Finite mixtures analysis in survey sampling problems

Shcherbina, Artem
Dept. of Probability Theory, Statistics and Actuarial Mathematics, Mechanics and Mathematics Faculty, National Taras Shevchenko University, Volodymyrska str., Kyiv, 01033, Ukraine
E-mail: artshcherbina@gmail.com

Maiboroda, Rostyslav
Dept. of Probability Theory, Statistics and Actuarial Mathematics, Mechanics and Mathematics Faculty, National Taras Shevchenko University, Volodymyrska str., Kyiv, 01033, Ukraine
E-mail: mre@univ.kiev.ua

Sugakova, Olena
Dept. of Probability Theory, Statistics and Actuarial Mathematics, Mechanics and Mathematics Faculty, National Taras Shevchenko University, Volodymyrska str., Kyiv, 01033, Ukraine
E-mail: sugak@univ.kiev.ua

Introduction

Observations of mixtures of different subpopulations are common in biological and sociological studies. For such purposes 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.

Such problems arise in analysis of data of sociological surveys concerning so called "sensitive questions", e.g. problems of drugs usage, school tests cheating and so on. Anonymous survey is usually used to avoid inadequate answers on the questions. answers on such questions. On the other hand, it is interesting to compare the obtained anonymous information on the proportion, say of cheaters in different groups of anonymous respondents to open information on their individual features, such as age, school marks, gender, etc. In this example one considers two subpopulations of cheaters and non-cheaters and estimate mean characteristics over these subpopulations.

Another example is an analysis of genetic and phenotype information in genomic imprinting studies.

For such purpose we consider some nonparametric estimates of the subpopulation means, such as weighted means with minimax and adaptive weights. Finite sample properties and asymptotic behaviour of these estimates are discussed. They are compared to maximum likelihood estimates for some parametric submodel.

Model description

To analyse such data finite mixture model (FMM) will be used. In classical FMM one observes a set of independent random vectors (variables) $X_1, \dots, X_K$ and the distribution of X_i is a mixture of L different probabilistic distributions:

$$\mathbf{P}\{Z_j \in A\} = p_1 H_1(A) + p_2 H_2(A) + \dots + p_L H_L(A), \tag{1}$$

where H_l, $l = 1, \dots, L$ is the distribution of observed variables for the units from the l-th component of the mixture, p_l is the probability to observe a unit from the l-th component (the mixing probability, the concentration of the l-th component in the mixture), A is any measurable subset of the observations space. In this model all concentrations are constant. Analysis of of FMMs was started by Newcomb [4]

and Pearson [5]. For recent results see McLachan and Pell [3].

Finite mixtures with varying concentrations were considered by Maiboroda, Sugakova and others [1, 2, 6] for independent observations.

In this paper we consider a case of dependent observations. Dependency arises due to a sampling scheme with replacement. We consider two classes of subjects, which are distributed over K groups, so that i–th group contains N_i^1 subjects of first class and N_i^2 subjects of second class. Total number of subjects in i–th group we will denote by N_i. We will treat group sizes as nonrandom known values.

Now, a random subject from group i will belong to class l with probability

$$w_i^l = \frac{N_i^l}{N_i}.$$

These probabilities will be called concentrations.

Let us denote by X some numeric characteristic of interest associated to each subject. We assume that values of X are generated by some stochastic mechanism independently for all subjects in the considered population. The distribution of X for subjects from one class is the same, but it is different in different classes (subpopulations).

In this paper we consider only the mean value estimation of X for subjects from both classes. Denote the mean value and variance of characteristic X for subjects from the class l by μ_l and σ_l respectively.

The observed sample contains values of X for n_i subjects selected from the i–th group ($i = 1, \dots, K$) by simple random sampling without replacement. Denote the value of characteristic X of the j–th subject from the i-th group in the sample by X_{ij}.

We will consider different estimators for mean value of characteristic X for both classes.

Linear estimate

Let us consider the following estimate for mean value of class l:

$$\hat\mu(a^l) = \frac{1}{K}\sum_{i=1}^{K} a_i^l T_i,$$

where T_i is mean value of characteristics in group i:

$$T_i = \frac{1}{n_i}\sum_{j=1}^{N_i} X_{ij},$$

and $a^l = (a_1^l, \dots, a_K^l)$ is some set of coefficients. Without lose of generality further we will consider estimation only for the first class and omit index 1 at a^1.

It can be shown, that the estimate for the first class $\hat\mu_1(a)$ is unbiased under the following conditions:

$$\frac{1}{K}\sum_{i=1}^{K} w_i^1 a_i = 1, \quad \frac{1}{K}\sum_{i=1}^{K} w_i^2 a_i = 0. \tag{2}$$

Such coefficients exist in case of not all concentrations are equal. To select the best set of coefficients we are to minimize the variance of $\hat\mu_1(a)$:

$$\mathbf{D}\,\hat\mu_1(a) = \frac{1}{K^2}\sum_{i=1}^{K}\frac{a_i^2}{n_i}\left(w_i^1\sigma_1^2 + w_i^2\sigma_2^2 + \frac{N_i - n_i}{N_i - 1} w_i^2 w_i^1 (\mu_1 - \mu_2)^2\right),$$

It depends on unknown population parameters, namely means and variances for both classes. In such case we can use minimax coefficients, introduced in Maiboroda [2]:

$$\bar{a}_{i,K} = \frac{(1 - \bar{r}_{1,K})w_i^1 + \bar{r}_{2,K} - \bar{r}_{1,K}}{\bar{r}_{2,K} - \bar{r}_{1,K}^2}, \quad i = 1, \dots, K.$$

where $\bar{r}_{j,K}$ — moments of concentrations w_i^1:

$$\bar{r}_{j,K} = \frac{1}{K}\sum_{i=1}^{K}(w_i^1)^j, \quad j = 1, 2.$$

They fulfil conditions (2) and minimize the sum of squares $\sum_{i=1}^{K} a_i^2$. Although minimax coefficients do not minimize the variance of $\hat{\mu}_1(a)$, 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 $C > 0, M \geq 1$, such that $\bar{r}_{2,K} - \bar{r}_{1,K}^2 > C$ for all K and $n_i \leq M$ for all $i \geq 1$. Then the estimate $\bar{\mu}_{1,K} = \hat{\mu}(\bar{a}_K)$ is consistent and distributions of*

$$\frac{1}{\sqrt{\mathbf{D}\,\bar{\mu}_{1,K}}}(\bar{\mu}_{1,K} - \mu_1)$$

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

To obtain the best coefficients a_i which minimize the variance of the estimate, rewrite expression for variance in the following form:

$$\mathbf{D}\,\hat{\mu}_1(a) = \frac{1}{K^2}\sum_{i=1}^{K} a_i^2 d_i,$$
$$d_i = \frac{1}{n_i}\left(w_i^1\sigma_1^2 + w_i^2\sigma_2^2 + \frac{N_i - n_i}{N_i - 1}w_i^2 w_i^1(\mu_1 - \mu_2)^2\right), \quad i = 1, \dots, K.$$

Considering conditions (2), the best coefficients can be found by following formula:

$$\tilde{a}_{i,K} = \frac{(\tilde{r}_{0,K} - \tilde{r}_{1,K})w_i^1 + \tilde{r}_{2,K} - \tilde{r}_{1,K}}{d_i(\tilde{r}_{2,K}\tilde{r}_{0,K} - \tilde{r}_{1,K}^2)}, \quad i = 1, \dots, K,$$

where $\tilde{r}_{j,K}$ are weighted moments of concentrations w_i:

$$\tilde{r}_{j,K} = \frac{1}{K}\sum_{i=1}^{K}\frac{(w_i^1)^j}{d_i}, \quad j = 0, 1, 2.$$

More detailed analysis of such estimates and proofs of theorems can be found at Shcherbina [7].

Theorem 2 *In case of all conditions of Theorem 1 are fulfilled, the estimate $\tilde{\mu}_{1,K}$ is consistent and distributions*

$$\frac{1}{\sqrt{\mathbf{D}\,\tilde{\mu}_{1,K}}}(\tilde{\mu}_{1,K} - \mu_1)$$

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

But the coefficients $\tilde{a}$ cannot be used for the estimation, since the values d_i depend on the vector of unknown parameters $\gamma = (\mu_1, \mu_2, \sigma_1, \sigma_2)$. Therefore using the minimax coefficients we compute a pilot estimate $\bar{\gamma}_K = (\bar{\mu}_{1,K}, \bar{\mu}_{2,K}, \bar{\sigma}_{1,K}, \bar{\sigma}_{2,K})$:

$$\bar{\mu}_{l,K} = \frac{1}{K}\sum_{i=1}^{K} \bar{a}^l_{i,K} \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \quad l = 1, 2, \tag{3}$$

$$\bar{\sigma}^2_{l,K} = \frac{1}{K}\sum_{i=1}^{K} \bar{a}^l_{i,K} \frac{1}{n_i}\sum_{j=1}^{n_i} Y^2_{ij} - \bar{\mu}^2_{l,K}, \quad l = 1, 2. \tag{4}$$

Then, the estimate for variance in group i equals

$$\bar{d}_{i,K} = \frac{1}{n_i}\left(w^1_i \bar{\sigma}^2_{1,K} + w^2_i \bar{\sigma}^2_{2,K} + \frac{N_i - n_i}{N_i - 1} w^1_i w^2_i (\bar{\mu}_{1,K} - \bar{\mu}_{2,K})^2 \right).$$

The estimate for weighted moments equals

$$\bar{r}_{i,K} = \frac{1}{K}\sum_{i=1}^{K} \frac{(w^1_i)^j}{\bar{d}_{i,K}}, \quad i = 1, \ldots, K.$$

Now, adaptive coefficients $\hat{a}_{i,K}$ are defined as

$$\hat{a}_{i,K} = \frac{(\bar{r}_{0,K} - \bar{r}_{1,K}) w^1_i + \bar{r}_{2,K} - \bar{r}_{1,K}}{\bar{r}_{0,K}\bar{r}_{2,K} - \bar{r}^2_{1,K}}, \quad i = 1, \ldots, K.$$

Hence, the adaptive estimate of mean will be

$$\hat{\mu}_{1,K} = \hat{\mu}(\hat{a}_K) = \frac{1}{K}\sum_{i=1}^{K} \hat{a}_{i,K} T_i.$$

A valuable property of the adaptive estimate is its asymptotic normality with the same asymptotic variance as for the best estimate.

Theorem 3 *In case of all conditions of Theorem 1 are fulfilled, the estimate $\hat{\mu}_{1,K}$ is consistent and distributions*

$$\frac{1}{\sqrt{\mathbf{D}\,\tilde{\mu}_{1,K}}}(\hat{\mu}_{1,K} - \mu_1)$$

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

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

Maximum likelihood estimate

To use the maximum likelihood approach we need some parametric model of the data distribution. Let us consider the simplest case of Bernoully distribution of X in both classes with different probabilities of success. I.e. X attains only the values 1 (success) or 0 (failure) with probability of success $\mathbf{P}\{X = 1\} = q_l$ for subjects, which belong to l-th class. The unknown parameter of this model is $q = (q_1, q_2)$.

We will treat group sizes (N_{i1}, N_{i2}) as independent random vectors with unknown distribution

$$G(n_1, n_2) = \mathbf{P}\left(N_{i1} = n_1, N_{i2} = n_2\right), \quad n_1, n_2 \in \mathbb{N}_0.$$

Let us consider total value for group i:

$$X_i = \sum_{j=1}^{N_i} X_{ij}.$$

It can be shown, that statistics $S_i = (X_i, N_{i1}, N_{i2})$ is sufficient for parameter $q = (q_1, q_2)$ estimation. Let us consider likelihood function

$$L(S,t) = \sum_{i=1}^{K} \ln f(S_i, t) = \sum_{i=1}^{K} l(S_i, t), \tag{5}$$

where $l(S_i, t) = \ln \mathbf{P}_t \left(X = X_i, N_1 = N_{i1}, N_2 = N_{i2}\right)$.

Then the maximum likelihood estimate of q is

$$\hat{q} = \underset{t \in [0,1]^2}{\operatorname{argmax}} L(S,t).$$

The following theorems establish consistency and asymptotic normality conditions of this estimate.

Theorem 4 *Let group sizes have finite mathematical expectations* $\mathbf{E}_q N_1 < \infty$. *If* $\mathbf{P}_q \left(N_{11} = N_{12}\right) < 1$ *or* $q_1 \leq q_2$, *then maximum likelihood estimate is strongly consistent.*

Theorem 5 *Let group sizes have finite second order moments* $\mathbf{E}_q N_1^2 < \infty$, *the true value of the parameter* $q \in (0,1)^2$, *and one of the following conditions is fulfilled*

1. $\mathbf{P}_q \left(N_{11} = N_{12}\right) < 1$,

   or

2. $q_1 < q_2$,

   or

3. $q_1 = q_2$ *and there is no constant* $C > 0$, *such that* $N_{11} = C N_{12}$ *a.s.*

Then the maximum likelihood estimate is asymptotically normal.

The advantage of this model is ability to provide estimates in case of constant concentrations. In the special case of equal sizes ($N_{i1} = N_{i2}$) model becomes simmetric and estimation of the model parameter (q_1, q_2) is possible only up to a permutation.

Although this parametric model is very restrictive, it is possible to use it for nonparametric estimation of means also. Really, let X_{ij} be arbitrary random variables. Then for any real constant C we may consider indicators $\mathbf{1}_{\{X_{ij} < C\}}$. Now we get our parametric case and compute estimates for probabilities of success for two classes. These probabilities correspond to distribution functions for two classes at point C. Further, when distribution functions are known, we can estimate all required characteristics.

Genomic imprinting data analysis

Let us consider the special case of parametric model, discussed in the previous section. Let each group consists of two elements, one from first and one from second classes, i.e. $G(1,1) = 1$.

This model can be useful for studying the genomic imprinting phenomenon. Genomic imprinting is a genetic phenomenon by which certain genes are expressed in a parent-of-origin-specific manner.

Let us consider expression of some gene of interest in organisms which are homozygous for this gene. We will assume that phenotypic features of the i-th considered organism allow us to conclude whether both alleles ($X_i = 2$), one ($X_i = 1$) or no allele ($X_i = 0$) of the gene are expressed in this

organism. In this case each organism can be considered as a group, that consists of two subjects: paternal and maternal chromosomes. Let X_{i1} (X_{i2}) be 1 if the paternal (maternal) allele is expressed in the i-th organism and zero otherwise. Then $X_i = X_{i1} + X_{i2}$. Although the values of X_{ij} are not observable separately, we can use the sufficient statistics X_i to construct maximum likelihood estimates for probabilities q_l of gene expression from the l–th type allele.

Let

$$f_i(t) = f(i,1,1,t) = g(i,1,1,t), \quad i = 0,1,2.$$

Then the likelihood can be represented as

$$L(S,t) = \sum_{i=1}^{K} \ln f_{X_i}(t) = K \sum_{l=0}^{2} \nu_l \ln f_l(t),$$

where ν_l — is the frequency of values l in sample $X_1, \ldots, X_K$.

In this case the maximum likelihood estimate can be written in the closed form:

$$\hat{q} = \left(\frac{\mu}{2} - \sqrt{\frac{\mu^2}{4} - \nu_2}, \frac{\mu}{2} + \sqrt{\frac{\mu^2}{4} - \nu_2} \right),$$

if $\nu_2 \le \mu^2/4$, and $\hat{q} = (\mu/2, \mu/2)$ otherwise.

According to Theorem 4 the estimate $\hat{q}$ is consistent. According to Theorem 5 it is asymptotically normal, if $0 < q_1 < q_2 < 1$. Note that if $q_1 > q_2$ one can enumerate the chromosomes in the reverse order, so the probabilities of the expression can be estimated for alleles of both types. But we are not able to identify what probability corresponds to the paternal or maternal chromosome in such case.

The asymptotic covariance matrix equals to the inverse of the information matrix

$$I = \frac{1}{q_1 + q_2 - 2q_1q_2} \begin{pmatrix} \dfrac{q_1 + q_2^2 - 2q_1q_2}{q_1 - q_1^2} & 1 \\ 1 & \dfrac{q_1^2 + q_2 - 2q_1q_2}{q_2 - q_2^2} \end{pmatrix}.$$

REFERENCES

1. Kubaychuk O. O. Estimation of moments by observations from mixtures with varying concentrations. Theory of Stochastic Processes, 2002, Vol. 8(24), N.3–4.— p. 226–232.

2. Maiboroda R., Statistical Analysis of Mixtures. Kyiv University, Kyiv, 2003. (Ukrainian).

3. McLachan G. J., Pell D. Finite Mixture Models. NY, Wiley, 2000.

4. Newcomb S. A generalized theory of the combination of observations so as to obtain the best result. Amer. J. Math., 1894, V.8.— p. 343–366.

5. Pearson K. Contribution to the mathematical theory of evolution. Trans. Roy. Sos. A. 1894, v. 185.— p. 71–110.

6. Shcherbina A., Maiboroda. R. Merging data from anonymous and open surveys: two-population problems. Proceedings of the Baltic–Nordic–Ukrainian Summer School on Survey Statistics, Kyiv, TBiMC, 2009.— 177 p.

7. Shcherbina A. Mean value estimation in the model of mixture with varying concentrations. Submitted to Probability and Mathematical Statistics, 2011.

8. Titterington D.M., Smith A.F.M., Makov U.E. Statistical analysis of finite mixture distributions. Wiley, New York, 1985, 243 p.