is defined as

$$\hat{\mu}_{ML}=\arg\max_{q\in[0,1]^2}\sum_{i=1}^{K}\ln f_i\left(\sum_{j=1}^{N_i}X_{ij},q\right).$$

The likelihood function can have several points of global maxima. In that case we can select either of them.

To establish asymptotic properties of this estimate we will treat subpopulation sizes in groups (N_{i1}, N_{i2}) as independent random vectors with unknown distribution G. This approach is analogous to the use of structural regression models in which regressors are considered as random variables. It doesn't affect the search of the likelihood extremal points, but simplify the asymptotic considerations, since the law of large numbers is now in force for empirical means of N_{ik} and functions from them.

We have to distinguish two cases:

There are groups with unequal subpopulation sizes, i.e. $\mathrm{P}(N_{i1} \neq N_{i2}) > 0$.

All the groups have equal subpopulations sizes, $N_{i1} = N_{i2}$ *a.s.*

In the first case the likelihood function have unique maximum. The following theorem establishes consistency and asymptotic normality of the maximum likelihood estimate in the first case.

Theorem 3. If the domain sizes have finite first moment $EN_i < \infty$, and $\mathrm{P}(N_{i1} \neq N_{i2}) > 0$ then the maximum likelihood estimate $\hat{\mu}_{ML}$ is consistent as $K \to \infty$. If the domain sizes have finite second moment $EN_i^2 < \infty$ and there is no constant $C > 0$ such that $N_{i1} = CN_{i2}$ *a.s.*, then the maximum likelihood estimate $\hat{\mu}_{ML}$ is asymptotically normal.

In the second case the likelihood function becomes symmetric with respect to the interchange of q_1 and q_2. That means that we can estimate parameters q_1 and q_2 up to a permutation only. Such data arise sometimes in genomic studies, see Scherbina (2011a). Let us constrict the parametric space to $\mu \in \{(q_1, q_2) \mid q_1 \le q_2\}$ and use the following maximum likelihood estimate

$$\hat{\mu}^*_{ML} = \arg\max_{0 \le q_1 \le q_2 \le 1} \sum_{i=1}^{K} \ln f_i\left(\sum_{j=1}^{N_i} X_{ij}, q\right).$$

Theorem 4. Assume that $\mu_1 \le \mu_2$. If the subpopulations sizes have finite first moment $EN_i < \infty$, then the maximum likelihood estimate $\hat{\mu}^*_{ML}$ is consistent as $K \to \infty$. If the domain sizes have finite second moment $EN_i^2 < \infty$ and $\mu_1 < \mu_2$, then $\hat{\mu}^*_{ML}$ is asymptotically normal.

Proofs of Theorems 3 and 4 are based on general theory of ML estimates (see Borovkov, 1998) and are presented in Shcherbina (2011a).

3.4. CDFS based estimates

Although the binomial distribution for X_{ij} described in the previous subsection is very restrictive, this technique can be used for arbitrary distributed variables. Let us fix any $x \in \mathbb{R}$ and replace each value X_{ij} with the indicator $I_{\{X_{ij}<x\}}$. We obtain a sample of binary variables and may use estimates for probabilities of success for two subpopulations described in Section 3.3. Clearly, these probabilities are the cumulative density functions $F_1(x)$ and $F_2(x)$ for first and second subpopulations. Thereby we get estimates $\hat{F}_1(x)$ and $\hat{F}_2(x)$ for them. Since they are not usual empirical CDFs, they can be not monotone. Despite of that, we can estimate μ by

$$\hat{\mu}_{CDF} = \left(\int_{\mathbb{R}} x d\hat{F}_1(x), \int_{\mathbb{R}} x d\hat{F}_2(x) \right).$$

To estimate means we calculate functions $\hat{F}_l(x)$ only at sample points $\{X_{ij}\}$. Let $\{X_i^*\}_{i=1}^N$ be the ordered sequence of observed characteristics, where N is the total sample size. We take $\hat{F}_l(x)$ as peacewise constant with jumps $\hat{F}_l(X_{i+1}^*) - \hat{F}_l(X_i^*)$ in points X_i^* for $i = 1,\ldots,N-1$ and jump $1 - \hat{F}_l(X_N^*)$ in point X_N^*. Now, the integrals can be easily computed:

$$\int_{\mathbb{R}} x d\hat{F}_l(x) = \sum_{i=1}^{N-1} X_i^* \left(\hat{F}_l(X_{i+1}^*) - \hat{F}_l(X_i^*) \right) + X_N^* \left(1 - \hat{F}_l(X_N^*) \right).$$

4. Simulation results

This section provides some simulation studies. Performance of the estimates considered in the previous section is compared on artificial data. We try different distributions for characteristic X, different subpopulation sizes and numbers of groups. Each iteration consists of the following steps:

1. Subpopulation sizes in groups N_{i1} and N_{i2} are generated from some distribution.

2. For each group N_{i1} and N_{i2} independent random variables are generated with distributions that correspond to the first and second subpopulations. Thus, we get the sample $\{X_{ij}, i = 1,2,\ldots,K,\ j = 1,2,\ldots,N_i\}$.

3. Estimates described in the previous chapter are computed.

Then, the mean square errors (MSE) of the estimates are computed. They are multiplied by K in plots.

Example 1. The variable X has Binomial distribution with parameters $\mu_1 = 0.2$ and $\mu_2 = 0.5$. Subpopulation sizes (N_{i1}, N_{i2}) can be $(1,2)$ or $(2,1)$ with equal probability. Number of groups K varies from 10 to 200. Number of simulations is equal 5000.

In the next table the correspondences between estimates and type of lines in plots are shown.

Estimate	Type of lines in plots
Weighted means	Dotted
Weighted least squares	Dashed
Maximum likelihood	Solid

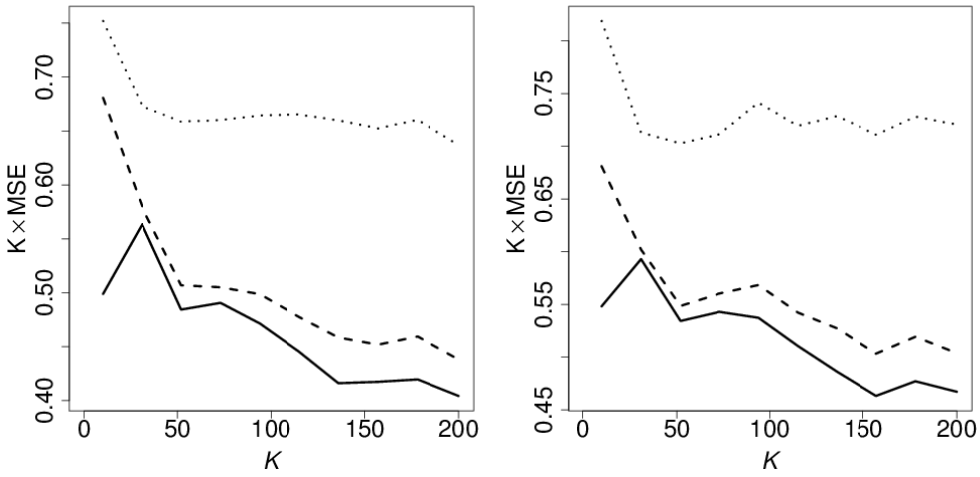

Figure 1. Normalized MSE of estimates for μ_1 and μ_2 for different number of groups K

Example 2. Consider the quality of CDF estimation described in Section 3.4. We took exponential distribution with intensity 1 on U_1 and standard normal distribution on U_2. Group sizes N_i are equal to 5. Size of the first subpopulation N_i^1 is uniformly distributed on $\{1,2,3,4\}$. Number of groups K varies from 20 to 200.

In the Figure 2 the CDFs (thin lines) and their estimates (solid and dashed lines) are represented for both subpopulations. As we can see, the deviations from the true values decrease rapidly when K increases. Although the estimated CDFs are not monotonous, they can be directly used in the estimator $\hat{\mu}_{CDF}$.

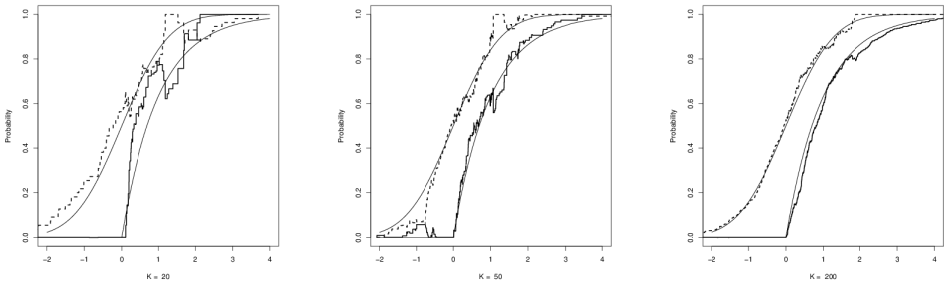

Figure 2. Estimation of CDF by maximum likelihood approach

Example 3. The variable X has Student t-distribution with 3 degrees of freedom and normal distribution with mean 2 and variance 4. Group sizes N_i are equal to 5. Size of the first subpopulation N_i^1 is uniformly distributed on $\{1,2,3,4\}$. Number of groups K varies from 10 to 200. Number of simulations equals to 1000.

In the next table the correspondences between estimates and type of lines in plots are shown.

Estimate	**Type of lines in plots**
Weighted means	Dotted
Weighted least squares	Dashed
CDF' based	Solid

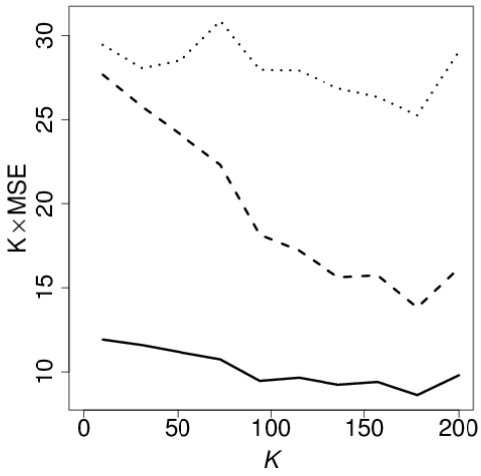

Figure 3. Normalized MSE of estimates for μ_1 for different number of groups K

5. Concluding remarks

The simulation studies indicate that the maximum likelihood estimates are the best in the parametric case, although the weighted least squares estimate is also quite good. Presence of the second order sums in weighted least squares estimates make them much better than the minimax estimates. In the nonparametric case the technique based on CDF estimation seems very promising. It shows the best performance in comparison with the other estimates. Weighted means estimate can be used as pilot estimates of means and variances or as a starting point in numerical calculations of the other estimates.

On the other hand, some caveats should be made about application of these estimates to real data analysis. One expects them to perform satisfactory only if the data satisfy the mixture model on which the estimates are based. Say, the division of the subjects into groups must not depend on their observed variable X. The values of X for subjects belonging to the same group must be independent given the subpopulations distribution in this group. There must be no outer factors shifting the distributions of X for subjects of the same population in different groups. So, a careful choice of the survey design is needed for efficient work of the proposed estimates.

Say, in our cheating example we may expect independence of the first year students' previous school marks and their distribution by academic groups if the groups were formed in random by the Dean's office. But it is not the case when we analyze second year students' current marks, since they may be dependent on their common experience which is different in different groups. In this case one needs more ingenious experiment design, e.g. using surveying of random groups of students attending a gym.

Despite these restrictions we hope that the proposed estimates will be useful in sensitive questions statistics and other problems connected with merging anonymous and non-anonymous surveys information.

Aknowlegements

The authors are thankful to the Referee and Editor for the fruitful discussion.

REFERENCES

BOROVKOV A.A. (1998). *Mathematical statistics,* Gordon and Breach Science Publishers, Amsterdam.

CHRISTOPHER C. HEYDE (1997) *Quasi-Likelihood And Its Application: A General Approach to Optimal Parameter Estimation,* Springer.

CHRISTOPHER G. SMALL, JINFANG WANG, (2003), *Numerical Methods for Nonlinear Estimating Equations*, Oxford.

COUTTS E. & JANN B. (2011), *Sensitive Questions in Online Surveys:Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)*, Sociological Methods & Research, 40, 169-193.

KERKVLIET J. (1994) Cheating *by economics students: A comparison of survey results.* The Journal of Economic Education, Vol. 25, No. 2, p.121-133.

MAIBORODA R. (1996) *Estimates for distributions of components of mixtures with varying concentrations.* Ukrainian Mathematical Journal, 48(4), 618-622.

MAIBORODA R. (1999) A*n asymptotically effective probability estimator constructed from observations of a mixture.* Theory Probab. Math. Stat. 59, 121-128

MAIBORODA R. & SUGAKOVA O. (2008) *Estimation and classification by observations from mixtures*, Kyiv University Publishers, Kyiv (in Ukrainian).

McLACKLAN G. J., PEEL D. (2000). *Finite Mixture Models*, Wiley, New York.

ONG A.D. and WEISS D.J. *The Impact of Anonymity on Responses to Sensitive Questions. Journal of Applied Social Psychology*. Volume 30, Issue 8, p. 1691–1708.

SHCHERBINA A. (2011) *Mean value estimation in the model of mixture with varying concentrations*. Teor. Imovir. Ta Matem. Statyst., No. 84, pp. 142-154. (In Ukrainian, English translation to appear in Theory Probab. Math. Stat).

SHCHERBINA A. (2011a) *Estimation of parameters of binomial distribution in mixture model*. Teor. Imovir. Ta Matem. Statyst. (In Ukrainian, English translation to appear in Theory Probab. Math. Stat).