International Institute for
Applied Systems Analysis
IIASA www.iiasa.ac.at

Structural Minimization of Risk on Estimation of Heterogeneity Distributions

Michalski, A. and Yashin, A.I.

IIASA Working Paper

WP-86-076

December 1986

Working Paper

Structural Minimization of Risk in Estimation of Heterogeneity Distributions

Anatoli Michalski
Anatoli Yashin

December 1986
WP-86-76

International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria

NOT FOR QUOTATION
WITHOUT THE PERMISSION
OF THE AUTHORS

Structural Minimization of Risk in Estimation of Heterogeneity Distributions

Anatoli Michalski
Anatoli Yashin

December 1986
WP-86-76

Working Papers are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute or of its National Member Organizations.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

Foreword

Population heterogeneity dynamics is one of the research directions in IIASA's Population Program. One typical and practical problem related to hidden heterogeneity is the estimation of the heterogeneity distribution.

This paper describes the approach to such an estimation which is based on the method of structural minimization of mean risk. It is shown how this method can be implemented to some real data. The main ideas of the method are also described.

Anatoli Yashin
Deputy Leader
Population Program

Contents

Structural Minimization of Risk in Estimation of Heterogeneity Distributions

Anatoli Michalski, Anatoli Yashin***

1. Introduction

Assume that there are two random variables z and T and both marginal distribution of T and conditional distribution of T given z are known. What can one say about the distribution density of z?

The version of this problem is known in econometrics and demography: T is interpreted as a random duration or death time, z is the latent (heterogeneity) variable which characterizes the individual's differences in susceptibility to transitions or death [1,2,3].

Denote by $f(z)$, $U(t)$ the probability density functions of random variables z and T respectively and by $k(t \mid z)$ the conditional distribution density function of T given z. We assume that all these densities exist.

It is easy to see that functions $U(t)$, $k(t \mid z)$ and $f(z)$ are related as follows

$$U(t) = \int k(t \mid z) f(z) dz \quad . \tag{1}$$

Formula (1) is the first kind integral Fredholm equation with respect to function $f(z)$ with kernel function $k(t \mid z)$. To find $f(z)$ when $U(t)$ and $k(t \mid z)$ are given means to solve the integral equation (1) with respect to $f(z)$. It turns out that the solution of this equation is unstable. It means that small disturbances in kernel function can produce big changes in $f(z)$. Moreover, if in addition the kernel function is also unknown then equation (1) can have a non-unique solution.

The last property has the important consequences for applications. It means, for instance, that one should use maximum ancilliary information to specify the kernel function $k(t \mid z)$ as precise as possible before the data processing.

*Anatoli Michalski, Institute of Control Sciences, Profsojusnaja 65, Moscow, USSR
**Anatoli Yashin, Population Program, IIASA, A-2361 Laxenburg, Austria

Another important remark is that in applications one usually does not have the precise knowledge of the distribution density $U(t)$. The typical information which come out of, say, clinical studies are the observed death times for a sample of n individuals. It is clear that such circumstances can only complicate the estimation problem of $f(z)$.

Recently many publications were devoted to the problems of modeling and estimation of heterogeneity in population analysis using other approaches. Shepard and Zeckhauser [4] showed that heterogeneity could be responsible for overestimates of the results of medical improvements. Keyfitz and Littman [5] demonstrated that ignoring heterogeneity leads to incorrect calculations of life expectancy. Vaupel and Yashin [2,3] described many paradoxes and puzzles which can be explained using the heterogeneity concept. Heckman and Singer [1] considered the identification problem in econometric models for duration data both for parametric and nonparametric cases. They have found in particular that the estimates of the model for duration data are sensitive to the assumptions about heterogeneity models. Manton et al. [6] came to the similar conclusion.

One idea which is discussed in our paper deals with the nature of such sensitivity. It turns out that very often the identification in the presence of hidden heterogeneity is an ill-posed problem, related to the solution of equation (1).

Some properties of this equation which are relevant to our study are discussed in chapter 2. In chapters 3 and 4 we describe the approach to the solution of equation (1) given the information about n death times. Chapter 3 focuses on the analysis of artificial data which were generated by the models of heterogeneous mortality. Chapter 4 demonstrates the results of the application of the developed approach to the real data. In both chapters the data processing algorithms were based on so-called structural minimization of mean risk approach. The main ideas and results of this approach are given in the Appendix.

2. Estimation of Hidden Heterogeneity as an Ill-Posed Problem

The three major mathematical problems are related to equation (1). The first is about when the solution of this equation exists. The second is about whether the solution is unique. The third is about how sensitive is the solution to the disturbances of the function $U(t)$.

In this paper we will not analyze the first problem, referencing publication [7] for those who are interested in a deeper understanding of the existence conditions. The nonunicity problem will be demonstrated in a particular case. Since the sensitivity problem is very important for the data analysis we will focus our main attention in this paper on this problem.

Let us consider an example of an ill-posed problem which can arise in demographic applications. Assume that the conditional density (kernel function) can be represented in the form

$$k(t\mid z) = z\lambda(t)\, exp(-z \int_0^t \lambda(s)ds) \quad . \tag{2}$$

It is well known that if the kernel function is smooth, then slight variations of $U(t)$ can produce the big changes in $f(z)$ [8]. One can see that if $\lambda(t)$ in (2) is smooth, then one can expect instability in the solution of equation (1).

The conditional density function $k(t\mid z)$ given by (2) corresponds to the well-known proportional hazard model of mortality, where z is a heterogeneity variable and $\lambda(t)$ is the underlined hazard. Assume that $\lambda(t) = \alpha\lambda_0(t)$ where α is some scale parameter. Let us show that for different values of α one can find the different solutions of the integral equation (1).

The equation (1) now will be

$$U(t) = \int z\alpha\lambda_0(t) exp(-z\alpha \int_0^t \lambda_0(\tau)d\tau) f(z)dz \tag{3}$$

where $U(.)$ is a density function for observed survival times.

Denote by $f_1(z)$ the solution of (3) for the case $\alpha = 1$. For any other value of α one can write

$$U(t) = \int u\lambda_0(t) exp(-u \int_0^t \lambda_0(\tau)d\tau) \frac{f(u/\alpha)}{\alpha}\, du \quad ,$$

where $u = \alpha z$. Since for any fixed α equation (3) has a unique solution one may write

$$f_1(z) = \frac{1}{\alpha} f(z/\alpha) \quad .$$

The relation between solution of (3) and function $f_1(.)$ follows from the next expression

$$f(z) = \alpha f_1(\alpha z) \quad .$$

The last statement shows, that using different values for parameter α we have different shapes for density of hidden heterogeneity variable, i.e., the solution of (3) is not unique when α is unknown.

3. Estimation of Hidden Heterogeneity

In this chapter the new approch to the solution of equation (3) is considered. The approach takes into account the instability property of the solution of equation (3) and the lack of information about the distribution density $U(t)$. The method is based on the structural minimization of mean risk. The ideas of this approach are outlined in the Appendix. To implement these ideas consider the family of functions $\{Q(x)\}$ where

$$Q(x) = -ln\left[\int_0^\infty z\lambda(t)\exp(-z\int_0^x \lambda(s)ds)f(z)dz\right] \tag{4}$$

and $f(z)$ is some distribution density function of z with $\lambda(t) > 0$. Let us take the mean risk functional in the form

$$G = -\int Q(x)U(x)dx \quad .$$

General theory of structural minimization of mean risk considers mean risk functional with nonnegative loss function $Q(x)$. In our case it is not so. However, assuming that the distribution of z is concentrated on a finite interval one can always add some positive constant to all functions from this family and make them positive without changing the optimal point of the functional.

The functional G with such $Q(x)$ is the particular case of so-called mixed entropy functional. It takes its minimal value on the solution of equation (1). The empirical risk functional will be as follows

$$G_L = -\frac{1}{L}\sum_{n=1}^{L} \ln\left[\int_0^\infty z\lambda(x_n)exp(-z\int_0^{x_n} \lambda(s)dsf(z)dz\right] \tag{5}$$

which coincides with the minus likelihood functional.

As a first example let us consider the families of functions $\{Q_i\}$ in the form of (4) where the functions $f_i(z)$ are supposed to be a histogram

$$f_i(z) = \frac{1}{i} \sum_{k=1}^{i} \alpha_{k,i} H_{k,i}(z) \tag{6}$$

where $\alpha_{k,i} \geq 0$, $\sum_{k=1}^{i} \alpha_{k,i} = 1$, and $H_{k,i}(z)$ are the step functions equal to $\frac{1}{z_{k+1,i} - z_{k,i}}$ when $z_{k,i} \leq z < z_{k+1,i}$ and equal to 0 otherwise, $z_{k,i}$, $k = 1,2,\dots,i$ are fixed points $z_{1i} = 0$, $z_{i+1,i} = 1$, $\alpha_{k,i}$ are the parameters of the histogram, i is the number of the parameters.

We used the values $z_{k,i} = (k-1)/i$ for creating the histogram. One can use any other set of $z_{k,i}$ if there is information on subinternational inside [0,1] where density function $f(z)$ changes fast. If there is no such preliminary information, then one should use equidistant points $z_{k,i}$.

The histogram approximation of densities is widely used in statistical practice. It presupposed the finiteness of the possible values of z. The number of intervals of the histogram will be determined during the structural minimization of risk procedure. We assume that the distributions $f(z)$ are all defined on the interval [0,1]. This interval can be changed if one has preliminary information on where the distribution of z is concentrated.

It is important to emphasize that we do not assume real distribution of heterogeneity parameter to be in form (6). Expression (6) gives only an approximation of real distribution and to implement structural risk minimization method we don't need to know the precise form of this distribution.

Now it is easy to construct functional families $\{Q_j\}$ by changing the number of parameters i in (6). So family $\{Q_1\}$ will be given by functions

$$f_1(z) = \alpha_{11} H_{11}(z) + \alpha_{21} H_{21}(z) \tag{7}$$

family $\{Q_2\}$ will be given by expression (8)

$$f_2(z) = \alpha_{12} H_{12}(z) + \alpha_{22} H_{22}(z) + \alpha_{32} H_{32}(z) \tag{8}$$

and so on. We will use the uniform greed $z_{1i}, z_{2i}, \dots, z_{ii}$ for which $(z_{k+1,i} - z_{k,i}) = 1/i$. In the case if one has more information on heterogeneity distribution, one can use other special greeds with different knots. The only thing is important that the grid is to be fixed before one starts to implement the structural risk minimization method, because the inequality (A5) in the Appendix is valid only in this case. If one will try to fit the greed to the experimental data, than one can have wrong result.

Substituting (6) into (5), one can see that in every family Q_i one is to minimize the functional

$$G_L = -\sum_{n=1}^{L} \ln\left[\frac{1}{i}\sum_{k=1}^{i}\frac{\alpha_{k,i}}{\Delta z_{k,i}}\frac{\lambda(x_n)}{\beta^2(x_n)}(e^{-\beta(x_n)z_{k,i}} - e^{-\beta(x_n)z_{k+1,i}} \right. \tag{9}$$

$$\left. + \beta(x_n)(z_{k,i}e^{-\beta(x_n)z_{k,i}} - z_{k+1,i}e^{-\beta(x_n)z_{k+1,i}}))\right]$$

where

$$\Delta z_{k,i} = z_{k+1,i} - z_{k,i} \quad ,$$

$$\beta(x_n) = \int_0^{x_n} \lambda(s)ds \quad ,$$

where $z_{k,i}$ are the knots in the greed for (6).

Following the structural minimization of mean risk approach one should minimize the functional of empirical risk (5), then compare the values of the functionals

$$B_i = -\frac{1}{L}\frac{\sum_{j=1}^{L} \ln \int_0^{\infty} z\lambda(t)exp(-z\int_0^{x_j}\lambda(s)ds)f_i^*(z)dz}{(1 - \sqrt{\frac{K_i(ln(L/K_i) + 1 - \ln q}{L}}}$$

for different i and choose the minimal value of B_i. Here $f_i^*(z)$ denotes the histogram constructed by minimizing functional (5) in the family of histograms with i parameters.

As a second example let us consider the situation when preliminary information is available on the heterogeneity distribution. Assume that heterogeneity variable z can take the finite number of known values. One needs to estimate the respective probabilities observing a sample of survival times $x_1, x_2, \ldots, x_L$. This approach corresponds to the case when the population under investigation consists of a finite number of homogeneous subgroups and we know the values of heterogeneity variable for each of these subgroups. This situation is simpler than above but it is relevant for many practical situations. In real life we can have information about surviving in, say, genetic subgroups and we may be interested in proportions of these subgroups in the total population.

To use our method for this case we rewrite expression (4) in the form

$$Q(x) = -ln\,(\sum_{j=1}^{i} Z_j P_j \lambda\, exp\,(-Z_j \int_0^x \lambda d\tau))$$

where $P_j = P(z = z_j)$.

As a matter of fact now we estimate not function but some numbers and instead of functional family $\{Q\}$, now one can use just i dimensional vector space, where i is number of fixed groups minus 1 because the sum of P_j is to equal 1.

Now one can check different hypotheses about subgroups in total population. When we consider different numbers of groups we have different families and minimizing expression

$$B_i' = -\frac{1}{L}\,\frac{\sum_{n=1}^{L} \ln(\sum_{j=1}^{i} Z_j P_j \lambda\, \exp(-Z_j \int_0^{x_n} \lambda d\tau))}{1 - \sqrt{\frac{i\,(ln\,(L\,/\,i) + 1) - \ln q}{L}}}$$

on proportion P_j and number of groups i we will find the best suitable number of subgroups and proportions for them.

To demonstrate the power of the method, we performed calculations with samples, generated with known probabilistic distributions. We considered the continuous distribution of heterogeneity variable with density function

$$f\,(z) = 1/\,(\beta z) \quad , \quad z \in [e^{-\beta},1]$$

where β is some known parameter. The density function corresponds to the case when the heterogeneity variable can be expressed in the form used in Cox's model [9]

$$z = e^{-\beta U}$$

and U is a random variable with uniform distribution on the interval [0,1]. For both examples the numerical calculations were provided.

In the first case we estimated the continuous density ($f\,(z)$) by histogram. The number of parameters in the histogram was determined on a given sample by the method described above. Typical estimate of continuous distribution ($f\,(z)$) is shown in Chart 1. In Table 1 we put the value of parameter β, sample size L, determined number of parameters in histogram i, probability of every subinterval in correspondence with $f(z)$, P, and estimated probability of every subinterval in correspondence with the histogram $\hat{P}$.

Table 1.

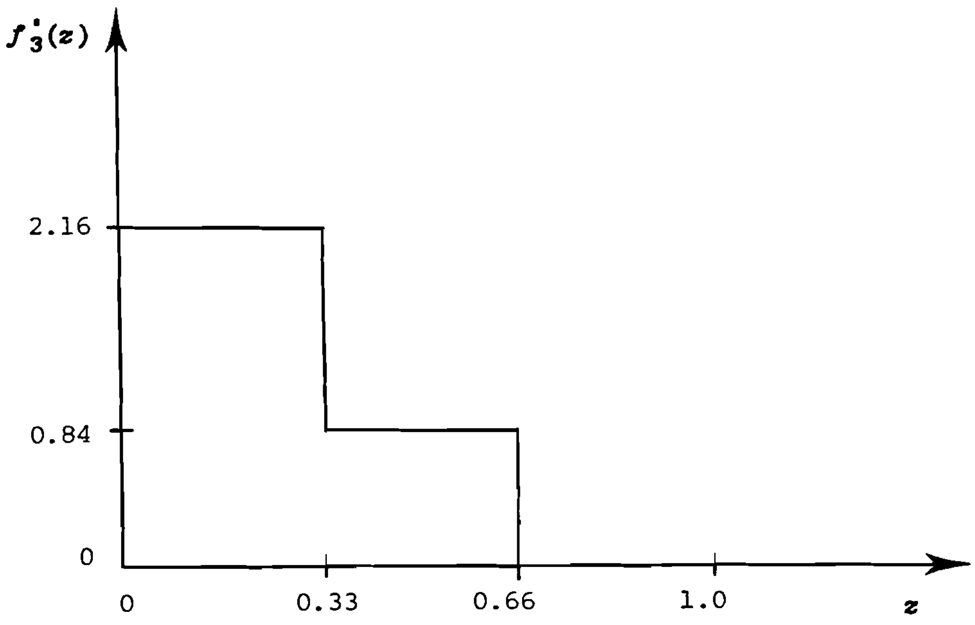
β	L	i	P	$\hat{P}$
1	50	3	0.45	0.41
			0.31	0.29
			0.24	0.30
	100	3	0.45	0.48
			0.31	0.36
			0.24	0.16
	300	5	0.29	0.33
			0.23	0.25
			0.19	0.13
			0.16	0.15
			0.13	0.14
3	50	3	0.66	0.69
			0.21	0.31
			0.12	0.00
	100	3	0.6	0.72
			0.21	0.28
			0.12	0.00
	300	5	0.52	0.50
			0.20	0.23
			0.13	0.15
			0.09	0.12
			0.06	0.00
9	50	4	0.85	1.00
			0.08	0.00
			0.04	0.00
			0.03	0.00
	100	5	0.82	0.88
			0.08	0.12
			0.04	0.00
			0.03	0.00
			0.02	0.00
	300	6	0.80	0.79
			0.08	0.21
			0.05	0.00
			0.03	0.00
			0.03	0.00
			0.01	0.00

Table 2.

N	L	P	$\hat{P}$
2	50	0.70	0.65
	100	0.70	0.71
		0.30	0.29
	300	0.70	0.70
		0.30	0.30
4	50	0.40	0.44
		0.30	0.32
		0.20	0.18
		0.10	0.05
	100	0.40	0.38
		0.30	0.31
		0.20	0.21
		0.10	0.10
	300	0.40	0.40
		0.30	0.30
		0.20	0.21
		0.10	0.09
6	50	0.30	0.24
		0.25	0.20
		0.20	0.22
		0.15	0.18
		0.05	0.06
		0.05	0.10
	100	0.30	0.34
		0.25	0.22
		0.20	0.19
		0.15	0.10
		0.05	0.08
		0.05	0.07
	300	0.30	0.33
		0.25	0.28
		0.20	0.20
		0.15	0.13
		0.05	0.02
		0.05	0.04

From Table 1 one can see that the larger the sample size, the better the estimation, but even in the case of small sample one still has a good estimation.

Chart 1.



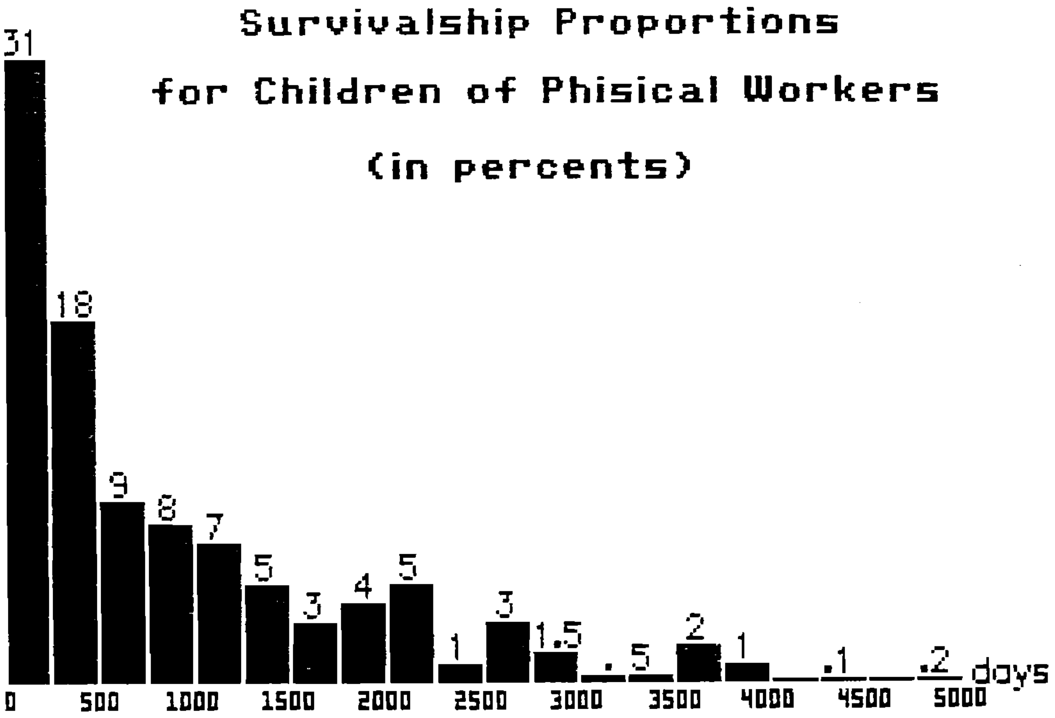
In the case of mix distribution when heterogeneity variable may have only fixed values we estimated probabilities of these values, or proportions between different states of heterogeneity variable. In Table 2 we put number of subgroups in population N, sample size L, real proportions P, and estimates $\hat{P}$.

Here again one can see that the larger the sample size, the better the estimation, but in small sample case the estimate is good either.

4. Experiments with Real Data

In this chapter we present the results, obtained by treatment of real data. The data file was extracted from the Umea Data Base with kind help of Gun Stenflo (Umea University, Sweden). The file included records of survival time for children born in one parish by mothers not older than 26 years in 1818-1895. That file was separated in two subfiles in accordance with parent's occupation. First subfile included records for children of farmers, workers, rural proletarians and cases with no occupational reference. The second subfile included the rest and in fact it was records with unknown occupation. We had 196 records in the first subfile and 579 in the second one. It was found that survivalship of children in these two files is different. For children of farmers, workers, rural proletarians and no occupational reference the mean value of survival time was 1180 days. 80% of this group survived more then 200 days, 50% survived more than 540 days and 20% survived more then 2000 days. For children of parents with unknown occupation the mean value of survival time was 427 days. 80% of this group survived more then 90 days, 50% survived more then 200 days and 20% survived more then 500 days. Histograms of survival time, based on these two files are presented on Charts 2 and 3.

Chart 2.



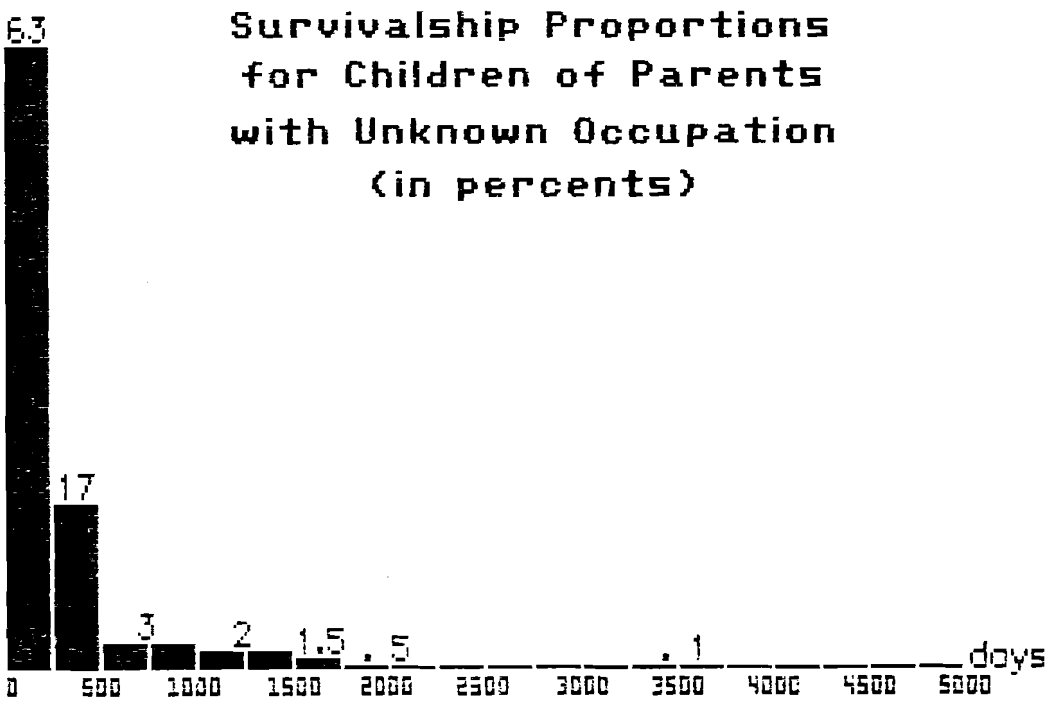
It is worth mentioning that the percent of dead children in the first subgroup is three times less than in the second one. In numbers per cents are 18.5% for the first subgroup and 53.0% for the second one. Such a situation could happen for instance, if the subgroup with unknown occupation has had more cases with bad feeding of the children and only "strong chaps" survive.

To demonstrate the use of the method we put back records from the two subgroups together. Information about surviving in those two subgroups, which we obtained on the preliminary investigation, was used as apriory information. We set a hypothesis that the general sample consists of two homogeneous sets. The value of hazard rate for the first set we assumed to be equal to the estimate of hazard rate, calculated on surviving times in records for children of physical workers. For the second set we put hazard rate equals to the estimate of it, calculated on surviving times in records with unknown occupation of parents. The numbers were 0.000847 and 0.00234 for the first and the second sets, respectively. For estimation of hazard rates we used maximum likelihood estimate in the form

$$\hat{\mu} = \frac{1}{1/L \sum_{i=1}^{L} t_i} .$$

Then we applied our method to estimate the proportion between two mentioned sets in the general sample. By calculations on IBM PC we estimated the proportion between first and second sets as 5/13. In our data file the relation between records with occupation more than four to records with occupation zero was 5/14. So the estimation is rather close to the original value. It means that the method can be successfully used for estimation of hidden heterogeneity.

Chart 3.



Appendix

Structural Minimization of Mean Risk in Small Sample Cases

Equation (1) can be solved using special probabilistic techniques for its solution. The approach is based on the methods of structural minimization of mean risk. Comprehensive analysis of this problem was developed by Vapnik [10]. More detailed consideration of integral equations' solution problems related to the mean risk minimization was done by Michalski [11].

The idea of mean risk minimization method is as follows. Let X be a random variable with distribution function $F(x)$. Let $\{Q\colon Q(x) \geq 0\}$ denote the class of all nonnegative functions such that for each function $Q(x)$ the functional

$$G = \int Q(x) dF(x) \tag{A1}$$

exists. The functional G is called the mean risk functional. To minimize the mean risk means to find the function Q^* from the family of functions $\{Q\}$ such that mean risk takes the minimal value on Q^*. Note that if the distribution function $F(x)$ is known, the approach to minimization of a mean risk is straightforward.

In many practical problems, however, the distribution function of X is unknown, but the sample of independent realizations of X is often available. If the sample is large enough the problem is equivalent to the mean risk minimization with a known distribution function. If the sample is small then one should use another approach to minimize the mean risk. Such approach is called the structural minimization of mean risk [10].

It turns out that the property of sample to be "small" or "large" depends on its size L and on the properties of functional family $\{Q\}$. This crucial property of functional family is called the "complexity" of this family.

The main idea of structural minimization of mean risk method is to substitute the unknown mean risk functional (A1) by the empirical risk functional G_L which is completely defined by the sample of random variable X:

$$G = \frac{1}{L} \sum_{j=1}^{L} Q(x_j) \quad , \tag{A2}$$

to structurize the functional family $\{Q\}$, selecting several classes of $\{Q_1\}, \{Q_2\}, \ldots, \{Q_n\}$ and making minimization within each class.

The .first step in this procedure seems to be natural since the sample of X is the only information about unknown distribution. The next step deserves special explanation.

Minimizing the empirical risk within the class $\{Q\}$ one should be sure that its minimizing function is close enough to the function that minimizes the mean risk. The guarantee of this closeness is the uniform convergence of the empirical risk functional to the mean risk functional when the size of the sample L tends to infinity.

The uniform convergence of empirical risk means that for any fixed ε the probability $P_{\varepsilon L}$

$$P_{\varepsilon L} = P \left\{ \sup_{Q \in \{Q\}} \frac{G - G_L}{G} \geq \varepsilon \right\} \tag{A3}$$

goes to zero when the size L of the sample tends to infinity. It turns out that probability $P_{\varepsilon L}$ depends on the property of a functional class $\{Q\}$. This property is represented by the notion of "complexity" of a class $\{Q\}$. The precise mathematical definition of the measure of complexity K of a functional class one can find in [10]. Later we will give the measure of complexity for some particular functional classes.

If the uniform convergence exists then probability $P_{\varepsilon L}$ can be estimated as follows

$$P_{\varepsilon,L} < (\frac{L}{K} e)^K e^{-L\varepsilon^2} \tag{A4}$$

where K is the complexity index. One can see from this inequality that the less K is, the better is approximation of mean risk by the empirical one. it means that in the "simple" classes of functions one can find more precise estimation of the mean risk.

To implement this result to the problem of mean risk minimization using the sample of values of random variable X, let us consider the system of functional classes $\{Q_1\} \subset \{Q_2\} \subset \cdots \{Q_n\}$ with the increasing indices of complexity. Let us show how in this case the inequality (A3) can be used. Taking into account (A4) we have

$$P_{\varepsilon L} = P\left\{ \sup_{Q \in \{Q_i\}} \left(\frac{G - G_L}{G}\right) \geq \varepsilon \right\} < \left(\frac{L}{K_i} e\right)^{K_i} e^{-L\varepsilon^2} \tag{A5}$$

where K_i is the complexity index of $\{Q_i\}$.

Denoting by q the right-hand side of inequality (A5) one can easily find the formula for ε when q, L, and K_i are given

$$\varepsilon = \sqrt{\frac{K_i\,(ln\,\frac{L}{K_i} + 1) - \ln q}{L}} \;.$$

Using this expression one can estimate the mean risk value by the empirical risk using formula

$$\frac{G - G_L}{G} < \varepsilon = \sqrt{\frac{K_i\,(ln\,\frac{L}{K_i} + 1) - \ln q}{L}}$$

or

$$G < \frac{G_L}{1 - \sqrt{\frac{K_i\,(ln\,\frac{L}{K_i} + 1) - \ln q}{L}}} \;. \tag{A6}$$

This formula makes sense for all functions from the class $\{Q_i\}$ if the denominator in the right-hand side is positive. Note that the reached value of mean risk in the class $\{Q_i\}$ has an upper bound B_i

$$B_i = \frac{\min\limits_{Q \in \{Q_i\}} G_L}{1 - \sqrt{\frac{K_i\,(ln\,\frac{L}{K_i} + 1) - \ln q}{L}}} \;.$$

Thus for each functional class $\{Q_i\}$ and given L and q one can calculate three variables: ε_i, G_L^i, and B_i which correspond to the value of relative uniform approximation error, minimum value of empirical risk in the class $\{Q_t\}$ and the upper bound of the reached value of the mean risk at the minimum point of the empirical risk in the class $\{Q_t\}$.

In the classes with small K_t the value of ε_t is small and the empirical risk gives a good approximation for the mean risk. However the minimum value of the empirical risk G_L^t can be high and consequently the reached value of the mean risk upper bound B_t can also be high.

With the increasing of the complexity of the class $\{Q_{t+1}\}$ the approximation of mean risk by the empirical risk became worse, the value of ε_{t+1} became larger but the maximum value of the empirical risk G_L^{t+1} is decreasing since $\{Q_t\} \subset \{Q_{t+1}\}$. As a result of that the upper bound B_{t+1} is also decreasing. Starting from some level of complexity of the class $\{Q_t\}$, say K_{t^*}, the growth of the error ε_t is not compensated by the decreasing of the value of the empirical risk G_L^t and the upper bound of the reached value of the mean risk starts to grow. It means that $\{Q_{t^*}\}$ can be chosen as a proper class in which the minimization of the empirical risk will guarantee the minimal value of the upper bound for the reached mean risk with given probability $1-q$.

One example for the system of classes $\{Q_t\}$ can be given by the algebraic polinoms of different degrees:

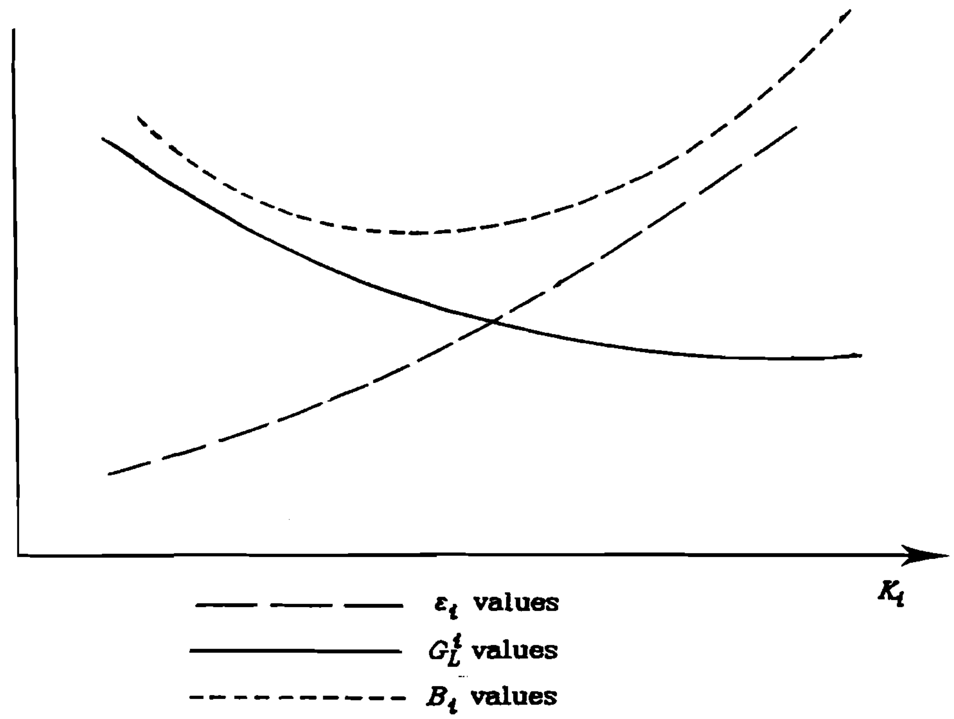
$$\{Q_t\} = \{Q(x,y) = (y - \sum_{j=0}^{t-1} \alpha_j x^j)^2\}$$

where α_j are the arbitrary parameters. If the sample of the couple (x,y) is given then one can calculate the value of the empirical risk and the value of B_t which we will identify with the estimation of the mean risk

By solution of the mean risk minimization problem using the finite sample of couple (x,y) we will understand the function Q^* which givees the minimum of the empirical risk in the class $\{Q_{t^*}\}$. This value depends on sample size L, sample values, and the validation value of the uniform approximation of the mean risk by the empirical one $1-q$. In practical calculations this value is often taken as .95. The typical situation is represented in Figure 1.

The important property of the structural mean risk minimization is that it does not require that the minimizing function belongs to the functional family $\{Q\}$. The method allows to make the best guaranteed approximation based on the finite size of the experimental sample and set of classes $\{Q_1\},\{Q_2\},\dots$. Moreover, it

Figure 1.



turns out that in the case of finite samples sometimes one should exclude the minimum point from the functional class [10].

Let us explain the notion o complexity index K for functional family $\{Q\}$. Assume that one has a sample $T = \{X_1,\dots,X_L\}$ of random variable X. For any given number $C > 0$ and function $Q(x)$ one can divide the sample T into two subsamples T' and T'' using the rule: number X_j belongs to subsample T' if $Q(X_j) > C$ and to subsample T'' if $Q(x_j) \le C$. Changing the number C and taking all possible functions $Q(x)$ from $\{Q\}$ one gets different subsamples. The maximal number of different divisions for all possible samples having the size L is called the complexity function of the class $\{Q\}$ on the samples having the size L. This function depends on the sample size and the functional family. We will use the notation $m_Q(L)$ for this function. It is clear that $m_Q(L) \le 2^L$. It turns out that the complexity function either equals 2^L or starting from some number K satisfies the inequality

$$m_Q(L) < (\frac{L}{K}e)^K$$

where K is the critical sample size. The variable K depends only on the properties of the functional family $\{Q\}$ and is called its complexity index.

The value of K in some cases can be easily calculated. If, for instance, $Q(x,y) = (y - \sum_{j=0}^{N-1} \alpha_j X^j)^2$ then $K = N$. Another example corresponds to the case when the function $Q(x)$ has not more than N extremums and x is scalar. In this case $K = N + 1$ [10].

Note that everywhere in this chapter the explanation of mean risk optimization was conducted in terms of functions of one or two random variables X and Y. One can easily see that the approach is appropriate for an arbitrary number of random variables.

REFERENCES

[1] Heckman, J.J. and B. Singer (1985) The Identification Problem in Econometric Models for Duration Data. In *Advances in Econometrics*, edited by Werner Hildenbrand. Cambridge University Press.

[2] Vaupel, J.W. and A.I. Yashin (1985) The Deviant Dynamics of Death in Heterogeneous Populations. Pages 179-211 in *Sociological Methodology 1985*, edited by Nancy B. Tuma. San Francisco: Jossey-Bass.

[3] Vaupel, J.W. and A.I. Yashin (1986) Heterogeneity Ruses: Some Surprising Effects of Selection on Population Dynamics. *The American Statistician* 39(3):176-185.

[4] Shepard, D. and R. Zeckhauser (1977) *Interventions in Mixed Populations: Concepts and Applications*. Discussion Paper Series. Harvard University, JKF School of Government.

[5] Keyfitz, N. and G. Littman (1979) Mortality in a Heterogeneous Population. *Population Studies* 33:333-342.

[6] Manton, K.G., E. Stallard, and J.W. Vaupel (1986) Alternative Models for the Heterogeneity of Mortality Risks Among the Aged. *Journal of the American Statistical Association* 81(385):635-644.

[7] Riesz, F. and B. Nagy (1955) *Functional Analysis*. Ungaz, New York.

[8] Tichonov, A.N. and Arsenin, V.A. (1974) Method for Ill-Posed Problems Solution. Manuscript.

[9] Cox, D.R. (1972) Regression Models and Life Tables. *Journal of the Royal Statistical Society*, Series B 34:187-202.

[10] Vapnik, V.N. (1982) *Dependencies Restoration on Base of Small Samples*. Springer-Verlag.

[11] Michalski, A.I.(1984) Algorithms for Dependencies Reconstruction. Moscow, Nauka.