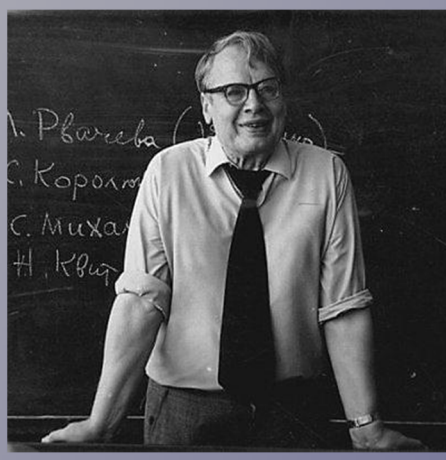
ELECTRONIC JOURNAL
OF INTERNATIONAL GROUP
ON RELIABILITY

Gnedenko Forum Publications



JOURNAL IS REGISTERED
IN THE LIBRARY
OF THE U.S. CONGRESS

•

ISSN 1932-2321

•

VOL.8 NO.2 (29)
JUNE, 2013

# RELIABILITY:
## THEORY&APPLICATIONS


San Diego

**ISSN** 1932-2321


http://www.gnedenko-forum.org/Journal/index.htm

# RELIABILITY:
## THEORY & APPLICATIONS

Vol.8 No.2 (29),
June, 2013

San Diego
2013

**Send your paper**

e-Journal ***Reliability: Theory & Applications*** publishes papers, reviews, memoirs, and bibliographical materials on Reliability, Quality Control, Safety, Survivability and Maintenance.

Theoretical papers have to contain new problems, finger practical applications and should not be overloaded with clumsy formal solutions.

Priority is given to descriptions of case studies.

General requirements for presented papers

1. Papers have to be presented in English in MSWord format. (Times New Roman, 12 pt , 1.5 intervals).
2. The total volume of the paper (with illustrations) can be up to 15 pages.
3. A presented paper has to be spell-checked.
4. For those whose language is not English, we kindly recommend to use professional linguistic proofs before sending a paper to the journal.

* * *

The Editor has the right to change the paper title and make editorial corrections.

The authors keep all rights and after the publication can use their materials (re-publish it or present at conferences).

Publication in this e-Journal is equal to publication in other International scientific journals.

Papers directed by Members of the Editorial Boards are accepted without referring.

The Editor has the right to change the paper title and make editorial corrections.

The authors keep all rights and after the publication can use their materials (re-publish it or present at conferences).

Send your papers to

the Editor-in-Chief ,
Igor Ushakov
igusha22@gmail.com

or

the Deputy Editor,
Alexander Bochkov
a.bochkov@gmail.com

# Table of Contents



Generalized Birnbaum-Saunders (GBS) distributions are proposed by Díaz-García et al. ([15], [16]) based on the family of elliptically contoured univariate distributions. This model is well-known as the highly flexible lifetime model by the difference in the degrees of kurtosis and asymmetry and processes uni-modality and bimodality. In this paper, a modifier Chi-squared goodness-of-fit test based on Nikulin-Rao-Robson statistics $Y_n^2$ is developed for the family of GBS distributions for the right censored data with unknown parameters by using the maximum likelihood estimation (MLE). Some applications of this model in survival analysis discuss also in the section of real study.



In this paper new algorithm of interval images recognition is suggested. This algorithm gives accuracy solution of considered problem but demands not linear but square complexity by a number of objects. Main motive of such construction is to analyze practically interesting case when there is preliminary silence before predicted events.



Parameter estimates, constructed by the minimum distance method, are briefly called the MD-estimates. The minimum distance method has been proposed by Wolfowitz (1957). An extensive bibliography was compiled and published by Parr (1981). In this paper the effectiveness of the shift parameter estimation based on the use of Cramer - von Mises weighted distance is discussed. The robustness of this kind of MD-estimates under various supermodels describing deviations from the Gaussian model is considered. Numerical results are given for the case of contaminated normal distributions.



In this paper multilevel approach to the issue of road safety level on the road network of European regions, classified as NUTS 2 in statistical databases of the European Union, has been presented. Following the pattern of many publications on road safety it has been assumed that the risk calculated as the number of death casualties in road accidents per 100,000 inhabitants of a given region has Poisson distribution. Therefore, generalized Poisson model has been assumed in the modelling process. Multilevel stochastic analysis was performed for the studied factor. Then a model was created that took into account the impact of different characteristics available on different level of aggregation, which may be helpful in the actions aimed at improvement of road safety in respective regions.

The new method of modeling of continuous random variables on empirical distributions offered. It shown, that discrepancy of accuracy of methods to shown requirements is shown at small number of realizations of random variables, reduced to not casual divergence of estimations of averages and average quadratic values empirical given and modeled samples.



Objective estimation of parameters of individual reliability is an indispensable condition of an opportunity of decrease in operational expenses for maintenance service and repair of the equipment and devices of electro power systems. The method of decrease in risk of erroneous classification of multivariate statistical data offered. The method based on imitating modeling and the theory of check of statistical hypotheses.

# CHI-SQUARED GOODNESS OF FIT TEST FOR GENERALIZED BIRNBAUM-SAUNDERS MODELS FOR RIGHT CENSORED DATA AND ITS RELIABILITY APPLICATIONS

**M. S. Nikulin[1] and X. Q. Tran[1]**

•

[1]Bordeaux University, IMB, UMR 5251, F-33400 Talence, France

E-mail: mikhail.nikouline@u-bordeaux2.fr,
xuanquang.tran@math.u-bordeaux1.fr

## ABSTRACT

Generalized Birnbaum-Saunders (GBS) distributions are proposed by Díaz-García *et al.* ([15], [16]) based on the family of elliptically contoured univariate distributions. This model is well-known as the highly flexible lifetime model by the difference in the degrees of kurtosis and asymmetry and processes uni-modality and bimodality. In this paper, a modifier Chi-squared goodness-of-fit test based on Nikulin-Rao-Robson statistics $\boldsymbol{Y}_n^2$ is developed for the family of GBS distributions for the right censored data with unknown parameters by using the maximum likelihood estimation (MLE). Some applications of this model in survival analysis discuss also in the section of real study.

**Keywords and phrases**: Birnbaum-Saunders distribution, Breast cancer, Carcinoma data, Chi-squared test, Censoring sample, GBS distributions, Goodness-of fit test, NRR test, Survival analysis.

## 1 Introduction

In 1969, Birnbaum and Saunders [9] have been proposed a model with two shape and scale parameters that is well known as Birnbaum-Saunders (BS) distribution. After their work, there was a lot of research work on this model and its applications in reliability and survival analysis. It must be mentioned as the work of Desmond [14] who strengthened the physical justification for the use of this distribution by relaxing some assumptions early bade Birnbaum-Saunders. Based on this distribution, Leiva *et al.* [23] has worked to model survival times of patients with multiple myeloma by using prognostic variables with censored data. A chi-squared test for this model is analyzed by Tahir [37] in 2012. In addition, in the recent research of Nikulin *et al.* ([30], [2], [29]) considered these applications of this model in the accelerated lifetimes (AFT) models and redundant systems. Nowadays, the BS distribution has known as cumulative damage distributions and it is a very useful in fatigue, reliability and survival analysis. However, its field of application has been extending beyond the original context of material fatigue and reliability analysis.

Therefore, studies to expand of the BS distribution have been looking for researchers in recent years, such as: Owen ([32], [33], [31]) proposed a three parameter Birnbaum-Saunders distribution, in 2000. Later, Volodin and Dzhungurova [38] developed a general family of fatigue life distributions denominated the crack distribution, which includes the Birnbaum-Saunders distribution as a particular case. In particular, we should be mention a generalized family of life distribution which is suggested by Díaz-García *et al.* [15] in their technical report in 2002, is called as the Generalized Birnbaum-Saunders (GBS) distributions. In his works, Díaz-García was obtained a distribution of the Birnbaum-Saunders type with different degrees of kurtosis, uni-modality, bimodality and absence of moments by basing on the family of elliptically contoured univariate distributions (which known as standard symmetrical distributions in $R$. A complete review about the

GBS distributions can be found in Sanhueza, Leiva, and Balakrishnan [36]. The purpose of this paper, we analyze a Nikulin-Rao-Robson $\boldsymbol{Y}_n^2$ goodness-of-fit tests for these distributions in the case of right censoring observations. We also demonstrate the applications of this model by applying it to reliability and survival data.

## 2 Generalized Birnbaum-Saunders distributions

As is already known, a random variable $T$ following the BS distribution allows the stochastic representation

$$T = \beta\left[\alpha\frac{Z}{4} + \sqrt{\alpha^2\frac{Z}{2} + 1}\right]^2 \approx BS(\alpha,\beta), \qquad \alpha > 0, \beta > 0, \tag{1}$$

where, $Z \approx N(0,1)$. Then the random variable $Z$ may be stochastically represented in the form

$$Z = \frac{1}{\alpha}\left[\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right] \approx N(0,1). \tag{2}$$

In 2002, Díaz-García *et al.* [15] were developed the BS distribution becomes GBS distributions which are related to standard symmetrical distributions in $R$, also known as elliptically contoured or simple Elliptic distribution ([1], [8], [21], [19], [12]).

If a random variable $Z$ follows an Elliptic distributions which correspond to all the symmetric distribution in $R$, denoted by $Z \approx EC(\mu,\ \sigma^2;\ g)$, the probability density function $f_Z(z)$ and cumulative distribution function $F_Z(z)$ of $Z$ given by,

$$f_Z(z;\ \mu,\sigma^2) = cg\left(\frac{(z-\mu)^2}{\sigma^2}\right), \qquad F_Z(z;\ \mu,\sigma^2) = \int_{-\infty}^{z} f_Z(u;\ \mu,\sigma^2)du\,, z \in\ R, |\mu| < \infty, \sigma\ > 0\,.$$

respectively, where, $g(\cdot)$ is the kernel of the probability density function of $Z$, $c$ is the positive normalization constant, such that $\frac{1}{c} = \int_{-\infty}^{+\infty} g(u^2)du$. The families Elliptic distributions include three sub-models: Kotz Type (KT), Pearson type VII (PVII) and type-III generalized logistic (LIII) distributions, for more details on these distributions is given by Anderson [1] Balakrishnan [8], Fang [19], Gupta [21], Cambanis [12] and others. The Normal, Cauchy, Laplace, Logistic, Power Exponential and $t(\nu)$-Student distributions are particular cases of these symmetric sub-classes in $R$. In table 1 below, we recall some results for kernel function $g(\cdot)$, the constant $c$ corresponding with standard symmetric distribution $EC(0,\ 1;\ g)$ in $R$.

| Distribution | Notation | $c$ | $g(z^2), z \in R$ |
|---|---|---|---|
| Normal | $N(0,1)$ | $\frac{1}{\sqrt{2\pi}}$ | $exp\left(-\frac{z^2}{2}\right)$ |
| $t(\nu)$-Student | $t(\nu)$ | $\frac{\Gamma\left\{\frac{\nu+1}{2}\right\}}{\Gamma(\nu+2)\sqrt{\nu\pi}}$ | $\left\{1+\frac{z^2}{\nu}\right\}^{-\frac{\nu+1}{2}}$ |
| Laplace | $L(0,1)$ | 0.5 | $e^{-\lvert z\rvert}$ |
| Logistic | $Log(0,1)$ | 1 | $\frac{e^{-z}}{(1+e^{-z})^2}$ |
| Cauchy | $C(0,1)$ | $\frac{1}{\pi}$ | $\frac{1}{1+z^2}$ |
| Power exponential | $PE(\nu)$ | $\frac{\nu}{(2\nu)^{\frac{1}{2\nu}}\Gamma\left(\frac{1}{2\nu}\right)}$ | $exp\left(-\frac{1}{2\nu}\lvert z\rvert^{2\nu}\right)$ |
| LIII | $LIII(q)$ | $\frac{\Gamma(2s)}{\Gamma^2(q)}$ | $\frac{e^{qz}}{(1+e^z)^{2q}}, q > 0$ |

| Distribution | Notation | $c$ | $g(z^2), z \in R$ |
|---|---|---|---|
| Pearson VII | $PVII(q,r)$ | $\frac{\Gamma(q)}{\Gamma(\frac{q-1}{2})\sqrt{r\pi}}$ | $\left[1+\frac{z^2}{r}\right]^{-q}, q>\frac{1}{2}, \nu>0$ |
| Kotz type | $KT(q,r,s)$ | $\frac{sr^{(2q-1)/2s}}{\Gamma(\frac{2q-1}{2s})}$ | $z^{2(q-1)}e^{-rz^{2s}}, q>\frac{1}{2}, r>0, s>0$ |

**Table 1: Kernel $g(\cdot)$ and normalization constants $c$ for some indicated distributions.**

Following Díaz-García *et al*. the random variable $T$ in (1) allows the GBS distributions, denoted by $T \approx GBS(\alpha, \beta; g)$,

$$T = \beta\left[\alpha\frac{Z}{4} + \sqrt{\alpha^2\frac{Z}{2}+1}\right]^2 \approx GBS(\alpha,\beta,g), \alpha>0, \beta>0,$$

iff the random variable $Z$ which is given by the expression

$$Z = \frac{1}{\alpha}\left[\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right] \approx EC(0,1; g).$$

So, the probability density function of $T$ can be written as

$$f_T(t,\alpha,\beta) = \frac{c}{2\alpha\beta}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}} + \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right\} g\left(\frac{1}{\alpha^2}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right]^2\right), t>0, \alpha>0, \beta>0, \quad (3)$$

the cumulative distribution function of $T \approx$ GBS(α, β; g) is expressed by

$$F_T(t,\alpha,\beta) = F_Z\left\{\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right]\right\}, t>0, \alpha>0, \beta>0, \quad (4)$$

the GBS hazard rate, survival and cumulative hazard functions are

$$\lambda_T(t,\alpha,\beta) = \frac{f_Z(a_t(\alpha,\beta))A_t(\alpha,\beta)}{1-F_Z(a_t(\alpha,\beta))},$$
$$S_T(t,\alpha,\beta) = 1 - F_Z(a_t(\alpha,\beta)), \text{and } \Lambda_T(t,\alpha,\beta) = -ln\{S_T(t,\alpha,\beta)\}, \quad (5)$$

respectively, where

$$a_t(\alpha,\beta) = \frac{1}{\alpha}\left[\left(\frac{t}{\beta}\right)^{\frac{1}{2}} - \left(\frac{\beta}{t}\right)^{\frac{1}{2}}\right]; \; A_t(\alpha,\beta) = \frac{1}{2\alpha\beta}\left[\left(\frac{\beta}{t}\right)^{\frac{1}{2}} + \left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right].$$

It is clear that the properties of GBS distributions depends on the kernel function $g(\cdot)$ and the unknown parameter $\boldsymbol{\theta} = (\alpha,\beta)^T$. The statistical theory and methodology of the GBS distributions, also some results for this flexible family of distributions mainly related to transformations, the hazard failure and censored data type II which can be found in the works of Sanhueza, Leiva et al.[36].

Table 2 below shown some probability density function of $T \approx GBS(\alpha, \beta; g)$, corresponding the specific symmetric distribution $EC(0,1; g)$ in Table 1.

The Figure 1, 2, 3 and 4 below illustrates some curve of the probability densities and hazard rate functions of $T \approx GBS(\alpha, \beta; g)$, allows with the kernel indicative.

| Kernel $g(\cdot)$ | Distribution | Probability density function of $T \approx GBS(\alpha, \beta; g)$, $f(t,\alpha,\beta;g), (t>0, \alpha>0, \beta>0)$ |
|---|---|---|
| *N(0, 1)* | GBS-Normal (BS) | $\frac{1}{2\alpha\beta\sqrt{2\pi}}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}} + \left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\exp\left\{-\frac{1}{\alpha^2}\left[\frac{t}{\beta}+\frac{\beta}{t}-2\right]\right\}$ |

| Kernel | Distribution | Probability density function of , |
|---|---|---|
| | GBS-Student | $\frac{\Gamma\left\{\frac{\nu+1}{2}\right\}}{2\alpha\beta\,\Gamma(\nu+2)\sqrt{\nu\pi}}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\left(1+\frac{1}{\nu\alpha^2}\left[\frac{t}{\beta}+\frac{\beta}{t}-2\right]\right)^{\frac{-(\nu+1)}{2}}$ |
| *L(0, 1)* | GBS-Laplace | $\frac{1}{4\alpha\beta}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\exp\left\{-\frac{1}{\alpha}\left\lvert\sqrt{\frac{t}{\beta}}-\sqrt{\frac{\beta}{t}}\right\rvert\right\}$ |
| *Log(0, 1)* | GBS-Logistic | $\frac{1}{2\alpha\beta}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\frac{\exp\left\{\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}}-\sqrt{\frac{\beta}{t}}\right]\right\}}{\left\{1+\exp\left\{\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}}-\sqrt{\frac{\beta}{t}}\right]\right\}\right\}^2}$ |
| *C(0, 1)* | GBS-Cauchy | $\frac{1}{2\pi\alpha\beta}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\left(1+\frac{1}{\alpha^2}\left[\frac{t}{\beta}+\frac{\beta}{t}-2\right]\right)^{-1}$ |
| *KT(q, r, s)* | GBS-KT | $\frac{sr^{\frac{2q-1}{2s}}\alpha^{2q-3}}{2\beta\Gamma\left(\frac{2q-1}{2s}\right)}\frac{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}}{\left(\frac{t}{\beta}+\frac{\beta}{t}-2\right)^{1-q}}\exp\left\{-\frac{r}{\alpha^{2s}}\left[\frac{t}{\beta}+\frac{\beta}{t}-2\right]^s\right\}$ |
| *PVII(q, r)* | GBS-PVII | $\frac{\Gamma(q)}{2\alpha\beta\sqrt{r\pi}\,\Gamma(q-1/2)}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\left\{1+\frac{1}{\nu\alpha^2}\left[\frac{t}{\beta}+\frac{\beta}{t}-2\right]\right\}^{-\frac{\nu+1}{2}}$ |
| *PE(r, s)* | GBS-PE | $\frac{sr^{\frac{1}{2s}}}{2\alpha\beta\Gamma(\frac{1}{2s})}\left\{\left(\frac{\beta}{t}\right)^{\frac{1}{2}}+\left(\frac{\beta}{t}\right)^{\frac{3}{2}}\right\}\exp\left\{-\frac{r}{\alpha^{2s}}\left[\sqrt{\frac{t}{\beta}}-\sqrt{\frac{\beta}{t}}\right]^{2s}\right\}$ |

**Table 2: The p.d.f of for some indicated distributions.**

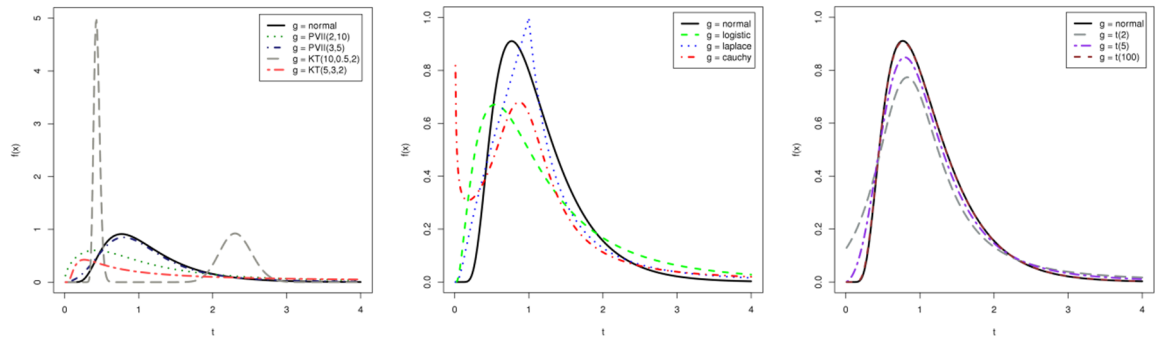



**Figure 1: Plots of densities for given kernel .**

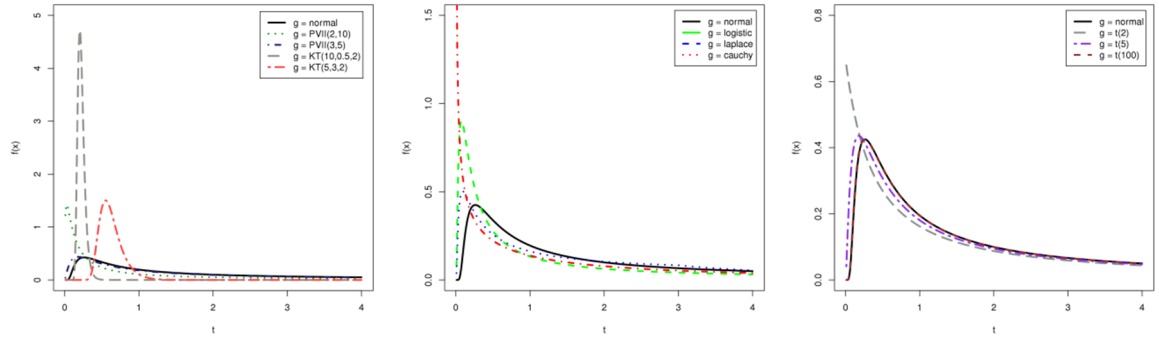



**Figure 2: Plots of densities for given kernel .**

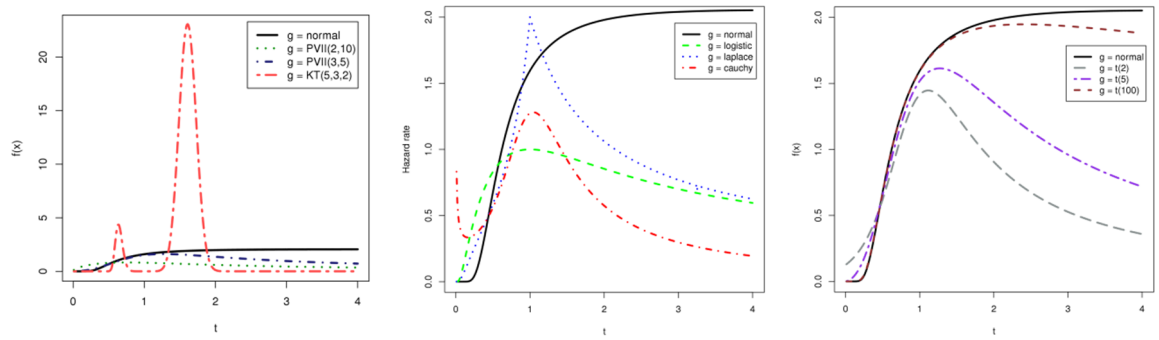



**Figure 3: Plots of failures rates for given kernel .**

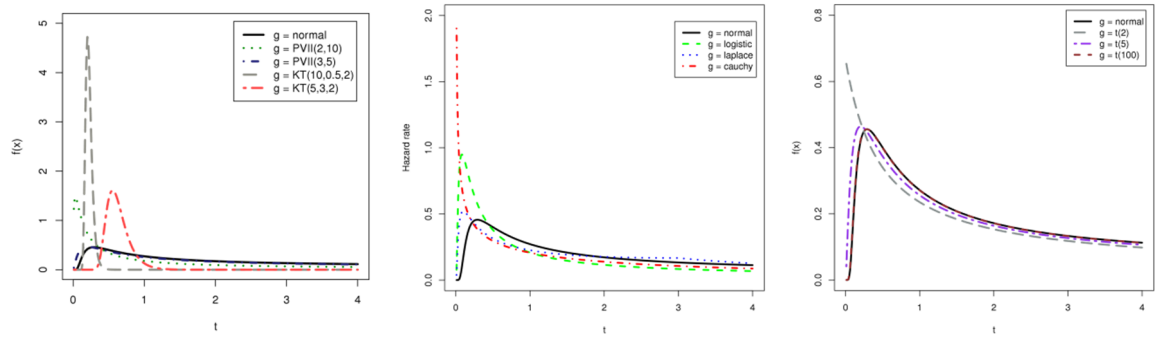



**Figure 4: Plots of failures rates for given kernel .**

## 3 Chi-squared type tests for right censored data

Following Bagdonavičius and Nikulin ([3], [4]), we describe a chi-squared test for testing composite parametric hypothesis when data are right censored.

Suppose that are failures time non-negative and independent and the probability density function of the random variable belong to a parametric family . The censoring variables are also non-negative and assumed to be random sample. Let us and are independent. We observed

(6)

where,

$$X_i = T_i \wedge C_i, \qquad \delta_i = 1_{\{T_i \le C_i\}}, i = 1, 2, \cdots, n.$$

Defined that

$$S(t,\boldsymbol{\theta}) = \boldsymbol{P}_{\boldsymbol{\theta}}(T > t); \;\; \lambda(t,\boldsymbol{\theta}) = \frac{f(t,\boldsymbol{\theta})}{S(t,\boldsymbol{\theta})}, \qquad \Lambda(t,\boldsymbol{\theta}) = -\ln\{S(t,\boldsymbol{\theta})\}, \boldsymbol{\theta} \in \Theta \subseteq R^m,$$

be the survival, hazard rate and cumulative hazard functions, respectively. Denote by $G_i$ and $g_i$ are the survival and the density function of the censoring time $C_i$, respectively. Supposing that the right censoring is non-informative which means that the function $G_i$ does not depend on $\boldsymbol{\theta}$. So in this case, we obtain the following expressions for the likelihood function $L(\boldsymbol{\theta})$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f^{\delta_i}(X_i,\boldsymbol{\theta}) S^{1-\delta_i}(X_i,\boldsymbol{\theta}) g^{1-\delta_i}(C_i) G^{\delta_i}(C_i).$$

So the members with $G_i$ and $g_i$ do not contain $\boldsymbol{\theta}$, so they can be rejected. The likelihood function is obtained

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f^{\delta_i}(X_i,\boldsymbol{\theta}) S^{1-\delta_i}(X_i,\boldsymbol{\theta}) = \prod_{i=1}^{n} \lambda^{\delta_i}(X_i,\boldsymbol{\theta}) S(X_i,\boldsymbol{\theta}). \tag{7}$$

The estimator $\widehat{\boldsymbol{\theta}}_n$ maximizing the likelihood function $L(\boldsymbol{\theta})$. The log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{\delta_i \ln \lambda(X_i,\boldsymbol{\theta}) + \ln S(X_i,\boldsymbol{\theta})\} = \sum_{i=1}^{n} \{\delta_i \ln \lambda(X_i,\boldsymbol{\theta}) - \Lambda(X_i,\boldsymbol{\theta})\}. \tag{8}$$

The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ satisfies the system equations

$$\dot{\ell}(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}_m,$$

where $\dot{\ell}(\boldsymbol{\theta})$ are the score vectors

$$\dot{\ell}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} \ell(\boldsymbol{\theta}) = \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2}, \cdots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_m}\right)^T.$$

The Fisher information matrix is defined as

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E_\theta \ddot{\ell}(\boldsymbol{\theta}),$$

where

$$\ddot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln \lambda(X_i,\boldsymbol{\theta}) - \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \Lambda(X_i,\boldsymbol{\theta}).$$

Supposing that $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$, under some regularity conditions, we have

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0; \;\; \sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) = i^{-1}(\boldsymbol{\theta}_0) \frac{1}{\sqrt{n}} \dot{\ell}(\boldsymbol{\theta}_0) + O_P(1), \frac{-1}{\sqrt{n}} \ddot{\ell}(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{P} i(\boldsymbol{\theta}_0),$$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N_m(0, i^{-1}(\boldsymbol{\theta}_0)); \;\; \frac{1}{\sqrt{n}} \dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} N_m(0, i(\boldsymbol{\theta}_0)),$$

where, $\widehat{\boldsymbol{\theta}}_n$ are the maximum likelihood estimation of $\boldsymbol{\theta}$ and the matrix

$$i(\boldsymbol{\theta}_0) = \lim_{n \to \infty} \frac{\boldsymbol{I}(\boldsymbol{\theta}_0)}{n}.$$

For any $t \ge 0$, set

$$N_i(t) = 1_{\{t \ge X_i, \delta_i = 1\}} = \begin{cases} 1, & \text{if } t \ge \mathrm{X}_i \text{ and } \delta_i = 1, \\ 0, & \text{if } 0 \le t \le \mathrm{X}_i. \end{cases}; \;\; Y_i(t) = 1_{\{t \le X_i\}} = \begin{cases} 1, & \text{if } t \le \mathrm{X}_i, \\ 0, & \text{if } t > \mathrm{X}_i. \end{cases}$$

$$N(t) = \sum_{i=1}^{n} N_i(t), \qquad Y(t) = \sum_{i=1}^{n} Y_i(t).$$

The process $N(t)$ shows the number of observed failures in the interval $[0, t]$ and the process $Y(t)$ shows the number of objects which are "at risk" just prior to time $t$. The sample (6) is equivalent to the sample

$$(N_1(t), Y_1(t), t \geq 0), (N_2(t), Y_2(t), t \geq 0), \cdots, (N_n(t), Y_n(t), t \geq 0). \quad (9)$$

For the sample (9), the parametric log-likelihood function can be written by expression follows

$$\ell(\boldsymbol{\theta}) = \int_0^{+\infty} \{\ln \lambda(u, \boldsymbol{\theta})\, dN(u) - Y(u)\lambda(u, \boldsymbol{\theta})\} du.$$

The score function is

$$\dot{\ell}(\boldsymbol{\theta}) = \int_0^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta}) \{dN(u) - Y(u)\lambda(u, \boldsymbol{\theta}) du\} = \int_0^{\infty} \frac{\partial}{\partial \theta} \ln \lambda(u, \boldsymbol{\theta})\, dM(u, \boldsymbol{\theta}),$$

and

$$\ddot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \int_0^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln \lambda(u, \boldsymbol{\theta}) dM_i(u, \boldsymbol{\theta}) - \sum_{i=1}^{n} \int_0^{\infty} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta})\right)^T \lambda(u, \boldsymbol{\theta}) Y_i(u) du,$$

where, $M_i(t, \boldsymbol{\theta}) = N_i(t) - \int_0^t Y_i(u)\lambda(u, \boldsymbol{\theta}) du$, $(\boldsymbol{\theta} \in \Theta)$ is the zero mean martingale with respect to the filtration generated by the data.

Suppose that the processes $N_i$ and $Y_i$ are observed for finite time $\tau > 0$, which means that at time $\tau$, observation on all surviving objects are censored, and so instead of using censoring time $C_i$. In this case, the matrix Fisher information can be written as

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}} \ddot{\ell}(\boldsymbol{\theta}) = E_{\theta} \sum_{i=1}^{n} \int_0^{\infty} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln \lambda(u, \boldsymbol{\theta})\right)^T \lambda(u, \boldsymbol{\theta}) Y_i(u) du.$$

Let be consider next the hypothesis

$$H_0 : F(x) \in \mathcal{F}_0 = \{F_0(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq R^m\},$$

here, $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_m)^T$ are an unknown $m$-dimensional parameters and $F_0$ is a known distribution function.

Subdividing the interval $[0, \tau]$ into $k > m$ smaller intervals $I_j = (a_{j-1}, a_j]$, with $a_0 = 0$, $a_k = \tau$, and denote by

$$U_j = N(a_j) - N(a_{j-1}),$$

the number of observed failures in the $j^{th}$ interval $I_j$, ($j = 1, 2, \cdots, k$). Let

$$e_j = \int_{a_{j-1}}^{a_j} \lambda(u, \widehat{\boldsymbol{\theta}}_n) Y(u) du.$$

A chi-squared test which was proposed by Bagdonavičius and Nikulin [21], based on the vector

$$\mathcal{Z} = (Z_1, Z_1, \cdots, Z_1)^T, \text{ with } Z_j = \frac{1}{\sqrt{n}}(U_j - e_j), j = 1, 2, \cdots, k. \quad (10)$$

Under the conditions

1) There exists a neighborhood $\Theta_0$ of $\boldsymbol{\theta}_0$ such that for all $n$ and $\boldsymbol{\theta} \in \Theta_0$, and almost all $t \in [0, \tau]$, the partial derivatives of $\lambda(t, \boldsymbol{\theta})$ of the first, second and third order with respect to $\boldsymbol{\theta}$ exist and are continuous in $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta_0$. Moreover, they are bound in $[0, \tau] \times \Theta_0$ and the log-likelihood function may be differentiated three times with respect to $\boldsymbol{\theta} \in \Theta_0$, by interchanging the order of integration and differentiation.

2) $\lambda(t, \boldsymbol{\theta})$ is bound away from zero in $[0, \tau] \times \Theta_0$.

3) A positive deterministic function $y(t)$ exists such that $\sup_{t \in [0,\tau]} \left|\frac{Y(t)}{n} - y(t)\right| \xrightarrow{P} 0$.

4) Under condition 1) - 3), the matrix $i(\boldsymbol{\theta}_0) = \lim_{n \to \infty} \frac{I(\boldsymbol{\theta}_0)}{n}$ is positive definite.

The statistic of Bagdonavičius and Nikulin given as

$$\boldsymbol{Y}_n^2(\widehat{\boldsymbol{\theta}}_n) = \mathcal{Z}^T\hat{\Sigma}^-\mathcal{Z}, \tag{11}$$

where, $\hat{\Sigma}^-$ is the general inverse matrix of the covariance matrix $\hat{\Sigma}$,

$$\begin{gathered}\hat{\Sigma} = \hat{\mathcal{A}} - \hat{\mathcal{C}}^T\hat{\boldsymbol{I}}^{-1}\hat{\mathcal{C}}, \\ \hat{\Sigma}^- = \hat{\mathcal{A}}^{-1} + \hat{\mathcal{A}}^{-1}\hat{\mathcal{C}}^T\hat{\mathcal{G}}^{-1}\hat{\mathcal{C}}\hat{\mathcal{A}}^{-1}, \qquad \hat{\mathcal{G}} = \hat{\boldsymbol{I}} - \hat{\mathcal{C}}\hat{\mathcal{A}}^{-1}\hat{\mathcal{C}}^T,\end{gathered} \tag{12}$$

$\hat{\mathcal{A}}$ is the diagonal $k \times k$ matrix with the elements $A_j = \frac{U_j}{n}$ on the diagonal, $\hat{\mathcal{A}}^{-1}$ is inverse matrix of $\hat{\mathcal{A}}$, and

$$\hat{\mathcal{C}} = [\hat{C}_{lj}]_{m\times k}\text{, with } \hat{C}_{lj} = \frac{1}{n}\sum_{i:X_i\in I_j}\delta_i\frac{\partial \ln\lambda(X_i,\theta)}{\partial\theta_l}, l = 1,2,\cdots, m,\ j = 1,\ 2,\cdots,\ k, \tag{13}$$

$$\hat{\boldsymbol{I}} = [\hat{\imath}_{ll'}]_{m\times m}\text{, with } \hat{\imath}_{ll'} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\partial \ln\lambda(X_i,\ \theta)}{\partial\theta_l}\frac{\partial \ln\lambda(X_i,\ \theta)}{\partial\theta_{l'}}, l, l' = 1,2,\cdots, \qquad m. \tag{14}$$

From the definition of $\mathcal{Z}$ in (10), the test statistic $Y_n^2(\hat{\theta}_n)$ should be written as

$$\boldsymbol{Y}_n^2(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{X}_n^2 + \boldsymbol{Q}, \tag{15}$$

where,

$$\boldsymbol{X}_n^2 = \sum_{j=1}^{k}\frac{(U_j - e_j)^2}{U_j}, \qquad \boldsymbol{Q} = \widehat{\mathcal{W}}^T\hat{\mathcal{G}}^-\widehat{\mathcal{W}}, \qquad \widehat{\mathcal{W}} = \hat{\mathcal{C}}\hat{\mathcal{A}}^{-1}\mathcal{Z}, \qquad \hat{\mathcal{G}} = \hat{\boldsymbol{I}} - \hat{\mathcal{C}}\hat{\mathcal{A}}^{-1}\hat{\mathcal{C}}^T.$$

Under the hypothesis $H_0$, the limiting distribution of the statistics $\boldsymbol{Y}_n^2(\widehat{\boldsymbol{\theta}}_n)$ is chi-squared with $r = rank(\Sigma^-)$ degrees of freedom that is,

$$\lim_{n\to\infty} P\{\boldsymbol{Y}_n^2(\widehat{\boldsymbol{\theta}}_n) > x \mid H_0\} = P\{\chi_r^2 > x\}, \text{for any } x > 0.$$

**Statistical inference for the hypothesis $\boldsymbol{H_0}$**: The null hypothesis $H_0$ is rejected with approximate significance level $\alpha$ if $\boldsymbol{Y}_n^2(\widehat{\boldsymbol{\theta}}_n) > \chi_\alpha^2(r)$ or $\boldsymbol{Y}_n^2(\widehat{\boldsymbol{\theta}}_n) < \chi_{1-\alpha}^2(r)$ depending on an alternative, where $\chi_\alpha^2(r)$ and $\chi_{1-\alpha}^2(r)$ corresponding are the upper and lower $\alpha$ percentage points of the $\chi_r^2$ distribution, respectively.

Using the method of interval selection which is proposed by Bagdonavičius, and Nikulin [20], we used $a_j$ as the random data function. Define

$$E_k = \sum_{i=1}^{n}\Lambda(X_i,\widehat{\boldsymbol{\theta}}_n),\ E_j = \frac{j}{k}E_k,\ j = 1,2,\cdots,k.$$

Denote by $X_{(1)},\ X_{(2)},\ \cdots,\ X_{(n)}$ the ordered sample from $X_1,\ X_2,\ \cdots,\ X_n$. Set

$$b_i = (n-i)\Lambda\big(X_{(i)},\widehat{\boldsymbol{\theta}}_n\big) + \sum_{l=1}^{i}\Lambda\big(X_{(l)},\widehat{\boldsymbol{\theta}}_n\big), \qquad i = 1,2,\cdots,n,$$

if $i$ is the smallest natural number verifying $b_{i-1} \le E_j \le b_i$ then $\hat{a}_j$ verifying the equality

$$(n-i+1)\Lambda\big(\hat{a}_j,\widehat{\boldsymbol{\theta}}_n\big) + \sum_{l=1}^{i-1}\Lambda\big(X_{(l)},\widehat{\boldsymbol{\theta}}_n\big) = E_j$$

So

$$\hat{a}_j = \Lambda^{-1}\left(\frac{E_j - \sum_{l=1}^{i-1}\Lambda(X_{(l)},\ \widehat{\boldsymbol{\theta}}_n)}{n-i+1}, \widehat{\boldsymbol{\theta}}_n\right);\ \hat{a}_k = \max\{X_{(n)},\tau\}, (j = 1,2,\cdots,k-1). \tag{16}$$

where $\Lambda^{-1}$ is the inverse of the function $\Lambda$. We have: $0 < \hat{a}_1 < \hat{a}_2 < \cdots < \hat{a}_k$, with this choice of intervals, then $e_j = \frac{E_k}{k}$, for all $j$.

**Application for GBS distributions**: In particular, we shall give chi-squared tests NRR for the hypothesis $H_0$ that the data $X_i$ are coming from the GBS distributions with the probability

density, cumulative distribution, hazard rate, survival and cumulative hazard functions give in formulas (3), (4) and (5), respectively.

The GBS log-likelihood functions $\ell(\boldsymbol{\theta}), (\boldsymbol{\theta} = (\alpha, \beta)^T)$ is

$$\ell(\boldsymbol{\theta}) = -\delta \ln \alpha - \delta \ln \beta + \sum_{i=1}^{n} \delta_i \ln\left\{\left(\frac{\beta}{X_i}\right)^{\frac{1}{2}} + \left(\frac{\beta}{X_i}\right)^{\frac{3}{2}}\right\} + \sum_{i=1}^{n} \delta_i \ln\{g(K_i(\alpha,\beta))\} + \sum_{i=1}^{n} \delta_i \ln\{1 - F_Z(a_i(\alpha,\beta))\}$$

Let $\widehat{\boldsymbol{\theta}}_n = \left(\hat{\alpha}, \hat{\beta}\right)^T$ be maximum likelihood estimations which are solutions of the non-linear system equations

$$\left(\dot{\ell}_\alpha(\boldsymbol{\theta}), \dot{\ell}_\beta(\boldsymbol{\theta})\right) = \mathbf{0}_2.$$

Using the formula (13) – (14), the elements $\hat{\imath}_{ll'}, (l, l' = 1, 2)$ of the Fisher information matrix $\hat{\boldsymbol{I}} = [\hat{\imath}_{ll'}]_{2\times 2}$ are

$$\hat{\imath}_{11} = \frac{1}{n\hat{\alpha}^2} \sum_{i=1}^{n} \delta_i \left[-1 + \widehat{K}_i(\hat{\alpha},\hat{\beta}) v\left(\widehat{K}_i(\hat{\alpha},\hat{\beta})\right) - \frac{\hat{A}_i(\hat{\alpha},\hat{\beta}) f_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}{1 - F_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}\right]^2,$$

$$\hat{\imath}_{22} = \frac{1}{n\hat{\beta}^2} \sum_{i=1}^{n} \delta_i \left[-1 + \frac{1}{2}\frac{1 + 3\frac{\hat{\beta}}{X_i}}{1 + \frac{\hat{\beta}}{X_i}} + \frac{1}{2}\hat{A}_i(\hat{\alpha},\hat{\beta})\hat{B}_i(\hat{\alpha},\hat{\beta}) v\left(\widehat{K}_i(\hat{\alpha},\hat{\beta})\right) - \frac{1}{2}\frac{\hat{B}_i(\hat{\alpha},\hat{\beta}) f_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}{1 - F_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}\right]^2,$$

$$\hat{\imath}_{12} = \frac{1}{n\hat{\alpha}\hat{\beta}} \sum_{i=1}^{n} \delta_i \left[-1 + \widehat{K}_i(\hat{\alpha},\hat{\beta}) v\left(\widehat{K}_i(\hat{\alpha},\hat{\beta})\right) - \frac{\hat{A}_i(\hat{\alpha},\hat{\beta}) f_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}{1 - F_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}\right] \times \left[-1 + \frac{1}{2}\frac{1 + 3\frac{\hat{\beta}}{X_i}}{1 + \frac{\hat{\beta}}{X_i}} + \frac{1}{2}\hat{A}_i(\hat{\alpha},\hat{\beta})\hat{B}_i(\hat{\alpha},\hat{\beta}) v\left(\widehat{K}_i(\hat{\alpha},\hat{\beta})\right) - \frac{1}{2}\frac{\hat{B}_i(\hat{\alpha},\hat{\beta}) f_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}{1 - F_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}\right],$$

and the matrix $\hat{C} = \left[\hat{C}_{lj}\right]_{2\times k}$ given by

$$\hat{C}_{1j} = \frac{1}{n\hat{\alpha}} \sum_{i:X_i \in I_j} \delta_i \left[-1 + \widehat{K}_i(\hat{\alpha},\hat{\beta}) v\left(\widehat{K}_i(\hat{\alpha},\hat{\beta})\right) - \frac{\hat{A}_i(\hat{\alpha},\hat{\beta}) f_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}{1 - F_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}\right],$$

$$\hat{C}_{2j} = \frac{1}{n\hat{\beta}} \sum_{i:X_i \in I_j} \delta_i \left[-1 + \frac{1}{2}\frac{1 + 3\frac{\hat{\beta}}{X_i}}{1 + \frac{\hat{\beta}}{X_i}} + \frac{1}{2}\hat{A}_i(\hat{\alpha},\hat{\beta})\hat{B}_i(\hat{\alpha},\hat{\beta}) v\left(\widehat{K}_i(\hat{\alpha},\hat{\beta})\right) - \frac{1}{2}\frac{\hat{B}_i(\hat{\alpha},\hat{\beta}) f_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}{1 - F_Z(\hat{A}_i(\hat{\alpha},\hat{\beta}))}\right].$$

where,

$$\hat{A}_i(\hat{\alpha},\hat{\beta}) = \frac{1}{\hat{\alpha}}\left\{\left(\frac{X_i}{\hat{\beta}}\right)^{\frac{1}{2}} - \left(\frac{\hat{\beta}}{X_i}\right)^{\frac{1}{2}}\right\}, \qquad \hat{B}_i(\hat{\alpha},\hat{\beta}) = \frac{1}{\hat{\alpha}}\left\{\left(\frac{X_i}{\hat{\beta}}\right)^{\frac{1}{2}} + \left(\frac{\hat{\beta}}{X_i}\right)^{\frac{1}{2}}\right\},$$

$$\widehat{K}_i(\hat{\alpha},\hat{\beta}) = \frac{1}{\hat{\alpha}^2}\left\{\frac{X_i}{\hat{\beta}} + \frac{\hat{\beta}}{X_i} - 2\right\}, i = 1, 2, \cdots, n.$$

and $f_Z(u) = c\, g(u^2)$, $F_Z(\cdot)$ are the probability density function and cumulative function of the random variable $Z \approx EC(0, 1;\ g)$ which follows a standard symmetrical distribution in $R$ with the kernel $g(\cdot)$, respectively, and

$$v(u) = -2w(u);\ w(u) = \frac{g'(u)}{g(u)},\ \ u > 0,$$

are the transformations functions of kernel function $g(u)$, and $w'(u)$ is the derivative of $w(u)$ ([15], [16]). Table 3 below shown some transformations functions $w(u)$ and its derivative $w'(u), (u > 0)$ corresponding with kernel $g(u)$ of indicated Elliptic distributions $EC(0, 1;\ g)$.

| | $N(0,1)$ | $t(\nu)$ | $L(0,1)$ | $Log(0,1)$ | $PE(\nu)$ |
|---|---|---|---|---|---|
| $w(u)$ | $\frac{-1}{2}$ | $-\frac{\nu+1}{2(\nu+u)}$ | $-\frac{1}{2\sqrt{u}}$ | $-\frac{1}{2\sqrt{u}}\tanh\left(\frac{\sqrt{u}}{2}\right)$ | $-\frac{\nu}{2}u^{\nu-1}$ |
| $w'(u)$ | 0 | $\frac{\nu+1}{2(\nu+u)^2}$ | $\frac{1}{4u\sqrt{u}}$ | $\frac{\sinh\sqrt{u}-\sqrt{u}}{4u\sqrt{u}[1+\cosh\sqrt{u}]}$ | $-\frac{\nu[\nu-1]}{2}u^{\nu-2}$ |

**Table 3: Transformations functions *w(u)* and its derivative *w'(u), (u > 0)* for kernel *g(u)* of the indicated Elliptic distributions $EC(\mathbf{0},\mathbf{1};\ g)$.**

Chi-squared test for GBS distributions: Under the hypothesis $H_0$, the matrix $\hat{\mathcal{G}}$ is non-degenerate. So, the hypothesis $H_0$ is rejected with approximate significance level $\alpha$ if

$$\boldsymbol{Y}_n^2(\widehat{\boldsymbol{\theta}}_n) > \chi_k^2(\alpha).$$

It is necessary to note that: $PE(\nu) \cong KT(1,0.5,\nu), N(0,\ 1) \cong KT(1,0.5,1), L(0,1) \cong PE(0.5), t(\nu) \cong PVII([\nu+1]/2\ ,\nu), Log(0,1) \cong LIII(1)$.

Thus, the next section, we consider goodness of fit test for following five distributions: GBS-$N(0,\ 1)$ which known as BS distribution, GBS-Laplace, GBS-Logistic, GBS-$t(\nu)$ and GBS-Cauchy distributions.

## 4 Real study

All distributions presented in the next two examples by using R statistics software, we analyze the goodness of fit test for the parametric generalized BS distributions in two studies to a data of breast cancer which set from research of Boag (1949) and the data from a laboratory investigation in which the vaginas of rats were painted with the carcinogen DMBA of Pike (1966).

### 4.1 Analysis of breast cancer data

Boag [10] was presented the survival times for 121 patients treated for cancer of the breast in one particular hospital during the years 1929-1938 which given in table below. The times are in months, and asterisks denote censoring times. This data included 66 observations and 55 censoring times.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.3 | 7.4* | 13.5 | 16.8 | 21.0 | 29.1 | 37* | 41 | 45* | 52 | 60* | 78 | 105* | |
| 129* | 0.3* | 7.5 | 14.4 | 17.2 | 21.1 | 30 | 38 | 41 | 46* | 54 | 61* | 80 | |
| 109* | 129* | 4.0* | 8.4 | 14.4 | 17.3 | 23.0 | 31 | 38* | 41* | 46* | 55* | 62* | |
| 83* | 109* | 139* | 4.0* | 8.4 | 14.4 | 17.3 | 23.0 | 31 | 38* | 41* | 46* | 55* | 62* |
| | 83* | 109* | 139* | 5.0 | 8.4 | 14.8 | 17.5 | 23.4* | 31 | 38* | 42 | 47* | 56 |
| | 65* | 88* | 111* | 154* | 5.6 | 10.3 | 15.5* | 17.9 | 23.6 | 32 | 39* | 43* | 48 |
| | 57* | 65* | 89 | 115* | 6.2 | 11.0 | 15.7 | 19.8 | 24.0 | 35 | 39* | 43* | 49* |
| | 58* | 67* | 90 | 117* | 6.3 | 11.8 | 16.2 | 20.4 | 24.0 | 35 | 40 | 43* | 51 |
| | 59* | 67* | 93* | 125* | 6.6 | 12.2 | 16.3 | 20.9 | 27.9 | 37* | 40* | 44 | 51 |
| | 60 | 68* | 96* | 126 | 6.8 | 12.3 | 16.5 | 21.0 | 28.2 | 37* | 40* | 45* | 51* |
| | 60* | 69* | 103* | 127* | | | | | | | | | |

Firstly, we consider the hypothesis $H_0$ that the survival times for 121 breast cancer patients belongs the Birnbaum-Saunders distribution. In this case, MLE's of the parameters $\boldsymbol{\theta} = (\alpha,\ \beta)^T$ of the BS distribution are $\widehat{\boldsymbol{\theta}}_n = (2.04798, 47.11415)^T$.

Choosing the sub-intervals $k = 6$. The values of $a_j$, the frequency vector $\mathcal{Z}$ and the elements of the matrix $\hat{\mathcal{C}}$ give in table follows.

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $a_j$ | 5.312384 | 9.476277 | 15.352179 | 24.889209 | 41.374625 | 154.000010 |
| $U_j$ | 2 | 8 | 9 | 19 | 15 | 13 |
| $e_j$ | 12.40306 | 12.40306 | 12.40306 | 12.40306 | 12.40306 | 12.40306 |
| $Z_j$ | -0.945732 | -0.400278 | -0.309368 | 0.599721 | 0.236085 | 0.054267 |
| $\hat{C}_{1j}$ | 0.149316 | 0.015361 | -0.008443 | -0.042270 | -0.052283 | -0.058149 |
| $\hat{C}_{2j}$ | -0.003347 | -0.000867 | -0.000594 | -0.001079 | -0.000913 | -0.001021 |

The matrix information of Fisher $\hat{\boldsymbol{I}} = [\hat{\imath}_{ll'}]_{2\times 2}$ and the matrix $\hat{\mathcal{G}}$ are

$$\hat{\boldsymbol{I}} = \begin{bmatrix} 2.62227 & -0.055142 \\ -0.055142 & 0.001280 \end{bmatrix}, \hat{\mathcal{G}} = \begin{bmatrix} 1.203914 & -0.025996 \\ -0.025996 & 0.000562 \end{bmatrix}.$$

We continued this data for GBS distributions for another kernel $g$: Logistic, Laplace, Cauchy and $t(\nu)$ distribution with the same sub-intervals. The results give in table 4 follow.

| Distribution | $\hat{\boldsymbol{\theta}}_n = (\hat{\alpha}, \hat{\beta})^T$ | $\boldsymbol{X}_n^2$ | $\boldsymbol{Q}$ | $\boldsymbol{Y}_n^2$ | $pv_{0.05}$-value |
|---|---|---|---|---|---|
| GBS-Cauchy | $(0.95391, 47.25487)^T$ | 148.2783 | 3.593348 | 151.8717 | 0 |
| BGS-$L(0,1)$ | $(1.36965, 51.99999)^T$ | 9.916332 | 0.772807 | 10.68914 | 0.0984723 |
| GBS-$Log(0,1)$ | $(0.95750, 54.63219)^T$ | 6.586821 | 1.861331 | 8.448152 | 0.2070737 |
| GBS-$t(100)$ | $(1.89550, 50.25370)^T$ | 20.70194 | 8.548100 | 29.25004 | $5.4553.10^{-5}$ |
| GBS-$t(5)$ | $(1.33215, 55.86467)^T$ | 5.652789 | 0.897494 | 6.550284 | 0.3644431 |

**Table 4: MLE's of $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^{\boldsymbol{T}}$, values of $\boldsymbol{Y}_{\boldsymbol{n}}^{\boldsymbol{2}}$ and $p$-values with indicated kernel distributions, data of Boag (1949).**

In this example, we suggest that GBS with kernel $g$: Normal, *t(100)* and Cauchy are strongly rejected and GBS with Logistic, Laplace and $t(5)$ kernels are very well in concordance with the survival times for breast cancer patients treated of Boag. Figure 5 below illustrates the curve of Kaplan-Meier estimate of survival function, with the curve of GBS survival functions corresponding with the kernel indicatives.

### 4.2 Analysis of the times until a carcinoma appeared

Pike [34] gave some data from a laboratory investigation in which the vaginas of rats were painted with the carcinogen DMBA, and the number of days $T$ until a carcinoma appeared was recorded. The data below are for a group of 19 rats (Group 1 in Pike's paper). The two observations with asterisks are censoring times.

143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246, 265, 304, $216^*$, $244^*$.

This data analyzed by Lawless [25], pp.188 where he suggested that probability plots for two parameters Weibull distribution. By using the NRR statistics for two parameters Weibull distribution, we obtain the value of NRR statistics $\boldsymbol{Y}_n^2 = 6.658669$ with $p$-value at level significance $\alpha = 0.05$ is $pv_{0.05} = 0.08361068$.

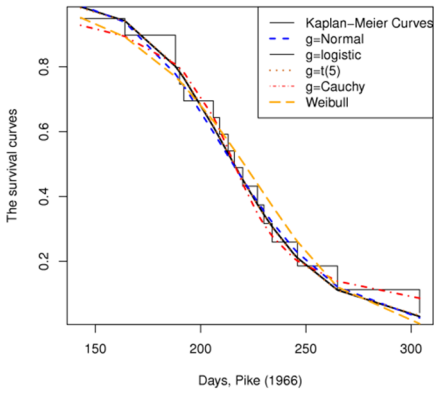



**Figure 5: GBS with indicated kernel, Weibull and Kaplan-Meier estimate of for data of Boag (1949).**

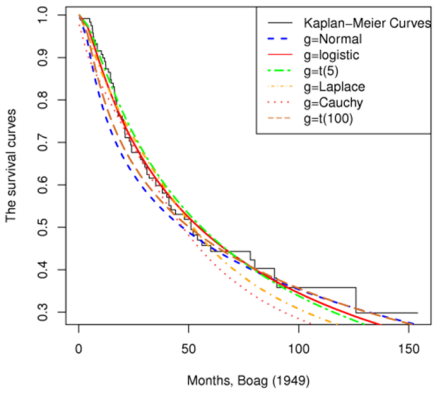



**Figure 6: GBS with indicated kernel and Kaplan-Meier estimate of for the data of Pike (1966).**

We consider next the hypotheses that the data above follows the GBS distributions in the cases of kernel : Standard Normal distribution , Standard Logistic distribution , Standard Cauchy distribution . Choosing grouping intervals , the results given in table 5 below.

| Distribution | | | | | -value |
|---|---|---|---|---|---|
| GBS- | | 3.125849 | 16.35656 | 19.48241 | 0.0006316 |
| GBS- | | 1.000369 | 2.885825 | 3.886194 | 0.4216269 |
| GBS- | | 0.749583 | 0.325662 | 1.075246 | 0.8981795 |
| GBS- | | 0.745397 | 0.3637746 | 1.109172 | 0.8928143 |

**Table 5: MLE's of , values of and *p*-values with indicated kernel distributions, data of Pike (1966).**

We plot the estimated GBS survivor functions correspond indicated kernel and the Kaplan-Meier estimate for the data of Pike (1966) in Figure 6.

In this example, it is clear that the data are the best in concord with GBS-Logistic, GBS-Cauchy and GBS- distributions, it also acceptes for two parameters Weibull distribution. However, these data contradict the BS distribution much very strongly.

## 5 Summary and conclusion

In this paper, we have presented a modifier Chi-squared goodness-of-fit test for generalized Birnbaum-Saunders distributions. The results obtained in our examples show that the considered families are in accordance with lifetimes data. In addition, its hazard rate functions can be uni-modal or bimodal by adjusting the values of its parameters and its kernel . So, it is necessary to use it as a baseline hazard rate functions in the parametric survival model. We would like to thank our colleagues, PhD. R. Tahir and N. Saaidia for valuable comments, which helped us improve the presentation.

## References

[1] Anderson, T. W. and Fang, K. T. Statistical inference in elliptically contoured and related distributions, Allerton Press New York, 1990.

[2] Bagdonavičius, V. B. and Nikulin, M. S. Statistical models to analyze failure, wear, fatigue, and degradation data with explanatory variables. *Communications in Statistics—Theory and Methods,* vol. 38, no. 16-17, pp. 3031-3047, 2009.

[3] Bagdonavičius, V. B. and Nikulin, M. S. Chi-squared tests for general composite hypotheses from censored samples. *Comptes Rendus Mathematique,* vol. 349, no. 3, pp. 219-223, 2011.

[4] Bagdonavičius, V. B. and Nikulin, M. S. Chi-squared goodness-of-fit test for right censored data. *International Journal of Applied Mathematics and Statistics,* vol. 24, no. SI-11A, pp. 30-50, 2011.

[5] Bagdonavičius, V. B., Krupois, J. and Nikulin, M. S. Non-parametric Tests for Censored Data, Wiley, 2011.

[6] Bagdonavičius, V. B., Levuliene, R. J. and Nikulin, M. S. Exact goodness-of-fit tests for shape-scale families and type II censoring. *Lifetime Data Analysis,* pp. 1-23, 2012.

[7] Bagdonavičius, V. B. and M. S. Nikulin, Statistical methods to analyse failures of complex systems in presence of wear, fatigue and degradation: An engineering perspective in accelerated trials. In *Electronic Instrument Engineering, 2008. APEIE 2008. 9th International Conference on Actual Problems of*, 2008.

[8] Balakrishnan, N. Handbook of the logistic distribution, vol. 123, CRC Press, 1992.

[9] Birnbaum, Z. W. and Saunders, S. C. A new family of life distributions. *Journal of Applied Probability,* pp. 319-327, 1969.

[10] Boag, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 11, no. 1, pp. 15-53, 1949.

[11] BolShev, L. N. and Mirvaliev, M. Chi Square Goodness-of-Fit Test for the Poisson, Binomial and Negative Binomial Distributions. *Theory of Probability \& Its Applications,* vol. 23, no. 3, pp. 461-474, 1979.

[12] Cambanis, S., Huang, S. and Simons, G. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis,* vol. 11, no. 3, pp. 368-385, 1981.

[13] Chernoff, H. and Lehmann, E. L. The use of maximum likelihood estimates in $\chi^2$ tests for goodness of fit. *The Annals of Mathematical Statistics,* vol. 25, no. 3, pp. 579-586, 1954.

[14] Desmond, A. Stochastic models of failure in random environments. *Canadian Journal of Statistics,* vol. 13, no. 3, pp. 171-183, 1985.

[15] Díaz-García, J. A. and Leiva-Sánchez, V. A new family of life distributions based on Birnbaum-Saunders distribution. Technical report I-02-17 (PE/CIMAT), Mexico. 2002.

[16] Díaz-García, J. A. and Leiva-Sánchez, V. A new family of life distributions based on the elliptically contoured distributions. *Journal of Statistical Planning and Inference,* vol. 128, no. 2, pp. 445-457, 2005.

[17] Drost, F. C. Asymptotics for generalized chi-square goodness-of-fit tests. *CWI Tracts,* vol. 48, pp. 1-104, 1988.

[18] Dzaparidze, K. O. and Nikulin, M. S. On a modification of the standard statistics of Pearson. *Theory of Probability & Its Applications,* vol. 19, no. 4, pp. 851-853, 1975.

[19] Fang, K. T., Kotz S. and Kai Wang, Ng. Symmetric Multivariate and Related Distributions, Monographs on Statistics and Applied Probability. 36, London: Chapman and Hall Ltd. MR1071174, 1990.

[20] Greenwood, P. E. and Nikulin, M. S. A guide to chi-squared testing. Wiley New York, 1996.

[21] Gupta, A. K. and Varga, T. Elliptically contoured models in statistics. Kluwer Academic Publishers, 1993.

[22] LeCam, L., Mahan, C., and Singh, A. An extension of a theorem of H. Chernoff and EL Lehmann. Recent advances in statistics, 303-332, 1983.

[23] Leiva, V., Barros, M., and Galea, G. A. Influence diagnostics in log-Birnbaum--Saunders regression models with censored data. *Computational Statistics & Data Analysis,* vol. 51, no. 12, pp. 5694-5707, 2007.

[24] Leiva, V., Riquelme, M., Balakrishnan, N. and Sanhueza, A. Lifetime analysis based on the generalized Birnbaum-Saunders distribution. *Computational Statistics & Data Analysis,* vol. 52, no. 4, pp. 2079-2097, 2008.

[25] Lawless, J. F. Statistical models and methods for lifetime data. New Jersey: John Wiley and Sons Publishers, 2003.

[26] Nikulin, M. S. Chi-square test for normality. In *Proceedings of International Vilnius Conference on Probability Theory and Mathematical Statistics*, 1973a.

[27] Nikulin, M. S. On a chi-square test for continuous distribution. *Theory of Probability and its Application,* vol. 19, pp. 638-639, 1973b.

[28] Nikulin, M. S. Chi-square test for continuous distributions with location and scale parameters. *Teoriya Veroyatnostei i ee Primeneniya,* vol. 18, no. 3, pp. 583-591, 1973c.

[29] Nikulin, M. S., Saaidia, N. and Tahir, R. Reliability analysis of redundant systems by simulation for data with unimodal hazard rate functions. *Journal| MESA,* vol. 2, no. 3, pp. 277-286, 2011.

[30] Nikulin, M. S., Saaidia, N. and Tahir, R. Recent Results in the Analysis of Redundant Systems. *Recent Advances in System Reliability,* pp. 181-193, 2012.

[31] Owen, W. J. A new three-parameter extension to the Birnbaum-Saunders distribution. *Reliability, IEEE Transactions on,* vol. 55, no. 3, pp. 475-479, 2006.

[32] Owen, W. J. and Padgett, W. J. A Birnbaum-Saunders accelerated life model. *Reliability, IEEE Transactions on,* vol. 49, no. 2, pp. 224-229, 2000.

[33] Owen, W. J. and Padgett, W. J. Power-law Accelerated Birnbaum-Saunders life models. *International Journal of Reliability, Quality and Safety Engineering,* vol. 7, no. 01, pp. 1-15, 2000.

[34] Pike, M. C. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics,* vol. 22, no. 1, pp. 142-161, 1966.

[35] Rao, K. C. and Robson, B. S. A chi-squabe statistic for goodies-of-fit tests within the exponential family. *Communications in Statistics-Theory and Methods,* vol. 3, no. 12, pp. 1139-1153, 1974.

[36] Sanhueza, A., Leiva, V. and Balakrishnan, N. The generalized Birnbaum–Saunders distribution and its theory, methodology, and application. Communications in Statistics—Theory and Methods, vol. 37, no. 5, pp. 645-670, 2008.

[37] Tahir, R. On Validation Of Parametric Models Applied In Survival Analysis And Reliabilty. Thesis of University Bordeaux I, 2012.

[38] Volodin, I. N. and Dzhungurova, O. A. On limit distributions emerging in the generalized Birnbaum-Saunders model. *Journal of Mathematical Sciences,* vol. 99, no. 3, pp. 1348-1366, 2000.

# IMAGE RECOGNITITON BY MULTIDIMENSIONAL INTERVALS

**G. Tsitsiashvili**

•

IAM, FEB RAS, Vladivostok, Russia
e-mails: guram@iam.dvo.ru

## ABSTRACT

In this paper new algorithm of interval images recognition is suggested. This algorithm gives accuracy solution of considered problem but demands not linear but square complexity by a number of objects. Main motive of such construction is to analyze practically interesting case when there is preliminary silence before predicted events.

## 1. PRELIMINARIES

In [1] the algorithm of interval images recognition is described. In a case of a single index which characterizes objects of first and second classes a minimal interval contained objects of the first class was constructed. Then an object with an index contained in this interval is considered as the first class object. In a case when each object is characterized by several indexes a one dimensional interval in a recognition procedure is replaced by a multidimensional interval constructed as a direct product of one dimensional intervals. An advantage of the interval images recognition before known algorithms is a linear (by a number of all objects and by a number of all indexes) calculation complexity. This algorithm is successfully used in manifold problems of medical geography and ecology, meteorology and fishing [2] – [9].

The algorithm is sufficiently satisfactory when a number of all objects in a sample is about 20-30 and a number of indexes is larger than 3. But in a problem arisen in a mining an emergence of a rock pressure cannot be predicted using a single interval. It means that there are first class objects which have predecessors and there are first class objects which have not predecessors. In this situation the single interval cannot characterize all first class objects because it goes past first class objects which may be described by a preliminary silence.

In this paper the method of interval recognition is developed in a direction of a consideration of such situation. It is based on a construction of few nonintersecting intervals which contain points characterized different first class objects. So first class objects are divided into some subclasses and their recognitions are realized separately. This algorithm is more complicated and has not linear but square complexity by a number of all objects.

## 2. MAIN RESULTS

Assume that first class objects are characterized by the set $B=\left\{b_j,\ 1\le j\le m\right\}$ and second class objects are characterized by the set $A=\left\{a_i,\ 1\le i\le n\right\}$, $-\infty\in A$, $\infty\in A$, of real numbers. Suppose that $m$ is much smaller than $n$. For real numbers $c,d$ ,$c\le d$, define the interval $(c,d)$ by the

condition $(c,d)=\{f: c<f<d\}$ *if* $c<d$ . If $c=d$ then the interval $(c,d)$ consists of the single point $c=d$. Construct the following rule of a recognition of the object $b\in B$. For each number $b\in B$ contrast two
numbers:

$$k(b)=\max\{a\in A: a\le b\},\ r(b)=\min\{a\in A: a\ge b\}.$$

As a result we construct for each number $b\in B$ the interval $(k(b),r(b))$.

**Theorem 1.** If $b_i,b_j\in B$, then the intervals $(k(b_i),r(b_i))$ , $(k(b_j),r(b_j))$ coincide or not intersect .

**Proof.** Assume that between the points $b_i,b_j$ there are not points of the set $A$. Then by a construction the intervals $(k(b_i),r(b_i))$, $(k(b_i),r(b_i))$ coincide. Vice versa if between the points $b_i,b_j$ there are points of the set $A$ then by a construction the intervals $(k(b_i),r(b_i))$, $(k(b_i),r(b_i))$ not intersect. So the points of the set $B$ are divided into classes of an equivalence by their belonging to coincide intervals.

Suppose now that the set $A$ consists of $n$ objects and its each object $i$ is characterized by $l$ - dimensional vector $a_i=(a_i^1,...,a_i^l)$. Analogously assume that the set $B$ consists of $m$ objects and its each object $j$ is characterized by $l$ - dimensional vector $b_i=(b_j^1,...,b_j^l)$. Define the interval $(k(b_j^t),r(b_j^t))$ by the equality

$$k(b_j^t)=\max\left\{a_i^t: a_i^t\le b_j^t,\ 1\le i\le n\right\},\ r(b_j^t)=\min\left\{a_i^t: a_i^t\ge b_j^t,\ 1\le i\le n\right\}.$$

Using these intervals construct $l$ - dimensional interval which is its direct product. $\otimes_{t=1}^{l}(k(b_i^t),r(b_i^t))$.

**Theorem 2.** *If* $1\le i\ne j\le m$, then $l$-dimensional intervals $\otimes_{t=1}^{l}(k(b_i^t),r(b_i^t))$, $\otimes_{t=1}^{l}(k(b_j^t),r(b_j^t))$ coincide or not intersect.

**Proof.** Indeed, by a construction for any $t$, $1\le t\le l$, one dimensional intervals $(k(b_i^t),r(b_i^t))$ , $(k(b_j^t),r(b_j^t))$ coincide or not intersect. If these one dimensional intervals for all $t$, $1\le t\le l$ coincide then their direct products $\otimes_{t=1}^{l}(k(b_i^t),r(b_i^t))$, $\otimes_{t=1}^{l}(k(b_j^t),r(b_j^t))$ coincide also. In opposite case there is $t$ so that appropriate intervals not intersect and consequently their direct products not intersect also. Consequently vectors from the set B are divided on subsets (equivalence classes) by their belonging to coincide $l$-dimension intervals.

Suppose that $(l+1)$ - dimensional vectors arrive in recognition system. The first component equals zero if this vector belongs to the set $A$ and equals one if this vector belongs to the set $B$. Assume that on the step 0 two $(l+1)$ - dimensional vectors $(0,+\infty,...,+\infty)$, $(0,-\infty,...,-\infty)$ are introduced into the system. Further on the step $n>0$ single vector $(\alpha_n,c_{n,1},...,c_{n,l})$ arrives. Denote $n_0$ the first vector for which $\alpha_n=1$. Then the first multidimensional interval containing the vector $(c_{n_0,1},...,c_{n_0,l})$ is constructed.

Further assume that on the step $n>n_0$ the vector $(\alpha_n,c_{n,1},...,c_{n,l})$ arrives in recognition system. Suppose that $\alpha_n=0$. Then if the vector $(c_{n,1},...,c_{n,l})$ does not belong to constructed intervals then the system of these intervals conserves. If $(c_{n,1},...,c_{n,l})$ belongs to one of constructed intervals then this interval is divided onto subintervals by described rule.

Assume that $\alpha_n=1$ then if the vector $(c_{n,1},...,c_{n,l})$ does not belong to constructed intervals then new interval containing this vector is constructed. If $(c_{n,1},...,c_{n,l})$ belongs to one of constructed intervals then the system of intervals does not change.

## REFERENCES

1. Tsitsiashvili G.Sh., Bolotin E.I. 2010. Construction of fast algorithm of images recognition in application to prognoze problems. *Informatics and control systems.* No. 2. P. 25-27. (In Russian).
2 Bolotin E.I., Tsitsiashvili G.Sh., Golycheva. 2002. Some aspects and perspectives of factor prognoze of epidemics of focus of tick-borne encephalitis foci on base of multidimensional time series. *Parazitology.* T. 36, No. 2. P. 89-95. (In Russian).
3. Bolotin E.I., Tsitsiashvili G.Sh., Burukhina I.G., Golycheva I.V. 2002. Possibilities of factor prognoze of tick-borne encephalitis foci in Primorye . *Parazitology.* T. 36, No. 4. P. 280-285. (In Russian).
4. Bolotin E.I., Tsitsiashvili G.Sh. 2003. Space time prognoze of tick-borne encephalitis foci. Bull. of the Far Eastern Branch of Russian Acad. of Sci. 2003. No.1. P. 5-19. (In Russian).
5. Shatilina T.A., Tsitsiashvili G.Sh, Radchenkova T.V. 2006. Experience of application of interval recognition method to prognoze extremal ice-covering of Tatar channel (Japanese sea). *Meteorogy and hydrology.* No. 10. P. 65-72. (In Russian).
6. Goriainov A.A., Shatilina T.A., Tsitsiashvili G.Sh., Lysenko A.V., Radchenkova T.V. 2007. Clymate causes of decrease of fish resources of Amur salmon in 20-th century. *Far eastern region - fishery.* No. 1,2 (6,7). P. 94-114. (In Russian).
7. Bolotin E.I., Tsitsiashvili G.Sh., Fedorova S.Yu., Radchenkova T.V. 2009. Factor time prognoze of critical levels of infection illnesses. *Human ecology.* No. 10. P. 23-29. (In Russian).
8. Bolotin E.I., Tsitsiashvili G.Sh., Fedorova S.Yu. 2010. Theoretical and practical aspects of factor prognoze of infection illnesses. *Human ecology.* No. 7. P. 42-47. (In Russian).
9. Bolotin E.I., Ananiev V.Yu., Tsitsiashvili G.Sh. 2010. Prognoze of infection illnesses: new approaches. *Population healthy and habitat.* No. 5. P. 15-19. (In Russian).

# ON ESTIMATION OF PARAMETERS BY THE MINIMUM DISTANCE METHOD

V. P. SHULENIN
•
Tomsk State University, Tomsk, RUSSIA
shulenin-vp@rambler.ru

## ABSTRACT

Parameter estimates, constructed by the minimum distance method, are briefly called the *MD*-estimates. The minimum distance method has been proposed by Wolfowitz (1957). An extensive bibliography was compiled and published by Parr (1981). In this paper the effectiveness of the shift parameter estimation based on the use of Cramer - von Mises weighted distance is discussed. The robustness of this kind of *MD*-estimates under various supermodels describing deviations from the Gaussian model is considered. Numerical results are given for the case of contaminated normal distributions.

### Statement of the problem

Let us consider first a case when the statistical model $(\mathrm{X}, \Im_\theta)$ is given in parametric form. $\mathrm{X} = \{\vec{x}\}$ denotes the sample space, the elements of which are realizations $\vec{x} = (x_1,...,x_n)$ of a random vector $\vec{X} = (X_1,...,X_n)$; $\Im_\theta = \{F : F(x,\theta), \theta \in \Theta\}$ is a parametric set of admissible probability distributions for the experiment considered; $X_1,...,X_n$ is a sequence of i.i.d. random variables with the distribution function $F(x,\theta)$ and the density $f(x,\theta)$, $x \in R^1$, $\theta \in \Theta$. The functional form of the distribution is defined up to an unknown parameter (scalar or vector), which belongs to a given parameter set $\Theta$. It is required to construct the estimate of an unknown parameter $\theta \in \Theta$ based on a sample $X_1,...,X_n$ from a distribution $F(x,\theta)$.

.

### The essence of the minimum distance method

If a distance $\rho(F,G)$ between any two distributions, $F, G \in \Im$, is given, then parameter $\theta$ may be estimated by minimization of the distance between the empirical distribution function $F_n(x)$, constructed from a sample $X_1,...,X_n$, and the distribution function $F_\theta(x) = F_X(x,\theta)$ adopted in the model $(\mathrm{X}, \Im_\theta)$. Thus, for a chosen distance $\rho(F,G)$ *MD*-estimator for $\theta$ is defined as $\hat{\theta} = \arg\min_\theta \{\rho(F_n, F_\theta)\}$. Various distances could be used for constructing *MD*-estimates (see Parr, and Schucany (1980)). For instance, the maximum likelihood method is based on a distance

$$\rho(F_n, F_\theta) = -\int \ln f(x,\theta) dF_n(x).$$

In this paper, we consider the estimates that are based on the weighted Cramer - von Mises distance

$$\rho_W(F_n, F_\theta) = \int [F_n(x) - F_\theta(x)]^2 W_\theta(x, F_\theta)\, dF_\theta(x) \qquad (1)$$

where $W_\theta = W(x, F_\theta)$ is a certain weight function, which may depend on d.f. $F_\theta$ (or on density $f_\theta$).

Assuming that $\rho_W(F_n,F_\theta)$ a differentiable function of the parameter $\theta$, its derivative is $\widetilde{\lambda}_{F_n}(\theta)=\partial\rho_W(F_n,F_\theta)/\partial\theta$. With this notations, the estimation $\theta_n$ for parameter $\theta$ based on the use of weighted Cramer-von Mises distance (1) is a solution of the equation

$$\widetilde{\lambda}_{F_n}(\theta)=-2\int[F_n(x)-F_\theta(x)]\frac{\partial F_\theta(x)}{\partial\theta}W_\theta(x)dF_\theta(x)+\int[F_n(x)-F_\theta(x)]^2\frac{\partial}{\partial\theta}[W_\theta f_\theta(x)]dx \qquad (2)$$

In this paper we consider the *MD* -estimation of the location parameter; in this case $F_\theta(x)=F(x-\theta)$. Let a family of reference distributions be designated as $\Im_0=\{F: F_\theta(x)=F_0(x-\theta),\ \theta\in R^1\}$, where $F_0$ is a distribution with density $f_0$. Rewrite (1) as

$$\rho_{F_n,F_0}(\theta,W)=\int[F_n(x)-F_0(x-\theta)]^2W(x-\theta)dx. \qquad (3)$$

Note that the choice of the weight function $W$ in the form of the density of reference distribution, i.e., in the form $W(x)=f_0(x)$, corresponds to the Cramer-von Mises distance; the choice of the weighting function $W(x)=f_0(x)/F_0(x)(1-F_0(x))$ gives the distance of Anderson-Darling (see for example, Boos (1981), Shulenin (1993a)). Assuming that $\rho_{F_n,F_0}(\theta,W)$ is a differentiable function of the parameter $\theta$, its derivative is $\lambda_{F_n}(\theta)=\partial\rho_{F_n,F_0}(\theta,W)/\partial\theta$. Then the equation $\lambda_{F_n}(\theta)=0$ for the obtaining the *MD* -estimation, may be written in the form

$$\frac{2}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n}-F_0(X_{(i)}-\theta)\right]W(X_{(i)}-\theta)=0, \qquad (4)$$

where $X_{(1)},...,X_{(n)}$ the ordered statistics of the sample $X_1,...,X_n$.

**Asymptotic normality of the *MD* -estimators**

The asymptotic properties of *MD* -estimators were studied by several authors (see, for example, Boos (1981), Wiens (1987), Shulenin (1992)). In this paper, we discuss the asymptotic properties of estimators $\theta_n$ of the parameter of location $\theta$, which, for a given reference d.f. $F_0$, and given weight function $W$, is a solution of equation (4). There are two variants of parameter estimating:

**Version 1**. The distribution function $F$ of the observations $X_1,...,X_n$ is *known* and it coincides with the reference distribution function $F_0$, that is $F=F_0$ (or $F\in\Im_0$).

**Version 2**. The distribution function of the observations is *not known* and it is not necessarily the same as the reference distribution function, that is $F\neq F_0$ (or $F\notin\Im_0$).

Note that the *MD* -estimator $\theta_n$ of the location parameter $\theta$, which is the solution of equation (4), can be written as a functional of the empirical distribution function, in the form of $\theta_n=\theta(F_n)$. Here the functional $\theta(F)$ is defined either by relation

$$\min_\theta\rho_{F,F_0}(\theta,W)=\rho_{F,F_0}(\theta(F),W),$$

or may be given implicitly (as functional $T(F)=\theta(F)$) by expression

$$2\int[F(x+T(F))-F_0(x)]f_0(x)W(x)dx-\int[F(x+T(F))-F_0(x)]^2W'(x)dx=0. \qquad (5)$$

For studying the asymptotic properties of the *MD* -estimators $\theta_n=\theta(F_n)$ for the location parameter $\theta$, we use the approach of Mises (see Serfling, R. J. (1980), Shulenin (2012)). Let us consider the expansion of the form

$$\theta(F_n)=\theta(F)+V_{1n}+R_{1n}, \qquad (6)$$

where $V_{1n}$ is approximation statistics, and $R_{1n}=\theta(F_n)-\theta(F)-V_{1n}$ is the remainder of the expansion (6). Let us start from defining approximation statistics $V_{1n}$ and the remainder $R_{1n}$. It is necessary to

compute the Gateaux differential of the first order $d_1T(F;G-F)$ for functional $T(F)$ defined by (5). Let $F_\lambda = F + \lambda(G-F)$, $0 \le \lambda \le 1$. Replacing the distribution function $F$ in (5) by the d.f. $F_\lambda$, we obtain the expression

$$2\int \{F(x+T(F_\lambda)) + \lambda[G(x+T(F_\lambda)) - F(x+T(F_\lambda))] - F_0(x)\} f_0(x) W(x) dx -$$
$$- \int \{F(x+T(F_\lambda)) + \lambda[G(x+T(F_\lambda)) - F(x+T(F_\lambda))] - F_0(x)\}^2 f_0(x) W'(x) dx = 0.$$

Differentiating the expression on $\lambda$, setting $\lambda = 0$, and taking into account that $d_1T(F;G-F) = \partial T(F_\lambda)/\partial\lambda|_{\lambda=0}$, $T(F_\lambda)|_{\lambda=0} = T(F) = \theta$, we get

$$d_1T(F;G-F) = \frac{\int [G(x)-F(x)]\{[F(x)-F_0(x-\theta)]W'(x-\theta) - f_0(x-\theta)W(x-\theta)\}dx}{\int f(x) f_0(x-\theta) W(x-\theta)dx - \int [F(x)-F_0(x-\theta)] f(x) W'(x-\theta)dx}$$

From this expression, after replacing $G$ by the empirical d.f. $F_n$, we get an approximation for statistics $V_{1n}$:

$$V_{1n} = d_1T(F;F_n - F) = n^{-1}\sum IF(X_i;F,F_0,W).$$

Here $IF(u;F,F_0,W) = d_1T(F;\Delta_u - F)$, $0 \le u < \infty$, is the Hampel influence function for the $MD$-estimator $\theta_n = \theta(F_n)$ of the location parameter $\theta$, which for a given reference d.f. $F_0$ and given weight function $W$ is a solution of equation (4). Note that the expression for the influence function also follows from the above formula by replacing d.f. $G$ by degenerated at the point $u$ distribution function $\Delta_u$. The resulting formulas, together with the expansion (6), are the basis for the proof of asymptotic normality of the $MD$-estimators, which are solutions of the equation (4).

Note that the general conditions of regularity (which impose restrictions on the behavior of the tails of d.f. $F$ and the weight function $W$) under which the expression $\sqrt{n}R_{1n} \to^p 0, n \to \infty$, and for which $MD$- estimator is consistent and asymptotically normal, given in Boos (1981). In addition, the considered here $MD$- estimates belong to the family of $MD_\alpha$- estimates whose asymptotic properties are described in Shulenin (1992).

To facilitate formulating further results, let us denote by $\Im_S$ a family of absolutely continuous symmetric distributions. Let the class of weight functions $W_S$ consists of differentiable and even functions, that is $W(-x) = W(x)$ and

$$\int \{F(x)(1-F(x))\}^p W(x+c)dx < \infty,\ p > 0,\ c \in (-\infty, +\infty).$$

**Theorem**. Let $(F,F_0) \in \Im_S, W \in W_S$. Then, under fulfillment of the inequalities $0 < \sigma^2(F;F_0,W) = \int IF^2(x;F,F_0,W)dF(x) < \infty$,
the asymptotic expression can be written in the form of

$$L\{\sqrt{n}[\theta(F_n) - \theta(F)]/\sigma(F;F_0,W)\} = N(0,1), n \to \infty.$$

The asymptotic variance of $MD$-estimate with the reference d.f. $F_0$ and the weight function $W$ under the distribution $F$ of observations $X_1,...,X_n$, is equal to $D(F;F_0,W) = \sigma^2(F;F_0,W)/n$; the Hampel influence function $IF(u;F,F_0,W) = -IF(-u;F,F_0,W)$ for the $MD$-estimates is calculated by formulas

$$IF(u;F,F_0,W)=A_{F,F_0}(u;W)/B_{F,F_0}(W),\ 0\le u<\infty, \tag{7}$$

$$A_{F,F_0}(u;W)=\int_0^u W(x)dF(x)-W(u)[F(u)-F_0(u)], \tag{8}$$

$$B_{F,F_0}(W)=\int_{-\infty}^{\infty} f_0(x)W(x)dF(x)-\int_{-\infty}^{\infty}[F(x)-F_0(x)]W'(x)dF(x). \tag{9}$$

The proof can be found in Boos (1981), Wiens (1987), Parr and de Wet (1981).

Note that for the first version of parameter estimation $\theta$, when $F\in\Im_0$ the influence function $IF(u;F,W)$, $0\le u<\infty$ is given by

$$IF(u;F,W)=\frac{\int_{-\infty}^{+\infty}\{F(x)-I[u\le x]\}W(x)dF(x)}{\int_{-\infty}^{+\infty}f^2(x)W(x)dx}=\frac{\int_0^u W(x)dF(x)}{\int_{-\infty}^{+\infty}f(x)W(x)dF(x)}=$$

$$=J^{-1}(F,W)\int_0^u f(x)W(x)dx\quad,0\le u<\infty\ , \tag{10}$$

and the asymptotic variance of $\sqrt{n}\,MD$-estimate is given by

$$\sigma^2(F,W)=\frac{\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty}\{F(y)-I[u\le y]\}W(y)dF(y)\right)^2dF(u)}{\left(\int_{-\infty}^{+\infty}f(x)W(x)dF(x)\right)^2}=\frac{\int_{-\infty}^{+\infty}\left(\int_0^x W(y)dF(y)\right)^2dF(x)}{\left(\int_{-\infty}^{+\infty}f^2(x)W(x)dx\right)^2}. \tag{11}$$

**Efficient *MD* - estimators**

For the first version of parameter $\theta$ estimation (when the distribution function $F$ of the observations $X_1,...,X_n$ is known and coincides with the reference function of a symmetric distribution $F_0$) there is an effective parameter estimate in the class of $MD$ - estimators. Its asymptotic variance is equal to the inverse of the Fisher information $I(f_0)$ about $\theta$ in distribution $F_0(x-\theta)$ with the density $f_0$. This score is determined by the effective weight function of the form

$$W^*(x)=a\,\frac{d^2\{-\ln f_0(x)\}}{dx^2}\cdot\frac{1}{f_0(x)}. \tag{12}$$

This effect was observed earlier in Boos (1981), Parr, De Wet (1981). Correctness of this fact can be seen from the following. Let us denote $\psi(x)=-f'(x)/f(x)$; then $\psi'(x)=d^2\{-\ln f(x)\}/dx^2$, and the expression (12) can be rewritten, taking into account that $F=F_0$, as $W(x)=a\psi'(x)/f(x)$. Substituting this weight function $W\in W_S$ in (11), and taking into account that $F\in\Im_S$, $\psi(0)=0$, we obtain

$$\sigma^2(F,W)=\frac{\int_{-\infty}^{+\infty}\left(\int_0^x W(y)dF(y)\right)^2dF(x)}{\left(\int_{-\infty}^{+\infty}f^2(x)W(x)dx\right)^2}=\frac{a^2\int_{-\infty}^{+\infty}\psi^2(x)dF(x)}{a^2\left(\int_{-\infty}^{+\infty}\psi'(x)dF(x)\right)^2}=\frac{I(f)}{I^2(f)}=\frac{1}{I(f)}.$$

**Example 1**. Note that the use of (12) allows to find the distribution function $F_0$, under which the Cramer - von Mises $MD$-estimator with the weighting function $W(x)=f_0(x)$ produces asymptotically efficient parameter estimates. In fact, solving the differential equation $d^2\{-\ln f_0(x)\}/dx^2=a\cdot f_0^2(x)$ under $W(x)=f_0(x)$, we obtain the density of the form

$$f_0(x) = 2/[\pi(e^x + e^{-x})] = (1/\pi)\sec h(x)\,,\ x \in R^1,$$

with the distribution function

$$F_0(x) = (2/\pi)\,arctg(e^x)\,,\ x \in R^1,$$

which is called the hyperbolic secant. Note that the Fisher information for the parameter $\theta$ in the density $f_0(x) = (1/\pi)\sec h(x)$ is hyperbolic secant as for the Cauchy distribution, and is equal $I(f_0) = 1/2$. Hence $\sigma^2(F_0, W = f_0) = 2$. Note, in addition, that the influence function for $MD$ - estimation with the weighting function $W \equiv 1$, with $F = F_0$ is *limited* and defined as

$$IF(x; F_0, W \equiv 1) = \frac{F_0 - (1/2)}{\int_0^1 f_0(F_0^{-1}(t))dt} = \frac{(2/\pi)arctg(e^x) - (1/2)}{(2/\pi^2)} = \pi\, arctg(e^x) - (\pi^2/4)\ ,x \in R^1.$$

The asymptotic variance of the $MD$ - estimate with weight function $W \equiv 1$ and $F = F_0$ is the same as the asymptotic variance of Hodges - Lehmann estimate $HL$, and for distribution $F_0(x) = (2/\pi)\,arctg(e^x)$ is given by

$$\sigma^2(F_0, W = 1) = \frac{1}{12\left(\int_0^1 f_0(F_0^{-1}(t))dt\right)^2} =$$

$$= \frac{1}{12\left((2/\pi)\int_0^1 \sin(\pi t/2)\cos(\pi t/2)dt\right)^2} = \frac{\pi^4}{48} \approx 2{,}029 = \sigma^2(F_0, HL)\quad .$$

**Example 2**. Let the supermodel $\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}$ be a finite set of distributions, where $F_{(1)} = \Phi$ is the standard normal distribution, Fisher information $I(f_{(1)}) = 1$; $F_{(2)}$ is logistic, $I(f_{(2)}) = 1/3$; $F_{(3)}$ is Laplace, $I(f_{(3)}) = 1$; $F_{(4)}$ is Cauchy , $I(f_{(2)}) = 1/2$; $F_{(5)}$ is hyperbolic secant, $I(f_{(5)}) = 1/2$. Optimal weight functions of the form (12) for these distributions are given in Table 1 and in Figure 1.

Table 1. Optimal weight functions of the form $W^*(x) = a \cdot \psi'(x)/f(x)$

| $F_{(1)}$ | $F_{(2)}$ | $F_{(3)}$ | $F_{(4)}$ | $F_{(5)}$ |
|---|---|---|---|---|
| $W^*_{(1)}(x) = 1/\phi(x)$ | $W^*_{(2)}(x) \equiv 1$ | $W^*_{(3)}(x) = 2e^{\lvert x\rvert}\delta(x-0)$ | $W^*_{(4)}(x) = (1-x^2)/(1+x^2)$ | $W^*_{(5)}(x) = (2/\pi)(e^x + e^{-x})^{-1}$ |

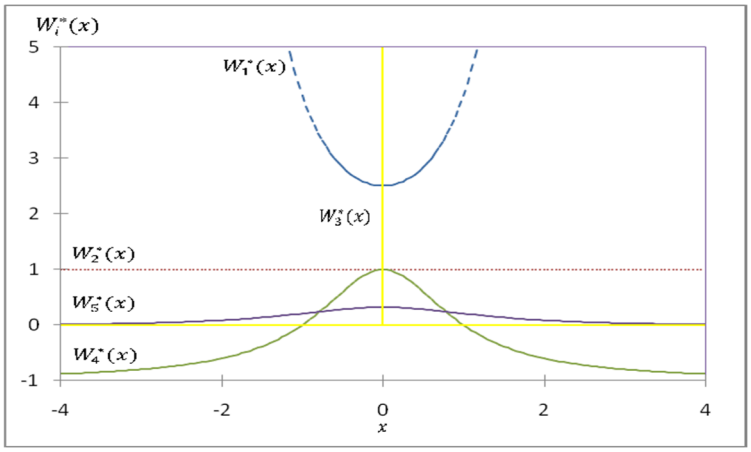



**Fig.1. Optimal weight functions-estimates for $F \in \Im_S^*$**

Note that the asymptotic variance of $MD$- estimate with the reference distribution $F_0(x)=F(x)$ and the weight function $W(x)=1/f(x)$ coincides with the asymptotic variance of the sample mean $\overline{X}$, and is calculated by the formula

$$\sigma^2(F,W=1/f)=\frac{\int_{-\infty}^{+\infty}\left(\int_0^x W(y)dF(y)\right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty}f^2(x)W(x)dx\right)^2}=\frac{\int_{-\infty}^{+\infty}\left(\int_0^x (1/f(y))f(y)dy\right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty}f^2(x)(1/f(x))dx\right)^2}=\int_{-\infty}^{\infty}x^2dF(x).$$

For the weight function $W(x)=1/\phi(x)$, where $\phi(x)$ is the standard normal density, $MD$- estimator is an efficient estimate of the location parameter $\theta$ of the normal distribution, but it has, like the sample mean $\overline{X}$, the *unlimited* influence function $IF(x;\Phi,W=1/\phi)=x$, $x\in R^1$ and its sensitivity to gross errors is not limited, that is $\gamma^*(\Phi,W=1/\phi)=\infty$. Note also that the choice of the weighting function $W(x)\equiv 1$ leads to asymptotically efficient $MD$- estimator for the logistic cdf $F_{(2)}$ (the variance in this case coincides with the variance of $HL$- estimator), and the absolute efficiency of the $MD$- estimator with weight function $W(x)=f_{(2)}(x)$ is equal to $AE(F_{(2)},W=f_{(2)})=[3{,}036(1/3)]^{-1}=0{,}988$. Recall that for the logistic distribution $F_{(2)}$ with density $f_{(2)}$, the equality $f_{(2)}=F_{(2)}(1-F_{(2)})$ holds, and therefore, the choice of the weighting function in the form inherent in $MD$- estimation based on the use of the Anderson-Darling distance, $W(x)=f_0/F_0(1-F_0)$, also leads to an effective $MD$-estimation for the logistic distribution. For the Laplace distribution with density $f_{(3)}(x)=(1/2)\exp(-|x|), x\in R^1$ function $\psi(x)=-f'_{(3)}(x)/f_{(3)}(x)=sign(x)$ and, therefore, the optimal weight function $W^*(x)=a\cdot\psi'(x)/f(x)$ defined by (12), takes the form $W^*_{(3)}(x)=\{sign(x)\}'/f_{(3)}(x)=\delta(x-0)/f_{(3)}(x)=2e^{|x|}\delta(x-0)$. Using this expression for the optimal weight function, and (11), one may see that the asymptotic variance of $MD$- estimate coincides with the asymptotic variance of the sample median $\overline{X}_{1/2}$, which is asymptotically efficient estimate of parameter $\theta$ for the Laplace distribution. In fact, from (11) with the weighting function $W(x)=\delta(x-0)/f(x)$, we obtain:

$$\sigma^2(F,W)=\frac{\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty}\{F(y)-I[u\le y]\}W(y)dF(y)\right)^2 dF(u)}{\left(\int_{-\infty}^{+\infty}f(x)W(x)dF(x)\right)^2}=$$

$$=\frac{\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty}\{F(y)-I[u\le y]\}\delta(y-0)dy\right)^2 dF(u)}{\left(\int_{-\infty}^{+\infty}f(x)\delta(x-0)dx\right)^2}=$$

$$=\frac{\int_{-\infty}^{+\infty}\{F(0)-I[u\le 0]\}^2 dF(u)}{f^2(0)}=$$

$$=\frac{(1/4)-\int_{-\infty}^{+\infty}I[u\le 0]dF(u)+\int_{-\infty}^{+\infty}I^2[u\le 0]dF(u)}{f^2(0)}=\frac{1}{4f^2(0)}=\sigma^2(F,\overline{X}_{1/2}).$$

Note that for the Cauchy distribution the optimal weight function $W^*_{(4)}(x) = a(1-x^2)/(1+x^2)$ is negative outside the interval $[-1, 1]$. This fact can be explained as follows. From (10) it follows that the weight function $W$ is expressed through the derivative of the influence function in the form $W(u) = J(F,W)\, IF'(u;F,W)/f(u)$, $0 \le u < \infty$. So, to "reduce" the influence outliers on the *MD* -estimation, it is necessary its influence function to decrease for large values of the argument and, consequently, the weight function should be *negative*, as is observed for the optimal weight function $W^*_{(4)}(x) = a(1-x^2)/(1+x^2)$ for the Cauchy distribution.

**Example 3**. Consider the family of t-distributions $\mathfrak{I}_r \in \mathfrak{I}_S$, for which the density distribution $f_r(x)$ with degrees of freedom $r$ can be written as

$$f_r(x) = A(r)(1+(x^2/r))^{-(r+1)/2}, x \in R^1, A(r) = \Gamma((r+1)/2)/\sqrt{r\pi}\,\Gamma(r/2).$$

Using (11), we can see that the optimal weight function for this family of distributions is calculated by the formula

$$W^*_r(x) = a \cdot r^{-(r+1)/2}(r+1)A^{-1}(r)(r-x^2)(r+x^2)^{(r-3)/2}.$$

Hence, under $r=1$ we obtain the optimal weight function for Cauchy distributions as $W^*_r(x)|_{r=1} = a \cdot 2\pi(1-x^2)/(1+x^2) = W^*_{(4)}(x)$. The case of $r \to \infty$ corresponds to the normal distribution. Given that under $r \to \infty$, the expressions $A(r) \to 1/\sqrt{2\pi}$ and $(1+(x^2/r))^{-(r+1)/2} \to e^{-x^2/2}$ are hold, from the general formula, we obtain:

$$\lim_{r\to\infty} W^*_r(x) = a \cdot \sqrt{2\pi}\, \exp(x^2/2) = a \cdot 1/\phi(x) = W^*_{(1)}(x).$$

**Robustness of the MD-estimators**

To study the properties of robustness, we consider two types of supermodels that describe deviations from the Gaussian model of observations. The first supermodel $\mathfrak{I}^*_S$, which was used in Example 2, is defined as a finite set of given distributions, that is,

$$\mathfrak{I}^*_S = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}.$$

Second supermodel $\mathfrak{I}_{\varepsilon,\tau}(\Phi)$ called Gaussian model with scale contamination, is determined as

$$\mathfrak{I}_{\varepsilon,\tau}(\Phi) = \{F : F_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\,\Phi(x/\tau)\}, 0 \le \varepsilon \le 1, \tau \ge 1,$$

where $\Phi(x)$ is the standard normal distribution function with density $\phi(x)$, $\varepsilon$ - the proportion of sample contamination, and $\tau$ is a parameter of the scale contamination.

**Example 4. The first option**. First, we consider the properties of *MD* -estimators within a supermodel under different types of reference cdf $F_0$ and weighting functions $W$. For the first version of parameter $\theta$ estimation (when the distribution function $F$ is known and equals the reference distribution function $F_0$, that is $F \in \mathfrak{I}_0$), the influence function of *MD* -estimation and its asymptotic variance are given by (10) and (11). Let us consider various types of the weighting function $W \in W_S$.

A) Let $W(x) \equiv 1$, $F(x) = F_0(x)$. Under these conditions the *MD* -estimators with the weight function $W(x) \equiv 1$ are $B$ -robust, that is, they have *limited* influence functions, which are defined as $IF(x;F,W \equiv 1) = \{2F(x)-1\}/2\int f^2(x)dx$. In the Gaussian case $F = \Phi$, the influence function is given by $IF(x;\Phi,W \equiv 1) = \sqrt{\pi}[2\Phi(x)-1]$. The sensitivity to gross errors

$\gamma^*(F,T)=\sup_x|IF(x;F,T)|$ of $MD$-estimators with the weighting function $W(x)\equiv 1$ is equal to $\gamma^*(\Phi,W\equiv 1)=\sqrt{\pi}\approx 1{,}77$.

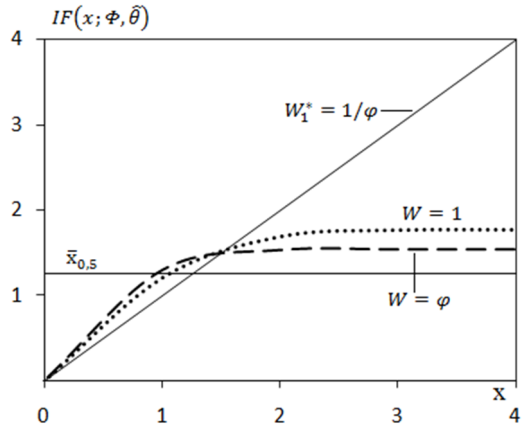



**Fig. 2. Influence function of $MD$-estimators for the normal distribution**

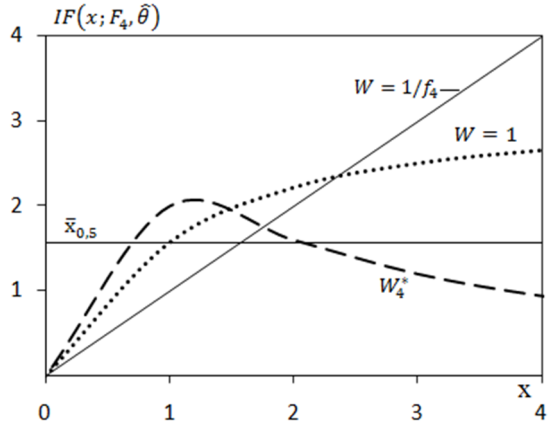



**Fig. 3. Influence function of $MD$-estimators for the Cauchy distribution**

**B)** Let the weight function coincides with the reference density, $W(x)=f_0(x)$, and $F(x)=F_0(x)$. Under these assumptions the asymptotic variance of the $MD$-estimation is given by

$$\sigma^2(F,W=f)=\frac{\int_{-\infty}^{+\infty}\left(\int_0^x f^2(y)dy\right)^2 dF(x)}{\left(\int_{-\infty}^{+\infty}f^3(x)dx\right)^2}.$$

Note that for a Gaussian distribution $F(x)=\Phi(x)$ and the weight function $W(x)=\phi(x)=(1/\sqrt{2\pi})\exp\{-x^2/2\}$ we obtain from (10) the *limited* influence function

$$IF(x;\Phi,W=\phi)=(\sqrt{3\pi}/2)\widetilde{\Phi}(x)=(\sqrt{3\pi}/2)[\,2\Phi(x\sqrt{2})-1\,],x\in R^1,$$

where $\widetilde{\Phi}(x)$ is the Laplace function given by

$$\widetilde{\Phi}(x)=(2/\sqrt{\pi})\int_0^x\exp\{-x^2\}dx,\ \widetilde{\Phi}(x)=2\Phi(x\sqrt{2})-1,\ x\ge 0,\qquad \Phi(x)=(1/\sqrt{2\pi})\int_{-\infty}^x\exp\{-x^2/2\}dx.$$

Sensitivity to gross errors $\gamma^*(F,T)$ of $MD$-estimation, with the weighting function $W(x)=\phi(x)$, is equal to $\gamma^*(\Phi,W=\phi)=\sqrt{3\pi}/2=1{,}53$. In this case, the asymptotic variance of $\sqrt{n}$ $MD$-estimation is

$$\sigma^2(\Phi,W=\phi)=2\int_0^\infty IF^2(x;\Phi,W=\phi)\,d\Phi(x)=\frac{3\pi}{2}\cdot\frac{1}{\sqrt{2\pi}}\int_0^\infty\widetilde{\Phi}^2(x)e^{-x^2/2}dx=$$

$$=(3/2)\,arctg(2/\sqrt{5})=1{,}095.$$

The asymptotic variance of the $MD$-estimators for the cases (A) and (B) were calculated for the following distributions: $F_{(1)}$ - normal, $F_{(2)}$ - logistic, $F_{(3)}$ - Laplace, $F_{(4)}$ - Cauchy, $F_{(5)}$ -hyperbolic secant. Numerical calculations derived from formulas for $F_0=F_{(i)},i=1,\ldots,5$ and with different weight functions are shown in Table 2.

Table 2. The asymptotic variance of $\sqrt{n}\,MD$-estimators for the supermodel $\Im_S^*$ at $F_0 = F_{(i)}$, $i = 1,...,5$

| The weight function | $F_{(1)} = \Phi$ | $F_{(2)}$ | $F_{(3)}$ | $F_{(4)}$ | $F_{(5)}$ |
|---|---|---|---|---|---|
| $W \equiv 1$ | 1,047 (0,96) | 3,000 **(1,00)** | 1,333 (0,75) | 3,287 (0,61) | 2,029 (0,98) |
| $W_{(i)}(x) = f_{(i)}(x)$ | 1,095 (0,91) | 3,036 (0,99) | 1,200 (0,83) | 2,573 (0,78) | 2,000 **(1,00)** |
| $\widetilde{W}_{(i)}(x) = f_{(i)} / F_{(i)}(1-F_{(i)})$ | 1,035 (0,97) | 3,000 **(1,00)** | 1,262 (0,79) | 2,317 (0,86) | 2,020 (0,99) |
| $W_{(i)}(x) = 1/f_{(i)}(x)$ | 1,000 **(1,00)** | 3,290 (0,91) | 2,000 (0,50) | $\infty$ (0,00) | 2,467 (0,81) |
| $W_{(4)}^*(x) = (1-x^2)/(1+x^2)$ | 1,109 (0,90) | 4,204 (0,71) | 1,230 (0,81) | 2,000 **(1,00)** | 2,103 (0,95) |

The absolute values of efficiency of $MD$-estimates are given in parentheses, they were calculated according to the formula $AE(F,W) = [\sigma^2(F,W)I(f)]^{-1}$. Note that for distributions with "heavy tails" (Cauchy and Laplace), the absolute efficiency of $MD$-estimators depends mainly on the choice of the weighting function $W$. For normal distribution, the optimal weight function is $W_{(1)}(x) = 1/f_{(1)}(x)$. Weight functions $W \equiv 1$ and $W_{(2)}(x) = f_{(2)}/F_{(2)}(1-F_{(2)})$ are optimal for the logistic distribution $F_{(2)}$. Weight function $W_{(4)}^*(x) = (1-x^2)/(1+x^2)$ is optimal for the Cauchy distribution. Weight function $W_{(5)}(x) = f_{(5)}(x)$ is optimal for distribution $F_{(5)}$ - hyperbolic secant.

**Example 5. The second option**. Consider the case when $F \neq F_0$, and the supermodel $\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}$ is the finite set of distributions, $F \in \Im_S^*$. In this case, the asymptotic variance of $\sqrt{n}MD$-estimators under the weight function $W = 1$ is given by

$$\sigma^2(F, F_0, W \equiv 1) = \frac{2\int_0^\infty [F_0(u) - (1/2)]^2 dF(u)}{\left(\int f_0(x) f(x) dx\right)^2}, F \in \Im_S^*. \qquad (13)$$

The numerical values of the asymptotic variance of $\sqrt{n}MD$-estimators for $F \in \Im_S^*$ and the weight function $W = 1$, calculated using the formula (13). are shown in Table 3.

Table 3. Asymptotic variance of $\sqrt{n}MD$-estimators, for $\hat{\theta}_{(i)} = \hat{\theta}(F_0 = F_{(i)}, W \equiv 1)$, $i = 1,...,5$, $F \in \Im_S^*$

| $\hat{\theta} \setminus F$ | $F_{(1)}$ | $F_{(2)}$ | $F_{(3)}$ | $F_{(4)}$ | $F_{(5)}$ | $d(\hat{\theta}, \Im_S^*)$ |
|---|---|---|---|---|---|---|
| $\hat{\theta}_{(1)}$ | 1,047 (0,96) | 3,051 (0,98) | 1,383 (0,72) | 2,911 (0,69) | 2,008 (0,99) | 0,42 |
| $\hat{\theta}_{(2)}$ | 1,016 (0,98) | 3,000 (1,00) | 1,524 (0,66) | 3,679 (0,54) | 2,069 (0,97) | 0,57 |
| $\hat{\theta}_{(3)}$ | 1,059 (0,94) | 3,048 (0,98) | 1,333 (0,75) | 2,957 (0,68) | 2,006 (0,99) | **0,41** |
| $\hat{\theta}_{(4)}$ | 1,046 (0,96) | 3,025 (0,99) | 1,385 (0,72) | 3,290 (0,61) | 2,017 (0,99) | 0,48 |
| $\hat{\theta}_{(5)}$ | 1,031 (0,97) | 3,011 (0,99) | 1,439 (0,70) | 3,276 (0,61) | 2,029 (0,98) | 0,49 |

Note that in the table (3) in parentheses the absolute efficiency estimates are presented, calculated by the formula $AE(F,\hat{\theta}) = \{\sigma^2(F, F_0, W \equiv 1)I(f)\}^{-1}$. In the last column of the table, the defects of the estimates in the supermodel $\Im_S^*$, calculated from (19), are given.

**Note 1**. One of convenient means for comparing qualities of estimates $\hat{\theta}_1,...,\hat{\theta}_k$ of a given parameter $\theta$ of a symmetric distribution $F$ is a concept of defect of the estimator (see, for example, Andrews *at all*. (972), Shulenin (2012)). Let $\hat{\theta}_1,...,\hat{\theta}_k$ be a finite set of asymptotically normal and unbiased estimates of the location parameter $\theta$, based on a sample $X_1,...,X_n$ from the distribution $F$, obeying the expression

$$L\left\{\frac{\sqrt{n}(\hat{\theta}_i-\theta)}{\sigma_F(\hat{\theta}_i)}\right\}=N(0,\ 1),\quad n\to\infty,\quad i=1,...,k.$$

Defect of estimator $\hat{\theta}_i, i=1,...,k$ among the compared parameter estimates $\hat{\theta}_1,...,\hat{\theta}_k$ for a symmetrical distribution $F$ is defined as

$$DE_F(\hat{\theta}_i)=1-\min\{\sigma_F^2(\hat{\theta}_1),...,\sigma_F^2(\hat{\theta}_k)\}/\sigma_F^2(\hat{\theta}_i),i=1,...,k. \qquad (14)$$

Note that if among the estimators $\hat{\theta}_1,...,\hat{\theta}_k$ there is an effective estimate, for which $\sigma_F^2(\hat{\theta}^*)=1/I(f)$ and, therefore, $\min\{\sigma_F^2(\hat{\theta}_1),...,\sigma_F^2(\hat{\theta}_k)\}=1/I(f)$, then the absolute defect of the estimator $\hat{\theta}_i$ is equal to one minus its absolute efficiency, i.e.,

$$ADE_F(\hat{\theta}_i)=1-A\mathcal{E}_F(\hat{\theta}_i),i=1,...,k. \qquad (15)$$

**Note 2**. Studying robustness of compared estimates $\hat{\theta}_1,...,\hat{\theta}_k$ of the location parameter $\theta$ in the supermodel $\mathfrak{I}$ consisting of a finite set of symmetric distributions, $\mathfrak{I}=\{F_1,...,F_r\}$, usually is made by observing the disposition of estimates' defects on the plane of two distributions. The defect for basic (ideal, usually a Gaussian) model is laid along the horizontal axis, and along vertical axis the defects for an alternative model, which is a part of a supermodel $\mathfrak{I}=\{F_1,...,F_r\}$, is laid. With this visual representation of the defects count on the plane of the two distributions, the preference is given to the estimate, which is closest to the origin. As examples, the absolute defects of estimates are presented on the plane of distributions "Gauss-Laplace" and "Gauss-Cauchy", see Figures (4) and (5).

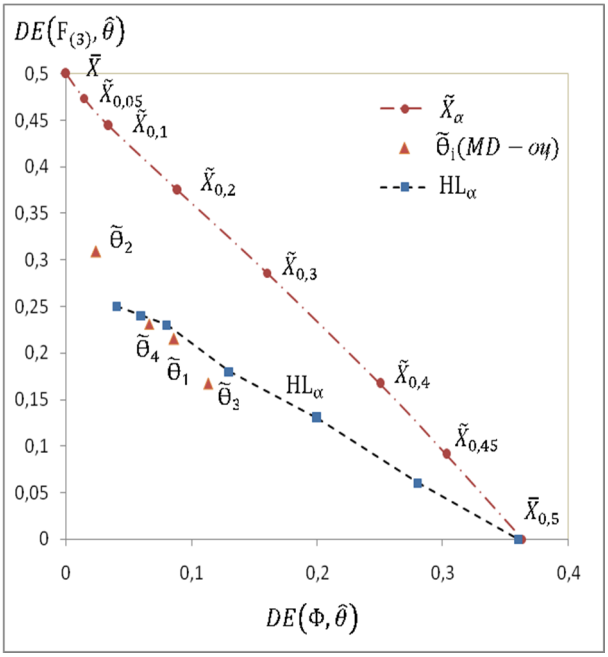



**Fig. 4. Defects estimates in the plane "Gauss-Cauchy"**

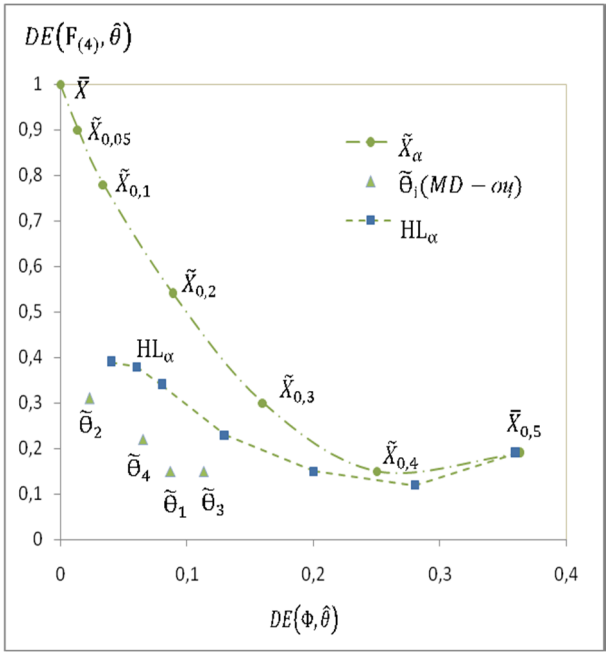



**Fig. 5. Defects estimates in the plane "Gauss-Laplace"**

The advantages of the $MD$ -estimates $\hat{\theta}_{(i)} = \hat{\theta}(F_0 = F_{(i)}, W = f_{(i)})$, $i = 1,...,5$ for $F \in \Im_S^*$ before the family $\widetilde{X}_\alpha$ - Winzor-means and family $HL_\alpha$ -estimates Hodges-Lehmann $0 \le \alpha \le 1/2$.
are clearly seen in these figures (they are placed closer to the origin).

**Note 3**. If we want to draw a conclusion on the preferenced estimator among compared estimates $\hat{\theta}_1,...,\hat{\theta}_k$ of the parameter $\theta$ within the entire supermodel $\Im = \{F_1,...,F_r\}$, we can use the Euclidean metric using the above notations:

$$d(\hat{\theta}_i; \Im) = \left\{ \sum_{j=1}^{r} [DE_{F_j}(\hat{\theta}_i)]^2 \right\}^{1/2}, \tag{16}$$

or

$$Ad(\hat{\theta}_i; \Im) = \left\{ \sum_{j=1}^{r} [ADE_{F_j}(\hat{\theta}_i)]^2 \right\}^{1/2}, \ i = 1,..., k. \tag{17}$$

The preference is given to the estimator $\hat{\theta}_i$ with the minimal value of $d(\hat{\theta}_i; \Im)$, that is

$$d(\hat{\theta}_i; \Im) = \min\{d(\hat{\theta}_1; \Im),...,d(\hat{\theta}_k; \Im)\}. \tag{18}$$

For the supermodel $\Im_S^* = \{F_{(1)}, F_{(2)}, F_{(3)}, F_{(4)}, F_{(5)}\}$, the formula (16) can be written as

$$d(\widetilde{\theta}_{(i)}, \Im_S^*) = \left( \sum_{j=1}^{5} [1 - \{\sigma^2(F_{(j)}, \widetilde{\theta}_{(i)}) I(f_{(j)})\}^{-1}]^2 \right)^{1/2} = \left( \sum_{j=1}^{5} [1 - A\Im(F_{(j)}, \widetilde{\theta}_{(i)})]^2 \right)^{1/2}, i = 1,..., 5. \tag{19}$$

According to the criterion (18), the preference among estimators $\widetilde{\theta}_{(1)},...,\widetilde{\theta}_{(5)}$ in the supermodel $\Im_S^*$, should be given to the $MD$ - estimator for $F_0 = F_{(3)}$ with reference Laplace distribution, and with weight function $W \equiv 1$, since this estimator has the minimum value of

$$d(\hat{\theta}_{(3)}, \Im_S^*) = \min\{d(\hat{\theta}_{(i)}, \Im_S^*),\ i = 1,...,5\} = 0{,}41$$

(see the last column of Table 3). Compare it with that of Hodges-Lehmann $d(HL, \Im_S^*) = 0{,}47$, of $\widetilde{X}_\alpha$ - Winzor-mean $d(\widetilde{X}_{0{,}45}, \Im_S^*) = 0{,}41$; of the sample median $d(\overline{X}_{1/2}, \Im_S^*) = 0{,}51$; of the sample mean $d(\overline{X}, \Im_S^*) = 1{,}14$, Shulenin (2012, p.256).

**Example 6. The second option**. Consider the Gaussian model with a scale contamination $\Im_{\varepsilon,\tau}(\Phi)$. Let the reference distribution be a normal distribution $F_0 = \Phi$, and the distribution of the observations is characterized by normal distribution with a scale contamination, $F \in \Im_{\varepsilon,\tau}(\Phi)$. Under these assumptions, the asymptotic variance of $\sqrt{n}\,MD$ -estimation for $W = 1$ is calculated by the formula

$$\sigma^2(F_{\varepsilon,\tau}, \Phi, W \equiv 1) = \frac{2\int_0^{+\infty} [\Phi(x) - (1/2)]^2 [(1-\varepsilon)\phi(x) + (\varepsilon/\tau)\phi(x/\tau)]dx}{\left( \int_{-\infty}^{\infty} \phi(x)[(1-\varepsilon)\phi(x) + (\varepsilon/\tau)\phi(x/\tau)]dx \right)^2} =$$

$$= \frac{[\pi(1-\varepsilon)/6] + [\varepsilon\, arctg(\tau^2/\sqrt{2\tau^2+1})]}{\{[(1-\varepsilon)/\sqrt{2}] + (\varepsilon/\sqrt{\tau^2+1})\}^2}.$$

For the weight function $W(x)=f_0(x)=\phi(x)$ the asymptotic variance of $\sqrt{n}\,MD$ -estimator is given by

$$\sigma^2(F_{\varepsilon,\tau},\Phi,W=\phi)=\frac{2\int_0^{\infty}\left(\int_0^{u}\phi(x)f_{\varepsilon,\tau}(x)dx-\phi(u)[F_{\varepsilon,\tau}(u)-\Phi(u)]\right)^2 dF_{\varepsilon,\tau}(u)}{\left(\int_{-\infty}^{\infty}\phi^2(x)dF_{\varepsilon,\tau}(x)-\int_{-\infty}^{\infty}\phi'(x)[F_{\varepsilon,\tau}(x)-\Phi(x)]dF_{\varepsilon,\tau}(x)\right)^2}=$$

$$=\frac{1}{4\pi^2\cdot\widetilde{B}^2(\varepsilon,\tau)}\sum_{i=1}^{20}A_i(\varepsilon,\tau),$$

where $\widetilde{B}(\varepsilon,\tau)$ and $A_i(\varepsilon,\tau), i=1,...,20$ are certain functions of the parameters $\varepsilon$ and $\tau$. The numerical values of the asymptotic variance of $\sqrt{n}\,MD$ -estimators for $F\in\Im_{\varepsilon,\tau}(\Phi)$ at different weight functions are given in the Table. 5.

Table 5. The asymptotic variance of $\sqrt{n}\,MD$ -estimators for $F\notin\Im_0$, $F=F_{\varepsilon,\tau}$, $F_0=\Phi$

| $W,\tau\ \backslash\ \varepsilon$ | 0,00 | 0,01 | 0,05 | 0,10 | 0,15 | 0,20 | 0,25 | 0,30 |
|---|---|---|---|---|---|---|---|---|
| $\tau=3$ | 1,047(0,95) | 1,071(0,96) | 1,171(0,97) | 1,307(0,97) | 1,458(0,95) | 1,625(0,94) | 1,811(0,93) | 2,019(0,93) |
| $W\equiv 1,\ 5$ | 1,047(0,95) | 1,078(0,95) | 1,210(0,93) | 1,395(0,90) | 1,607(0,86) | 1,851(0,83) | 2,132(0,80) | 2,459(0,78) |
| $\tau=3$ | 1,095(0,91) | 1,117(0,92) | 1,209(0,93) | 1,333(0,94) | 1,470(0,95) | 1,620(0,95) | 1,786(0,95) | 1,972(0,96) |
| $W=\phi\ ,5$ | 1,095(0,91) | 1,122(0,92) | 1,237(0,91) | 1,393(0,90) | 1,562(0,89) | 1,749(0,88) | 1,956(0,87) | 2,187(0,87) |

The absolute efficiency of $MD$ -estimates calculated using the formula $AE(F_{\varepsilon,\tau},\hat{\theta})=\{\sigma^2(F_{\varepsilon,\tau},W)I(f_{\varepsilon,\tau})\}^{-1}$, where $I(f_{\varepsilon,\tau})$ is the Fisher information about the location parameter of distributions from the supermodel $\Im_{\varepsilon,\tau}(\Phi)$, are given in the table in parentheses.

Fig. (6) shows the absolute efficiency of estimates for $F\in\Im_{\varepsilon,\tau}(\Phi)$. It is clearly seen that $MD$ -estimates with the reference function $F_0=\Phi$ and the weight function $W(x)=\phi(x)$, as well as the weight function $W(x)=1$, provide high absolute efficiency when $0\le\varepsilon\le 0{,}3$. The absolute efficiency of the sample mean $\overline{X}$ decreases sharply, and the median for the sample $\overline{X}_{1/2}$ is slowly growing, remaining at low levels.

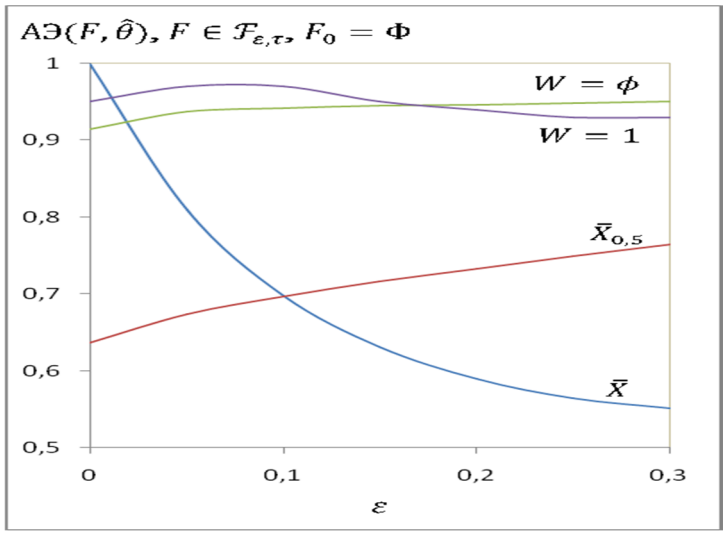



**Fig. 6. Absolute efficiency estimates for $F\in\Im_{\varepsilon,\tau}(\Phi)$, $\tau=3$**

**Example 7. Adaptive version**. Properties of the $MD$ - estimates depend strongly on the choice of the weighting function $W$ for distributions with "heavy tails". Therefore, the study of the properties of the efficiency and robustness of $MD$ -estimates (for the case $F \notin \Im_0$) opens the possibility of an adaptive approach to the choice of the reference distribution $F_0$ and weighting function $W$ within the given supermodel, based on the sample estimates of functionals that determine the "degree of heaviness of tails" of distributions (see Shulenin (1993a)). Adaptive selection of the weighting function can provide the required quality of $MD$ -estimates for a given supermodel.

Let us consider an example of the supermodel $\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\}$. We assume that the proportion of contamination $\varepsilon$ may vary in limits $0 \le \varepsilon \le 0{,}3$, and the scale parameter $\tau$ is $\tau = 3$. For this supermodel with the reference function $F_0 = \Phi$, let us define an adaptive weighting function $\hat{W}$ as

$$\hat{W}(x;\ X_1,...,X_n) = \begin{cases} 1/\phi(x), & 1{,}71 < Q(F_n) \le 1{,}76 \\ 1, & 1{,}76 < Q(F_n) \le 1{,}86 \\ \phi(x), & 1{,}86 < Q(F_n) \le 1{,}91 \end{cases}, \qquad (20)$$

where $Q(F_n)$ is the sample estimate of the functional $Q(F;\nu,\mu)$ which characterizes the "degree of heaviness of the distribution tails" and is defined in Shulenin (1993a). Sample estimate of $Q(F_n)$ is based on a sample $X_1,...,X_n$ and may be written as

$$Q(F_n;\ \nu,\mu) = \frac{m}{k}\left(\sum_{i=n-k+1}^{n} X_{(i)} - \sum_{i=1}^{k} X_{(i)}\right) \Big/ \left(\sum_{i=n-m+1}^{n} X_{(i)} - \sum_{i=1}^{m} X_{(i)}\right), k = [\nu n],\ m = [\mu n]. \qquad (21)$$

Here the parameters $\nu$ and $\mu$ satisfy inequalities $0 < \nu < \mu \le 0{,}5$, $\nu = 0{,}2$, $\mu = 0{,}5$ and $X_{(1)},...,X_{(n)}$ are the order statistics of the sample $X_1,...,X_n$.

Note that the choice of the weighting function in the form of (20), the absolute efficiency of adaptive $MD$ - estimates do not fall below the level of 0.95 when the proportion $\varepsilon$ of contamination is $0 \le \varepsilon \le 0{,}3$. It means that within a given supermodel the absolute efficiency satisfies inequalities $0{,}95 \le A\Im(\Phi_{\varepsilon,\tau}, \hat{W}) \le 1$ if $\tau = 3$, $0 \le \varepsilon \le 0{,}3$, $n \ge 40$ (see Figure 6). If we choose not to adapt the weighting function, and use, for example, the Anderson - Darling weight function in form of $\widetilde{W}(x,\phi) = \phi(x)/\Phi(x)(1-\Phi(x))$, then the absolute efficiency of $MD$ - estimates with such a weight function in the framework of the supermodel $\Im_{\varepsilon,\tau}(\Phi)$ could fall to the level of 0.47.

**Conclusion**

We studied the asymptotic properties of the $MD$ - estimators of the location parameter $\theta$, based on the use of a weighted Cramer - Mises distance. It is shown that these estimates are $B$ - robust, that is, their influence functions are limited, and therefore, they are "protected" against outliers in the sample. For the case $F \in \Im_0$, the optimal weight functions are given that make $MD$ - estimates asymptotically efficient. For the Gaussian model with a scale contamination (for $F \in \Im_{\varepsilon,\tau}(\Phi), \tau = 3$) the absolute efficiency of $MD$ - estimates with the weight function $W \equiv 1$ does not fall below 0.93 at $0 \le \varepsilon \le 0{,}3$, and it increases from 0.91 to 0.96 for the weight function $W \equiv \phi$.

Summarizing, we note that there is a close connection of $MD$ -estimators of parameter $\theta$ with the other robust $M$ -, $L$ -, and $R$ - estimators (see Shulenin and Tarasenko (1994), Shulenin and Serykh (1993), Shulenin(1995)). Properties of $MD$ - estimators in some cases coincide with those of many well-known estimates of the location parameter $\theta$; for example, with the properties of the Hodges - Lehmann estimates, the sample mean and median. Note also that the abovementioned asymptotic results is quite good approximation for properties of $MD$-estimators for sample sizes

$n \geq 20$. This is confirmed by the numerous computer simulation results. Studied properties of the efficiency and robustness of $MD$-estimates open (for the case $F \notin \Im_0$) the possibility to use an adaptive approach to the choice of the reference distribution function $F_0$ and the weighting function $W$ within the given supermodel, based on sample estimates of functionals that determine the "degree of heaviness of tails" of distributions (see Example 7 and Shulenin (2010), Shulenin(2010a)).

**Acknowledgement**

The author expresses gratitude's to Prof. F.P. Tarasenko for valuable comments and help in preparing English version of the paper.

## REFERENCES

Andrews D.F., Bickel P.J., Hampel F.R., Huber P. J., Rogers W.H., Tukey J.W. (1972). Robust estimation of location: survey and advances. Princeton. N.Y.: Princeton Univ. Press. 1972, 375 p.

Bickel, P. J. (1976). Another look at robustness: a review of reviews and some new development. *Scand. J. Statist. Theory and App1*. **3**, 145-168.

Boos, D. D. (1981). Minimum distance estimators for location and goodness of fit. *J. Amer. Statist. Assoc*. **76**, N.375, 663-670.

Parr, W.S. and Schucany, W.R. (1980). Minimum distance and robust estimation. J. Amer. Statist. Assoc. **75,** No. 371, 616 – 624.

Parr, W. C. (1981). Minimum distance estimation: a bibliography. Comm. Statist. **A10**, 1205-1224.

Parr, W.C., De Wet. (1981). On minimum weighted Cramer-von Mists statistical estimation. *Comm. Statist*. **A10(12),** 1149 – 1166.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. N. – Y.: Wiley, 371p.

Shulenin, V. P. (1992). ( Шуленин В. П. Асимптотические свойства и робастность MD-оценок. Теория вероятностей и её применение. 1992 т. 37, в. 4, с. 816-818).

Shulenin, V. P., Serykh, A.P. (1993). (Шуленин В. П. , Серых А. П. Робастные и непараметрические алгоритмы обработки данных физических экспериментов. Известия вузов Физика. 1993, № 10, с.128-136).

Shulenin, V. P.(1993a). (Шуленин В. П. Введение в робастную статистику. Томск: Изд-во Том. ун-та,1993.-227с.)

Shulenin, V. P. and Tarasenko, F. P. (1994). Connections of MD-estimates with classes of robust estimates of location parameter. 12th Prague Conf. on Inform. Theory. August 29 – September 2, 220-223.

Shulenin, V. P. (1995). ( Шуленин В. П. Границы эффективности оценок, построенных методом минимума расстояния Крамера-Мизеса. Известия вузов Физика. 1995, № 9, с.84-89).

Shulenin, V. P. (2010). ( Шуленин В.П. Свойства адаптивных оценок Ходжеса – Лемана в асимптотике и при конечных объемах выборки. Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2010, № 2(11), с. 96 – 112).

Shulenin, V. P. (2010a). (Шуленин В.П. Эффективные и робастные MD-оценки Крамера–Мизеса. ВестникТомского государственного университета. Управление, вычислительная техника и информатика. 2010, № 3(12), с. 107 – 121).

Shulenin, V. P. (2012). (Шуленин В.П. Математическая статистика. Ч. 3. Робастная статистика: учебник. – Томск: Изд-во НТЛ, 2012. – 520 с.)

Wiens, D.P. (1987). Robust weighted Cramer-von Mises estimators of location, with minimax variance in $\varepsilon$ - contamination neighbourhoods. *The Canadian Journal of Statistics*. **15,** N. 3, 269-278.
Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist*. **28**, 75-88.

# MODEL OF MULTILEVEL STOCHASTIC ANALYSIS OF ROAD SAFETY ON REGIONAL LEVEL

**J.Wachnicka**
•
Gdansk University of Technology, Poland
e-mail: joanna.wachnicka@pg.gda.pl

**L. Smolarek**
•
Gdynia Maritime University, Poland
e-mail: leszsmol@am.gdynia.pl

## ABSTRACT

In this paper multilevel approach to the issue of road safety level on the road network of European regions, classified as NUTS 2 in statistical databases of the European Union, has been presented. Following the pattern of many publications on road safety it has been assumed that the risk calculated as the number of death casualties in road accidents per 100,000 inhabitants of a given region has Poisson distribution. Therefore, generalized Poisson model has been assumed in the modelling process. Multilevel stochastic analysis was performed for the studied factor. Then a model was created that took into account the impact of different characteristics available on different level of aggregation, which may be helpful in the actions aimed at improvement of road safety in respective regions.

*Key words: road safety, factors, modelling, Europe, regions*

## 1 INTRODUCTION

In 1771 the first accident involving motor vehicle, a steam powered one, was reported. Since then several hundred million accidents have occurred, in which over 60 million people have died. Despite the activities being carried out with an aim to improve safety, over 1.2m people die on roads each year, and even up to 50 million are injured [1]. This is then a global issue of an epidemiologic nature. Scientists in Western Europe and United States for quite a long time have been searching for the cause of this situation. The issue is complex enough to be addressed by scientists from different fields: economy, mathematics, transport or medicine. However, so far they have focused on researches covering data for respective countries, without going deeper into differences between respective regions of a given country. Most frequently analysed were the figures of changes in number of casualties over time, by means of time series [2]. In researches aimed at finding factors that could influence fatality on roads national product per capita [3] and transport activity [4] have been indicated. Unfortunately, transport activity is unavailable in regional databases. Therefore, the scientists often point to population density as a good substitute index, which may replace transport activity [5]. Literature studies showed that the researchers focus either on national characteristics or on regional characteristics alone, and do not combine both. In this paper the combination of national and regional characteristics in one model has been presented.

## 2 MULTILEVEL MODEL – METHODOLOGY OF THE APPROACH

In order to create model that combines national and regional characteristics, data concerning the number of death casualties in a given region have been collected as well as additional characteristics that describe regions and countries. The reason for this approach was the fact that there were characteristics available on national level that could effectively differentiate safety levels in the regions of respective countries, though unfortunately they were unavailable on regional level of aggregation. On the other hand, respective regions of a given country differ among each other in terms of population density or road network concentration and these elements are worth considering in the model. Since significant dispersion of fatality rate has been observed, it was decided to model demographic index of fatalities on roads (FATALR) calculated as the number of killed per 100,000 inhabitants. The assumption was that the model should have the following formula:

$$FATALR = \alpha \cdot MODEL_{NATIONAL}^{\beta_1} \cdot MODEL_{REGION}^{\beta_2} \cdot NPPC_{NATIONAL}^{\beta_3} \quad (1)$$

where:
*FATALR – demographic rate of fatalities in road accidents in a given region [fat./100 thou. inhab.]*
$MODEL_{NATIONAL}$ *- model for national data*
$MODEL_{REGION}$ *- model for regional data*
$NPPC_{NATIONAL}$ *- model describing changes of average national product per capita*
$\alpha, \beta_1, \beta_2, \beta_3$ *- estimated parameters*

FATALR estimation is based on the assumption that this parameter has Poisson distribution. National models were created on the basis of data from 11 European countries, whereas in the case of regional models in this paper the focus has been on two European countries that substantially differ in terms of road safety level: Great Britain, where the actions for improvement of road safety had a long tradition, and Poland, where the average fatality rate is more than double British figure, likely attributed to cultural, political and economic differences. Histograms of FATALR value in regions of comparable countries in the analysed period of 1999-2008 presented on Fig.1 show that there are no grounds for rejecting the hypothesis of Poisson distribution, frequently assumed in the analyses of safety level [6]. In further analyses, according to this assumption, FATALR index will be alternatively referred to as λ.

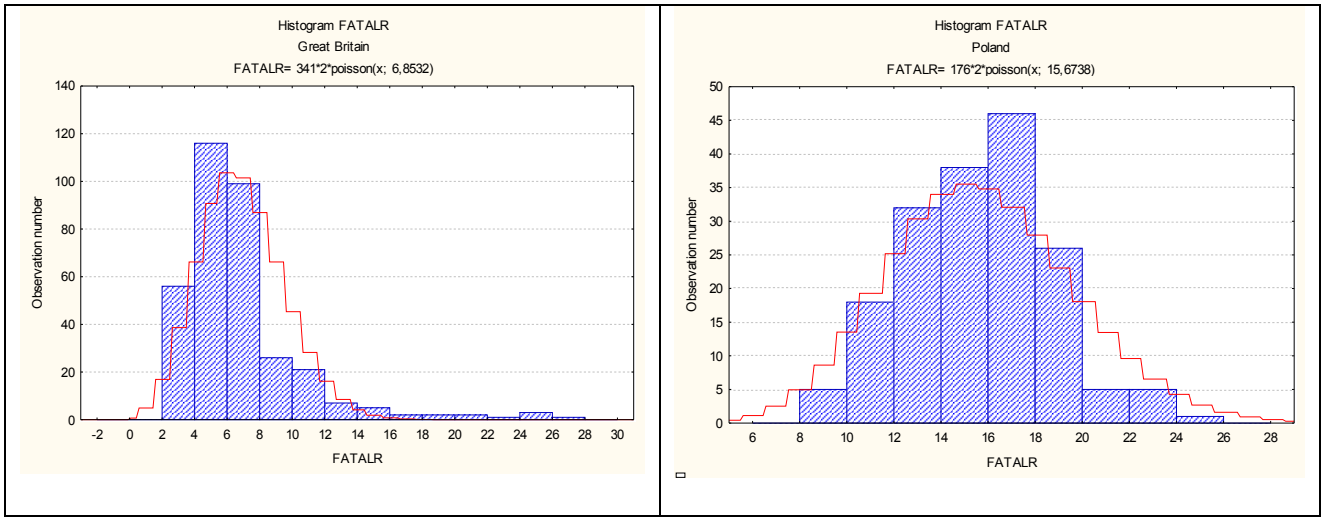



**Figure 1.** Histograms of analysed FATALR indices in the regions of Great Britain and Poland.

## 3 NATIONAL MODEL

Taking into account the assumptions, λ parameter has been evaluated with 95% probability for the set of data from comparable countries. Then the set of independent variables, available only on national level of aggregation, which characterize the country, used for development of descriptive models for evaluated λ parameters ($MODEL_{NATIONAL}$), has been created. In this paper the impact of such factors were analysed as: corruption index - COR (the higher the value, the better a country is perceived, i.e. as less corrupted), percentage of passenger cars older than 10 years – OLD in the total fleet of the country, calculated as average from 10 years.

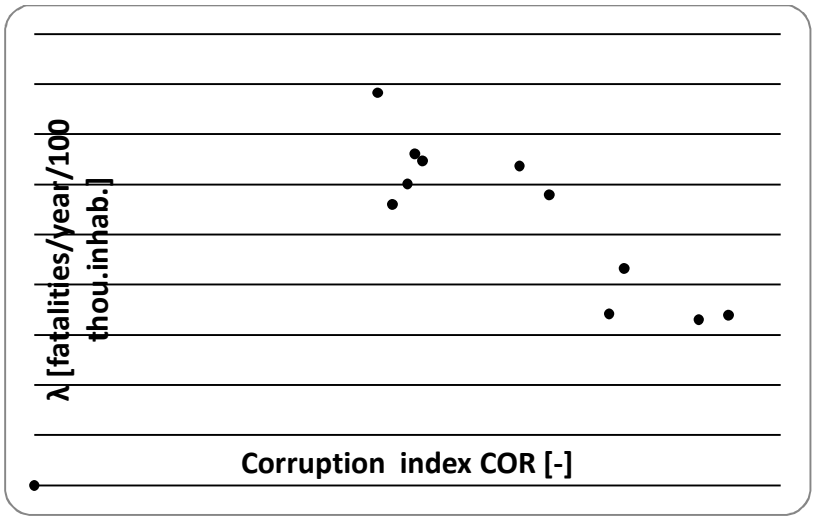



**Figure 2.** Graph of dependence of λ on the value of corruption index in a given country

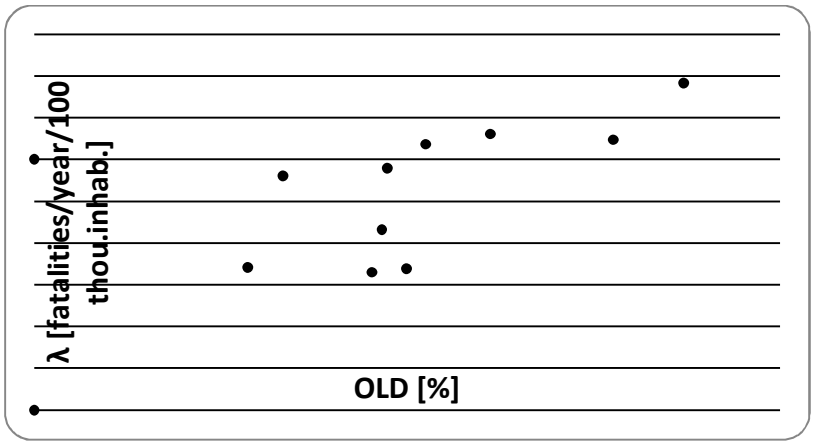



**Figure 3.** Graph of dependence of λ on the percentage of passenger cars older than 10-years old

As Figures 2 and 3 show these indices may impact the studied parameter. The tendency of falling λ parameter with the increase of corruption index has been observed, whereas the reverse correlation has been seen in the case of the percentage of old passenger cars. The developed model has a shape of linear model:

$$\lambda = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 \quad (2)$$

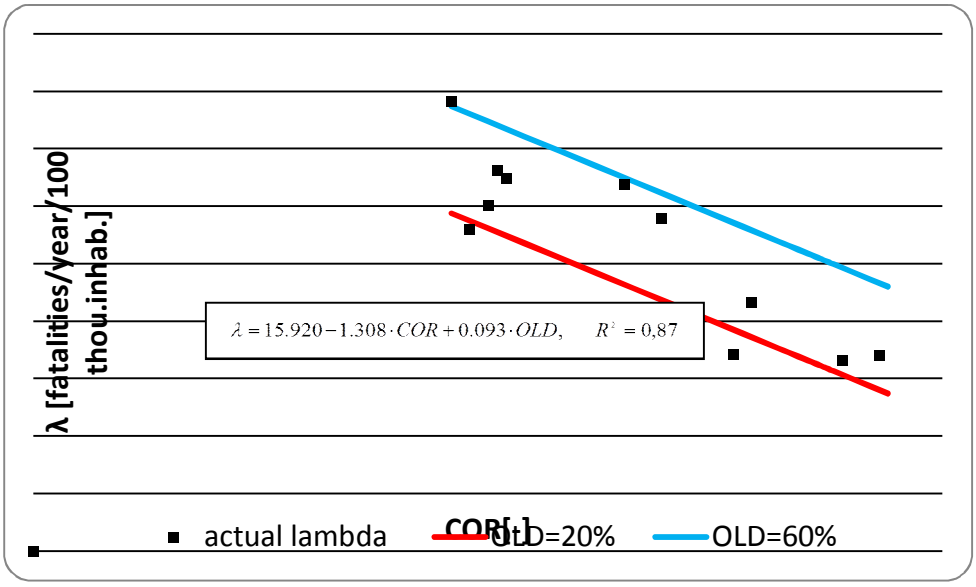



**Figure 4.** Illustration of linear model that takes into account corruption index and the percentage of cars older than 10 years old in the total fleet

Q factor for the model equals 0.87. Together with the increase of corruption index – COR, λ parameter is falling, however if there is a large percentage of old cars in total fleet of passenger cars, then λ parameter will be higher.

## 4 REGIONAL MODEL

The next step was development of models describing the impact of regional characteristics in respective countries on FATALR values in these regions. For this purpose, separated base of regional data was created for each country and the attempts were made to develop a model of impact of respective variables on modelled dependent variable. In the case of all countries one type of the model was checked, which was initially prepared based on joined database from all European regions. General shape of this model has been presented below (3), whereas in table 1, calculated indices in the model in analysed countries have been listed. Cluster analysis allowed specification of classes of correlated variables. In individual models the impact of respective classes have been taken into account through selection of their representatives.

$$MODEL_{REGION} = \alpha \cdot (lnNPPC)^{\beta} \cdot NPPC^{\gamma_6} \cdot DPOP^{\gamma_1} \cdot VEHD^{\gamma_2} \cdot e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)} \quad (3)$$

where:
*$MODEL_{REGION}$ - model for regional data*
*NPPC - average national product per capita in a given region, in a given year [thou. EUR]*
*DPOP - population density [no. of people/km$^2$]*
*VEHD – total vehicle density [veh./km$^2$]*
*CARP – percentage of passenger cars in total fleet of cars [%]*
*ROADC – intensity of the total number of roads per one inhabitant [km/person]*
*UNEMP – unemployment index*
*$\alpha$, $\beta$,$\gamma_1$,$\gamma_2$,$\gamma_3$,$\gamma_4$, $\gamma_5$, $\gamma_6$ - estimated parameters*

Table 1. List of parameters of regional models for selected countries

| Model | $R^2$ | α | β | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\gamma_4$ | $\gamma_5$ | $\gamma_6$ |
|---|---|---|---|---|---|---|---|---|---|
| Great Britain | 0,71 | 0,012 | 32,318 | -0,413 | 0,266 | | 0,364 | -0,039 | -9,501 |
| Poland | 0,54 | 154,361 | -0,179 | 0,123 | -0,149 | -0,031 | 0,358 | | |

On Figures 5 and 6 the regional model has been shown

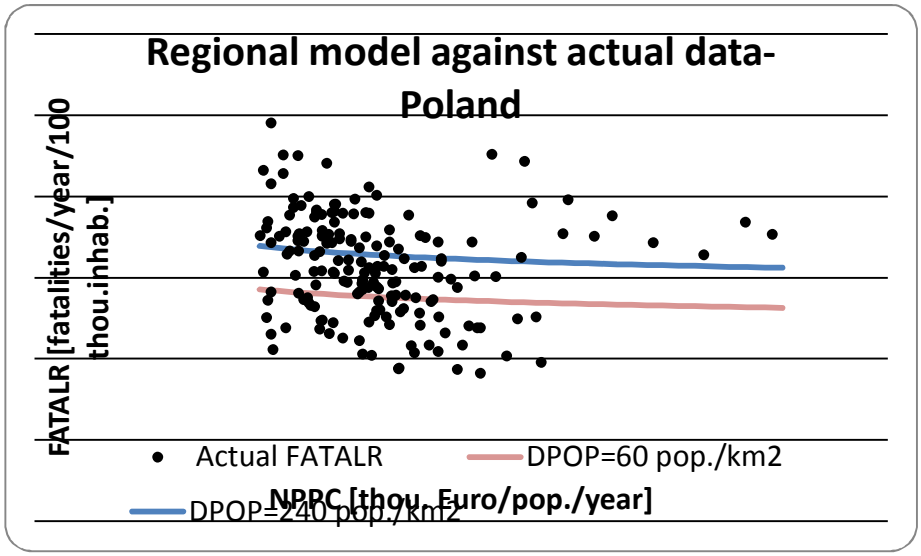



**Figure 5.** Graph of prepared regional FATALR model in relation to population density DPOP against the actual data for Poland - remaining variables from the model assumed as average.

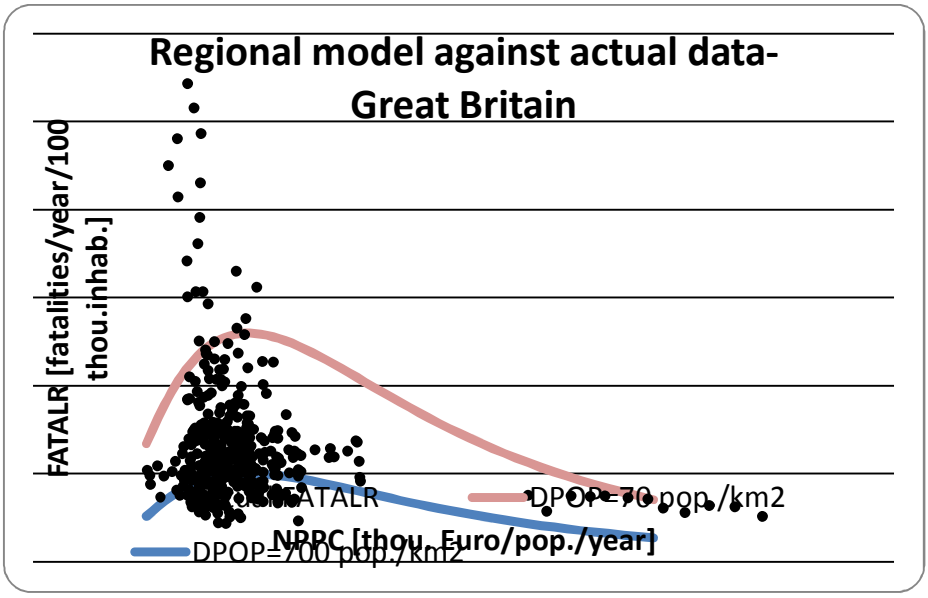



**Figure 6.** Graph of prepared regional FATALR model in relation to population density DPOP against the actual data for Great Britain - remaining variables from the model assumed as average.

## 5 MODEL FOR TREND OF NATIONAL PRODUCT PER CAPITA VARIATION IN TIME

Global national and local regional models are the ones, in which the majority of independent variables are characterized by slight variability in time. The only variable dynamically changing in time, and at the same time occurring in almost all models, is national product per capita. Analysis of variations of average national product per capita NPPC over time proved that analysed countries are characterized by two types of NPPC change trends in time.

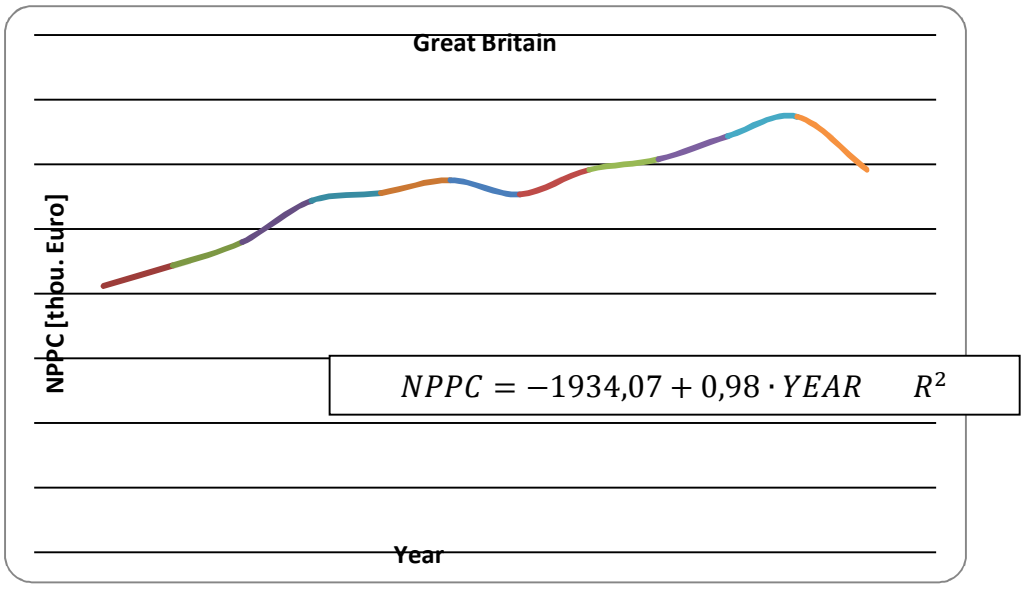



**Figure 7.** Graph of dependence of NPPC changes in Great Britain in the analysed years

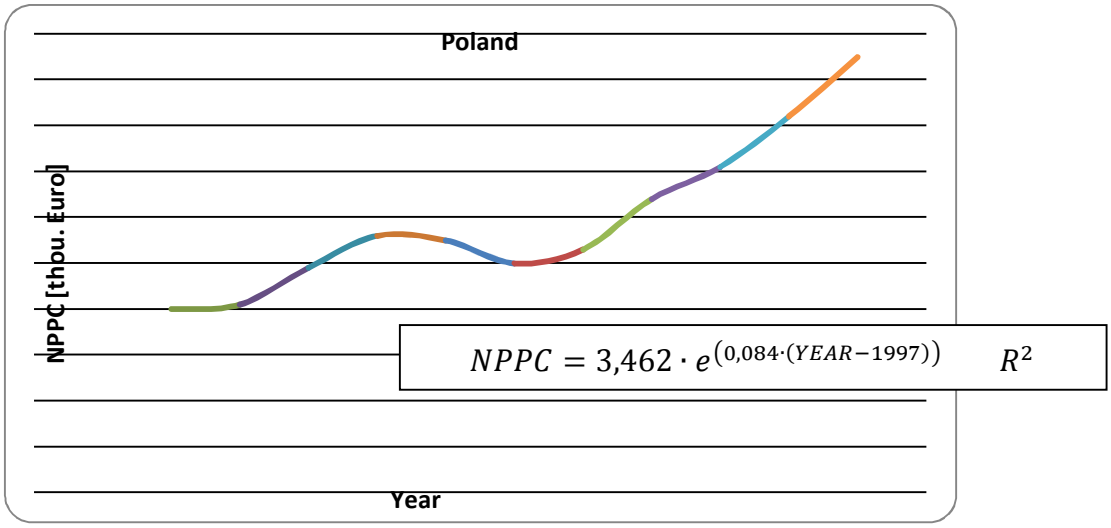



**Figure 8.** Graph of dependence of NPPC changes in Poland in the analysed years

First one is a linear, which characterizes mainly the countries of the "Old Europe", where the economic situation is stabilized and standard of living is improving. The second one, however, is nonlinear trend, probably resulting from dynamic changes occurring in these countries after their access to the European Union, Fig. 7 and Fig. 8.

## 6 Aggregated model

After preparation of partial models, the model comprising all partial models has been created and it has the following formula:

$$ATALR = 5{,}353 \cdot MODEL_{NATIONAL}^{-0{,}285} \cdot MODEL_{REGION}^{0{,}818} \cdot JPKB_{NATIONAL}^{-0{,}212} \quad (4)$$

It is a multiplicative model, elements of which have Q factors ranging between 0.89 and 0.54, while resultant model has Q factor equal to 0.76. This is a result of using average annual data as input data, in order to eliminate momentary fluctuations that could obscure the character of effects of respective influences. Received results may be considered satisfactory and visualisation of model adjustment to the actual data has been presented in Fig. 9 and 10.

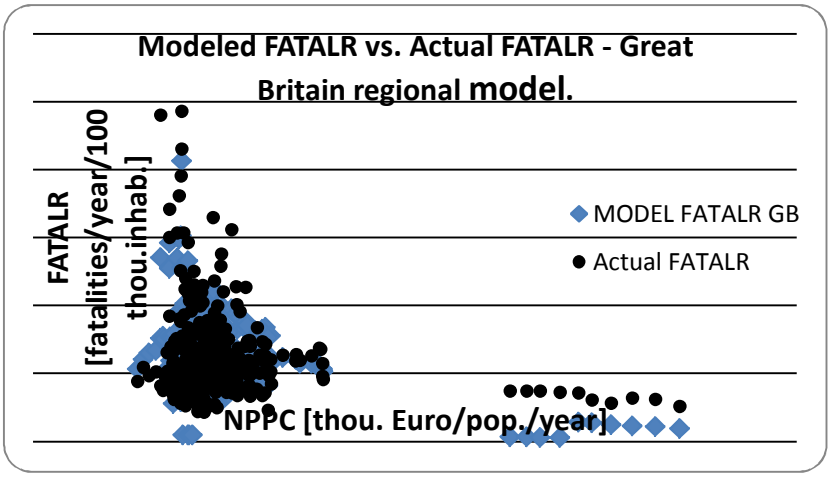



**Figure 9.** Graph of the actual and modelled data in relation to NPPC in a given region for regions in Great Britain.

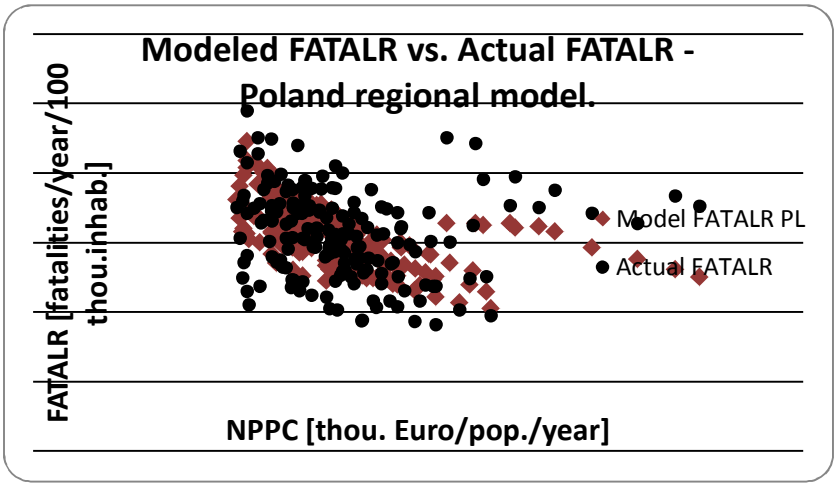



**Figure 10.** Graph of the actual and modelled data in relation to NPPC in a given region for regions of Poland.

## 7 ANALYSIS OF THE IMPACT OF RESPECTIVE VARIABLES ON FATALR MODEL

Analysis of the model's sensitivity is based on analysing changes in FATALR occurring as a result of changes in the values of variables being the arguments of model function, based on the observation of this function's derivatives against respective variables.
Furthermore, significant information on the way of potential impact on FATALR are obtained from the information on the values of sensitivity coefficients and the direction as well as size of variations of these coefficients caused by the changes of function's arguments. Sensitivity analysis is one of the basic instruments of risk assessment in decision-making process. Sensitivity analysis allows defining accuracy level, which is necessary at estimation of respective parameters for the model to be precise enough.

Since hierarchic model of FATALR is a multiplicative model, in which respective factors are raised to different powers, then the impact of each factor, estimated by elasticity function, is constant and equal to exponent of the power [7]. Hence performing sensitivity analysis for $\text{Model}_{\text{REGION}}$ is particularly important, especially that it takes into account local specificity as concerns variables existing in the model. Depending on the number of changed factors, the impact of which is taken into consideration, univariate and multivariate analysis is distinguished. In the case of univariate sensitivity analysis, the reaction of the model to a change of one of the factors is studied, while assuming constant level of the other ones.

Most frequently used sensitivity measures of gain include measures based on the concept of point elasticity of function towards respective variables. This approach yields correct results, when we consider slight change of a given factor.
Similarly to referenced article [7], by calculation of elasticity of regional model for respective variables we receive:

$$E_M(CARP) = \gamma_3 \cdot CARP, E_M(ROADC) = \gamma_4 \cdot ROADC, E_M(UNEMP) = \gamma_5 \cdot UNEMP$$
$$E_M(DPOP) = \gamma_1, E_M(VEHD) = \gamma_2$$
$$E_M(NPPC) = \frac{\beta + lnNPPC \cdot \gamma_6}{lnNPPC} = \frac{\beta}{lnNPPC} + \gamma_6.$$

Set of independent variables may then be divided into three groups of variables:

- fixed level of elasticity, DPOP and VEHD variables;
- linearly dependent elasticity level, CARP, ROADC, UNEMP variables;
- inversely proportional elasticity level, NPPC variable.

Variables of the first two groups are characterized by relatively small dynamics of change, as in a period of about 10 years it is difficult to observe significant changes in demography of a given region, and similarly in the total vehicle fleet or total length of roads. It is different in case of unemployment index, although this variable is present only in one of the models prepared and turned out to be useful in the case of British regions. The most interesting proved the impact of national product per capita in respective regions. Its changes turned out to have the biggest impact on the modelled FATALR values in the regions.
When building the model special attention must be paid to separable homogenous groups of elements, especially in respect of their variability.
The purpose of sensitivity analysis of FATALR models is checking how the value of the model changes as a result of combined changes of factors that impact the model. In multivariate sensitivity analysis the impact of combined changes of several factors is taken into account. Therefore, it is necessary to apply differential for the assessment of the total change of the model's value. By calculating partial derivatives for respective variables, we receive

$$\frac{\partial Model_{Region}}{\partial CARP} = \gamma_3 \cdot \alpha \cdot (lnNPPC)^{\beta} \cdot NPPC^{\gamma_6} \cdot DPOP^{\gamma_1} \cdot VEHD^{\gamma_2} e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)}$$
$$\frac{\partial Model_{Region}}{\partial ROADC} = \gamma_4 \cdot \alpha \cdot (lnNPPC)^{\beta} \cdot NPPC^{\gamma_6} \cdot DPOP^{\gamma_1} \cdot VEHD^{\gamma_2} e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)}$$
$$\frac{\partial Model_{Region}}{\partial UNEMP} = \gamma_5 \cdot \alpha \cdot (lnNPPC)^{\beta} \cdot NPPC^{\gamma_6} \cdot DPOP^{\gamma_1} \cdot VEHD^{\gamma_2} e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)}$$
$$\frac{\partial Model_{Region}}{\partial DPOP} = \gamma_1 \cdot \alpha \cdot (lnNPPC)^{\beta} \cdot NPPC^{\gamma_6} \cdot DPOP^{\gamma_1 - 1} \cdot VEHD^{\gamma_2} e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)}$$
$$\frac{\partial Model_{Region}}{\partial VEHD} = \gamma_2 \cdot \alpha \cdot (lnNPPC)^{\beta} \cdot NPPC^{\gamma_6} \cdot DPOP^{\gamma_1} \cdot VEHD^{\gamma_2 - 1} e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)}$$

$$\frac{\partial Model_{Region}}{\partial NPPC} = [\beta + lnNPPC \cdot \gamma_6](lnNPPC)^{\beta - 1} \cdot NPPC^{\gamma_6 - 1} \cdot \alpha \cdot DPOP^{\gamma_1} \cdot VEHD^{\gamma_2} e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)}$$

By calculating differential we receive a formula, which describes joined impact of independent variables on the value of $Modelu_{Region}$:

$$dModel_{Region} = \alpha\{[\beta + lnNPPC \cdot \gamma_6]DPOP \cdot VEHDdNPPC + lnNPPC \cdot NPPC \cdot [DPOPVEHD(\gamma_3 dCARP + \gamma_4 dROADC + \gamma_5 dUNEMP) + VEHD\gamma_1 dDPOP + DPOP\gamma_2 dVEHD]\} \cdot (lnNPPC)^{\beta - 1} \cdot NPPC^{\gamma_6 - 1} \cdot DPOP^{\gamma_1 - 1} \cdot VEHD^{\gamma_2 - 1} e^{(\gamma_3 \cdot CARP + \gamma_4 \cdot ROADC + \gamma_5 \cdot UNEMP)}$$

## 8 SUMMARY

Presented heuristic and statistical analysis points to advisability of using local characteristics for the assessment of safety in road transport. Due to differences in availability of registered data, classes of "close", in terms of cluster analysis, variables and their representatives for specified groups of effects (impacts) have been distinguished.

λ parameter in national model is inversely proportional to corruption index and directly proportional to percentage of old passenger cars. These mutually competing effects (impacts) may be used to specify a strategy of actions aimed at reducing road accident fatalities index.

Models describing the impact of regional characteristics in respective countries on the FATALR values in such regions, and analysis of sensitivity, point to NPPC variable as dynamic and controlled element that drives FATALR changes.

In multivariate analysis of sensitivity the impact of concurrent change of several independent factors is analysed. That is why it is necessary to check correlation of variables used in aggregated model as well as in partial models. It will allow development of more effective mechanisms of influencing FATALR index.

**Acknowledgement**

The paper has been funded by grant-in-aid for carrying out scientific research or development work and related tasks, aimed at the progress of young scientists and doctoral students.

## 9 REFERENCES

[1] "Global status report on road safety:time for action," Geneva, 2009.

[2] J. J. F. Commandeur, F. D. Bijleveld, R. Bergel-Hayat, C. Antoniou, G. Yannis, and E. Papadimitriou, "On statistical inference in time series analysis of the evolution of road safety.," *Accident; analysis and prevention*, pp. 1–11, Dec. 2012.

[3] A. E. Van Beeck, E., Mackenbach, J. P., Looman, C. W., & Kunst, "Determinants of Traffic Accident Mortality in the Netherlands: A Geographical Analysis.," *International Journal of Epidemiology*, vol. 20, no. 3, pp. 698–706, 1991.

[4] K. Jamroz, "The impact of road network structure and mobility on the national traffic fatality rate," no. Jamroz 2011, pp. 1–7.

[5] L. Fridstrøm, J. Ifver, S. Ingebrigtsen, R. Kulmala, and L. K. Thomsen, "Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts.," *Accident; analysis and prevention*, vol. 27, no. 1, pp. 1–20, Feb. 1995.

[6] Firth, D. (1991) Generalized linear models. In: Hinkley V, Reid N, Snell EJ, editors. *Statistical theory and modelling: in honour of Sir David Cox, FRS*. London: Chapman and Hall; p. 66-93.

[7] Wachnicka J.,Smolarek L, " The Multivariate Multilevel Analysis Of Different Regional Factors Impact on Road Safety in European Country Regions" Journal of KONBIN, No 4 (24) 2012, pp. 141-148, ISSN 1895-8281

# COMPARISON METHODS OF MODELING CONTINUOUS RANDOM VARIABLES ON EMPIRICAL DISTRIBUTIONS

**Farhadzadeh E.M., Farzaliyev Y.Z., Muradaliyev A.Z.**

•

Azerbaijan Scientific-Research and Design-Prospecting
Institute of Energetic AZ1012, Ave. H.Zardabi-94,
e-mail:fem1939@rambler.ru

## ABSTRACT

The new method of modeling of continuous random variables on empirical distributions offered. It shown, that discrepancy of accuracy of methods to shown requirements is shown at small number of realizations of random variables, reduced to not casual divergence of estimations of averages and average quadratic values empirical given and modeled samples.

## STATEMENT OF THE PROBLEM.

One of the basic stages of imitating modeling is formation of random variables and casual events with the set law of distribution. In conditions of electro power systems (EPS) examples of random variables are: duration of emergency repair of the equipment and devices, intervals of time between non-working conditions of power units, the maintenance of soluble gases in transformer oil, etc. Casual events: short circuits on transmission lines, refusal in switching-off of the switch, false work of relay protection or automatics, etc. the Analytical form of laws of distribution here is in most cases unknown. Laws of change of a continuous random variable set by statistical (empirical) function of distribution (s.f.d.), and a discrete random variable – proceeding from those or other assumptions of probability of occurrence of casual event. This feature brings the certain interrelation between number of intervals s.f.d. F*(X), and number of intervals m at discrete representation of continuous empirical function of distribution F(X). If for F(X) the number of intervals m gets out equal (10÷20), for F*(X) m=n.

Objectivity of imitating modeling in many respects depends on that, how much realizations of modeled random variables (events) will appear casual and will reflect the set laws of distribution. It is necessary to note also, that in practice often aspire to present set of statistical data one of known laws of distribution. Actually, the law of distribution of the statistical data concerning a class multivariate represents an uncertain composition of many distributions. In other words, difficulties of representation observable s.f.d. objective analytical law in many respects increase.

**Methods of statistical modeling.** By development of these methods, the greatest attention given a condition when the type of function of distribution of continuous random variable X known. Statistical modeling on empirical distribution is carried out by two methods. According [1] s.f.d. represented the following equations:

$$F_1^*(X)=\begin{cases}0 & \text{если } X=X_0=0\\ \dfrac{1}{n}+\dfrac{(X-X_i)}{n(X_{i+1}-X_i)} & \text{если } X_0<X<X_n\\ 1 & \text{если } X\geq X_т\end{cases} \quad (1)$$

where i=0,(n-1)

If to designate realization of a random variable with uniform distribution in an interval [0,1] through ξ, that according to (1) calculation corresponding ξ realizations of random variable X it is carried out under the formula:

$$X=X_i+(X_{i+1}-X_i)\cdot(\xi\cdot n-i) \quad (2)$$

where i=0,(n-1)

Intuitively clearly, that if the divergence $(X_n\text{-}X_1)$ is commensurable with $X_1$, modeling s.f.d. $F_1^{**}(X)$ under the formula (2) leads to regular distinction $F_1^*(X)$ and $F_1^{**}(X)$. This distinction shown in following parities of averages (accordingly $M_1^*(X)$ and $M_1^{**}(X)$) and average quadratic (accordingly $G_1^*(X)$ and $G_1^{**}(X)$) values of random variable X:

$$\begin{aligned}&M_1^*(X)>M_1^{**}(X)\\ &G_1^*(X)<G_1^{**}(X)\end{aligned} \quad (3)$$

Graphic illustration of this method is resulted on fig.1a.

In the second method [2] s.f.d. represented the following equation:

$$F_2^*(X)=\begin{cases}0 & \text{если } X<X_1\\ \dfrac{i-1}{n-1}+\dfrac{(X-X_i)}{(n-1)\cdot(X_{i+1}-X_i)} & \text{если } X_1\leq X<X_n\\ 1 & \text{если } X\geq X_n\end{cases} \quad (4)$$

Calculation of realization of random variable X spent under the formula:

$$X=X_i+(X_{i+1}-X_i)\cdot[\xi\cdot(n-1)-(i-1)] \quad (5)$$

where i=1,(n-1)

In [2] it is marked, that obvious lack of this method is modeling random variable X in interval $X_1<X<X_n$, in other words, size X never can be less $X_1$ and more $X_n$, that the brings the certain error of an estimation $M_2^{**}(X)$. Graphic illustration $F_2^*(X)$ and components of the formula (5) is resulted on fig.1в.

Features of calculation under formulas (2) and (5) have caused expediency of specification of these methods of modeling. S.f.d. recommends presenting the following the equation [4]:

$$F_3^*(X)=\begin{cases}0 & \text{если } X\leq X_1\\ \dfrac{i-1}{n+1}+\dfrac{(X-X_i)}{(n+1)\cdot(X_{i+1}-X_i)} & \text{если } X_1<X<X_{n+1}\\ 1 & \text{если } X\geq X_{n+2}\end{cases} \quad (6)$$

where i=1,(n+1)

Thus, calculation of realization of random variable X carried out under the formula:

$$X=X_i+(X_{i+1}-X_i)\cdot[\xi\cdot(n+1)-(i-1)] \quad (7)$$

where i=1,(n+1)

The graphic illustration of components $F_3^*(X)$ is resulted on fig.1c

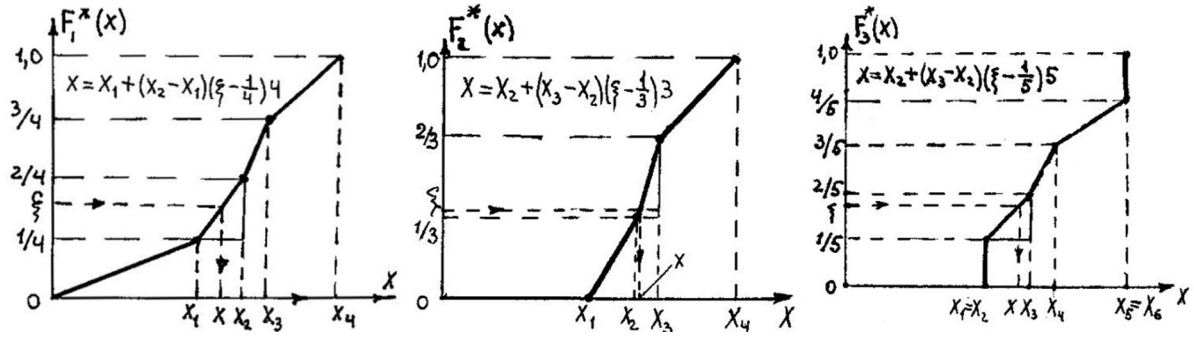



Fig.1. Illustration methods of modeling continuous random variables on empirical distribution. a - method [1]; b – method [2]; c – method of authors

**Algorithm of comparison of methods of statistical modeling.** The basic requirement shown to methods of statistical modeling, accuracy of conformity of distribution $F_j^{**}(X)$ to initial distribution F*(X), where j=1,3. Most simple way of the control of a degree of such conformity at small values n is comparison of estimations of average values $M_E^*(X)$ and $M_j^{**}(X)$, and also average quadratic values $G_E^*(X)$ and $G_j^{**}(X)$.

The block scheme of modeling algorithm is resulted on fug.2.

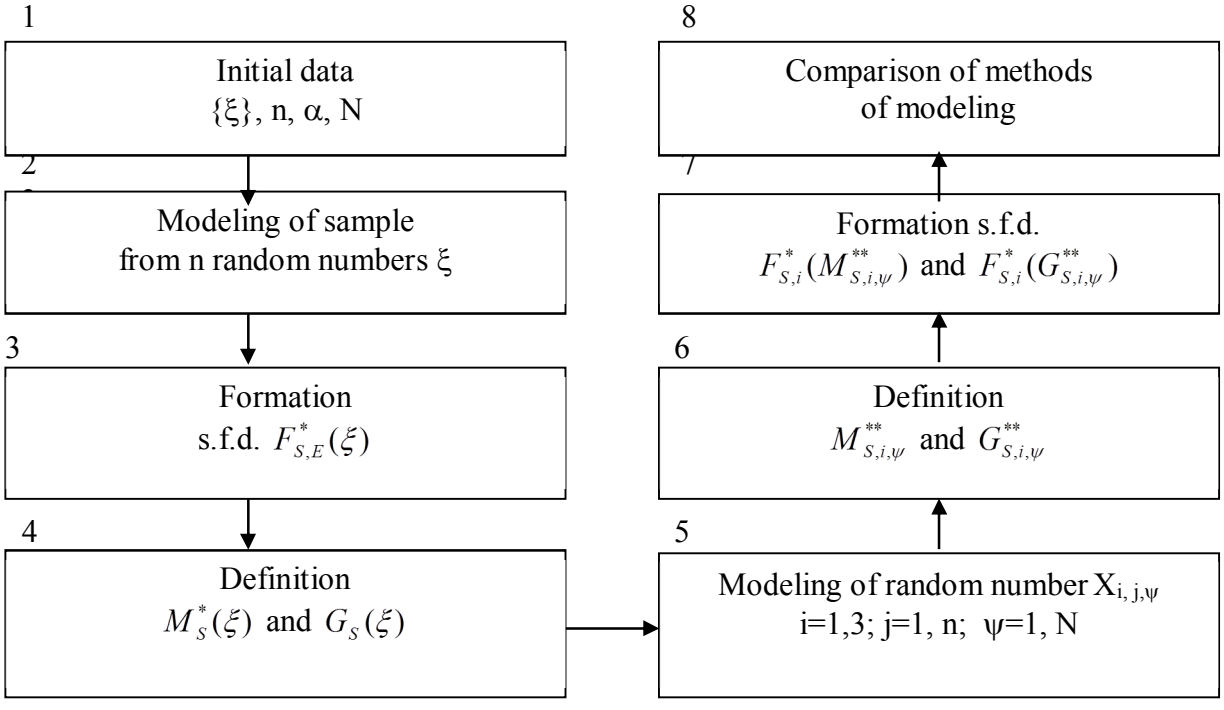



Fig 2. The integrated block diagram of algorithm of comparison of methods of modeling of continuous random numbers

Let's consider features of this algorithm by way of numbering blocks of its block diagram (fig.2.)

1. Initial data are:

– set of pseudo-random numbers {ξ} with uniform distribution in an interval [0,1];

- n – number of random numbers ξ in sample from {ξ}. Change n allows to establish its influence on result of comparison of methods of modeling of random variables X;
- α - significance value. Allows to estimate influence of a degree of conformity s.f.d. F*(ξ) to the uniform law on result of comparison of methods of modeling of random variables X;
- N – number of imitations of modeled sample $\{X\}_n$

2. Under program RAND, (ξ) is formed n pseudo-random numbers ξ, corresponding the uniform law of distribution in an interval [0,1];
3. Considering, that X=ξ, average M*(X) and an average quadratic G*(X) values on sample is calculated ξ.
4. Under formulas (1), (4) both of (6) and sample $\{\xi\}_n$ are formed s.f.d. $F_1^*(X)$, $F_2^*(X)$ and $F_3^*(X)$;
5. Sample from n pseudo-random numbers with an opportunity of the control of conformity of distribution F*(ξ) is formed to the uniform law with the set significance value α. The method of the general random numbers forms three samples from n random numbers X on distributions $F_1^*(X)$, $F_2^*(X)$ and $F_3^*(X)$. Calculations are spent for significance values (a errors I type) α, and number of realizations of sample N=1000;
6. Estimations of an average $M_{i,\psi}^*(X)$ and average quadratic $G_{i,\psi}^*(X)$ values of modeled random variables on i- th to a method are calculated for ψ- th samples with i=1,3 and ψ=1,N;
7. Are formed s.f.d. $F_i^*\left[M^*(X)\right]$ and $F_i^*\left[G^*(X)\right]$ for each of three methods i=1,3;
8. Comparison of methods is carried out by comparison M*(X) and G*(X) with similar parameters of distributions $F_i^*\left[M^*(X)\right]$ and $F_i^*\left[G^*(X)\right]$, i.e. with $M_i^{**}(X) = M_i^*[M_{i,\Psi}^*(X)]$ and $G_i^{**}(X) = M_i^*[G_{i,\Psi}^*(X)]$ i=1,3. Advantage is given a method for which the deviation from $M_E^*(X)$ and $G_E^*(X)$ is minimal

## RESULTS OF CALCULATIONS

It is established:

1. Influence of a method of modeling on accuracy of reproduction of distribution F*(ξ) it is shown only for small n. Already at n≥20 divergence between M*(X) and $M_i^{**}(X)$, as well as G*(X) and $G_i^{**}(X)$ with i=1,3 does not exceed 1%. Notice, that at n=4 the divergence between M*(X) and $M_i^{**}(X)$ makes 12%, and between G*(X) and $G_i^{**}(X)$ makes 28.5%;
2. The size of a divergence including the greatest, between F*($X_j$) and $F_i^{**}(X_j)$ (designate this size as $St_j$) does not depend on law of change F*($X_j$) and $F_i^{**}(X_j)$, and depends on random variables of sample $\{\xi\}_n$, their numbers n and a way of modeling i=1,3. As an example on fig.3 the graphic illustration of independence $St_j$ with j=1,n from type F*(X) is resulted. In it finds reflection known nonparametric character of criterion of the greatest divergence [3]
3. Comparison of methods of modeling shows, that

$$M^*(X) = M_2^{**}(X) = M_3^{**}(X) >> M_1^{**}(X)$$

$$M^*[G_2^*(X)] << G^*(X) = M^*[G_3^*(X)] << M^*[G_1^*(X)]$$

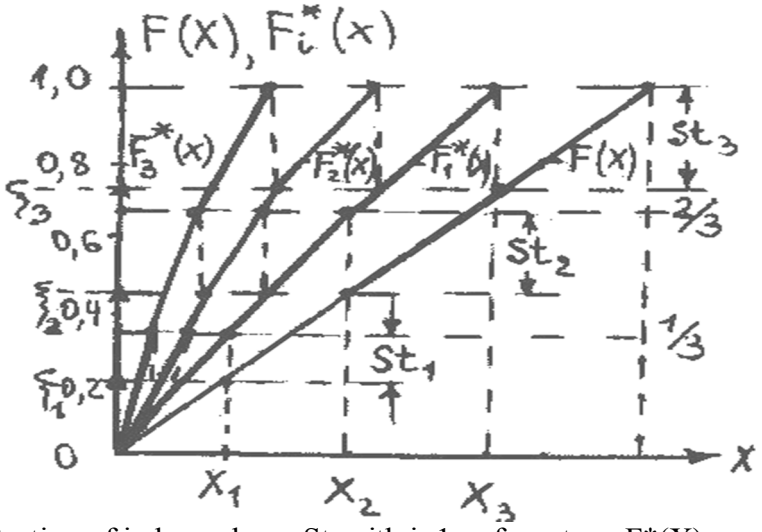
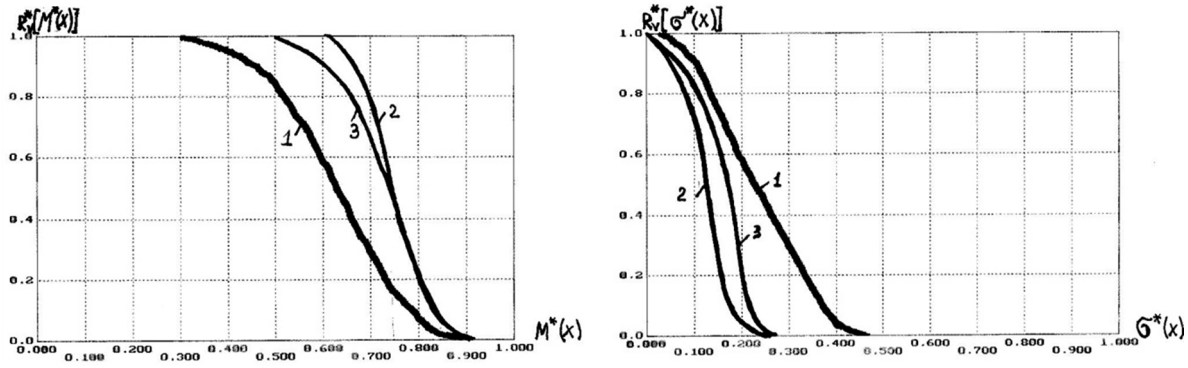



Fig. 3. Graphic illustration of independence $St_j$ with j=1, n from type F*(X)

In other words, the first method does not meet shown requirements to accuracy of calculation both on size $M_i^*(X)$, and on value $G_i^*(X)$. Average values $M_2^{**}(X)$ and $M_3^{**}(X)$, calculated at modeling samples random variables, accordingly, the second (i=2) and the third (i=3) methods, are practically indiscernible and equal to M*(X). However, average quadratic values modeled samples for the second method of modeling $\{X\}_n$ essentially differ from a reference value $G^*(X)$ while the size $M^*[G_3^*(X)]$ practically does not differ from $G^*(X)$.

Graphic illustration of distinction s.f.d. $R^*[M_i^*(X)] = 1 - F^*\left[M_i^*(X)\right]$ and $R_i^*[(G_i^*(X)] = 1 - F^*\left[G_i^*(X)\right]$ for various methods (i=1÷3) and α=0 it is resulted on fig.4.




a) b)

Fig.4. Illustration of distinction s.f.d. averages (a) and averages quadratic (b) values of realizations выборок modeled i=1,2 и3 methods

4. With increase α:

- Average value $M_i^*[M_i^*(X)]$ with i=1,3 i.e. for each method of modeling aspires to the true value and allows to compare with methods more full. At α=0.8 following values are received: $M_1^*[M_1^*(X)] = 0.629$, $M_2^*[M_2^*(X)] = 0.751$ and $M_3^*[M_3^*(X)] = 0.738$ at $M^*(X) = 0.739$;
- The disorder of realizations $M_i^*(X)$ with i=1,3 decreases. If for α=0 for $M_1^*(X)$ it made $G_1^*[M_1^*(X)] = 0.143$, at α=0.8 size $G_1^*[M_1^*(X)] = 0.066$, i.e. disorder of realization $M_1^*(X)$ decreases in 2,2 times. The same reduction of disorder observed for the second and third methods;
- Distinction value of realizations $G_1^*(X)$, on the average, practically invariable also does not exceed 10% for n=4 and 3% for n=16. In the illustrative purposes on fig. 5 distributions $F^*[M_i^*(X)]$

and $F^*[G_i^*(X)]$ are resulted at $\alpha$=0.8, parities of considered methods confirming independence from $\alpha$

- Consequences from increase $\alpha$ are similar to consequences of artificial increase in number of modeled random variables n on size $(1-\alpha)^{-1}$.

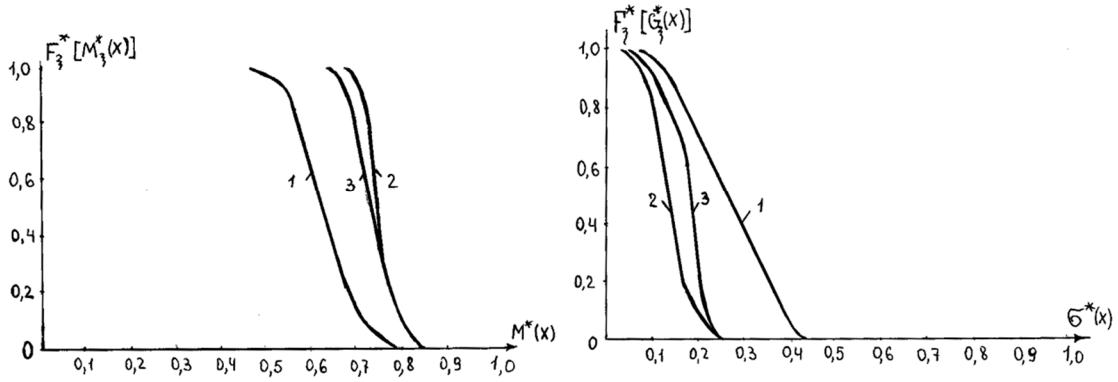



Fig.5. Graphic illustration s.f.d. $F^*[M_i^*(X)]$ and $F^*[G_i^*(X)]$ at $\alpha$=0.8

## CONCLUSION

The lead complex analysis has allowed establishing:

1. Discrepancy of accuracy of methods of modeling of continuous random variables on empirical distributions to shown requirements is shown only at small number of realizations of sample of random variables (n<20)
2. Comparison of methods of modeling can be lead by comparison of modeled estimations of averages and average quadratic values of random variables to empirical values of estimations of these parameters
3. Modeling of continuous random variables on the empirical distribution calculated under the formula (1), at small n leads to essential distinction of averages and average quadratic values of random variables of sample from empirical values, and under the formula (4) – average quadratic values of sample
4. Increase in a significance value $\alpha$ conformity of sample from n pseudo-random numbers to the uniform law on the consequences to similarly artificial increase in number n on size $(1-\alpha)^{-1}$
5. Statistical modeling of random variables on empirical distributions is expedient for spending under the formula (7).

## REFERENCE

1. Pole Y.G. Likelihood modeling on electronic computers.–M.: Sov. radio, 1971, 400c.
2. Averill Law, V. Devid's, Kelton M. Imitating modeling. Classics CS,. – SPb.: Peter; Kiev: Publishing group BHV, 2004 847 p.
3. Gnedenko B.V., Beljaev J.K.,Solovyov A.D. Mathematical nightingales methods in the theory of reliability. "Science", 1965, 524 p.
4. Farhadzadeh E.M., Muradaliev A.Z., Rafiyeva N.K., Nazirova U.K. Method of statistical modeling of random variables on empirical distributions. Kazan: News of High schools. Problems of Power №9-10, 2008, 112-120 p

# DECREASE IN RISK ERRONEOUS CLASSIFICATION THE MULTIVARIATE STATISTICAL DATA DESCRIBING THE TECHNICAL CONDITION OF THE EQUIPMENT OF POWER SUPPLY SYSTEMS

**Farhadzadeh E.M., Farzaliyev Y.Z., Muradaliyev A.Z.**

•

Azerbaijan Scientific-Research and Design-Prospecting
Institute of Energetic AZ1012, Ave. H.Zardabi-94,
e-mail:fem1939@rambler.ru

## ABSTRACT

Objective estimation of parameters of individual reliability is an indispensable condition of an opportunity of decrease in operational expenses for maintenance service and repair of the equipment and devices of electro power systems. The method of decrease in risk of erroneous classification of multivariate statistical data offered. The method based on imitating modeling and the theory of check of statistical hypotheses.

## I. INSTRUCTION

Estimation parameters of individual reliability of the equipment of power supply systems provides classification of final population of multivariate statistical data of operation, tests and restoration of deterioration on the set versions of attributes (VA) [1].

VA reflects features of a design, a condition of operation, feature of occurrence of refusals and carrying out of repairs of the equipment. Expediency of classification on each of population VA is established by comparison of statistical functions of distribution (s.f.d.) final population of statistical data $F^*_{\Sigma}(X)$ and s.f.d. samples n random variables from this population on i versions of V attribute $F^*_{V,i}(X)$, where v=1, k; k-number of attributes of random variable X (for example, durations of emergency repair); i=1, rk; $r_k$- number of versions k an attribute. If s.f.d. $F^*_{\Sigma}(X)$ and $F^*_{V}(X)$ differ not casually, in other words, sample $\{X\}_n$ where n-number of random variables of sample, it is not representative classification of data at an estimation of parameters of individual reliability is expedient and on the contrary. It is necessary to note, that unlike sample of a general data population (analogue: infinite set of random variables with uniform distribution in an interval [0,1]), which imposing appearance is set by some significance value $\alpha$, sample of final population of multivariate data on set VA is not casual, as a matter of fact, and it can appear only representative. In particular, sample can appear representative, if for considered data set VA not significant.

## II. RECOMMEND METHOD

In a basis of comparison $F^*_{\Sigma}(X)$ and $F^*_{V}(X)$ there is a statistical modeling (by means of computer program RAND) n pseudo-random numbers $\xi$, random variables of sample equal to number, with uniform distribution in an interval [0,1].

Indispensable condition thus is consistency s.f.d. $F^*_{V}(\xi)$ to the uniform law of distribution $F_{\Sigma}(\xi)$, in other words, casual character of distinction $F_{\Sigma}(\xi)$ and $F^*_{V}(\xi)$. It is obvious, that from the uniform law of change of random numbers $\xi$ at all consistency does not follow the uniform law

s.f.d. $F_V^*(\xi)$ with the set significance value $\alpha$. Use at modeling statistical analogue $F_V^*(X)$ s.f.d. $F_V^*(\xi)$, essentially differing from $F_\Sigma(\xi)$, leads to erroneous increase in value of the greatest divergence of distribution of this analogue $F_V^{**}(X)$ from $F_\Sigma^*(X)$ and by that to growth of probability of the erroneous decision at classification of data.

Representative character of sample $\{\xi\}_n$ at the decision of a problem of an estimation of expediency of classification of multivariate data it was supervised Kolmogorov's by criterion [2]. According to this criterion sample $\{\xi\}_n$ it is unpresentable, if

$$D_n > d_{n,(1-\alpha)} \quad (1)$$

where:

$$D_n = \max(D_n^+, D_n^-) \quad (2)$$

$$D_n^+ = \max\{D_i^+\}; \quad 1 \le i \le n \quad (3)$$

$$D_i^+ = \left(\frac{i}{n} - \xi_i\right) \quad (4)$$

$$D_n^- = \max\{D_i^-\}; \quad 1 \le i \le n \quad (5)$$

$$D_i^- = \left(\xi_i - \frac{i-1}{n}\right) \quad (6)$$

$d_{n,(1-\alpha)}$ – critical value of statistics Dn provided that $F_\Sigma(\xi)$ and $F_V^*(\xi)$ differ casually

In [3] it is marked, that estimation $D_n$ under the formula

$$D_n^/ = \max\left\{\left|D_i^+\right|\right\}, \quad 1 \le i \le n \quad (7)$$

leads to incorrect decisions on a parity $F_\Sigma(\xi)$ and $F_V^*(\xi)$.

The similar remark can be found and in [4]. The reason of such discrepancy does not stipulate. At uncertain in advance n, decrease in time of calculation, according to [3], is reached by application of exact approach Stephens, which tabulated critical values $d_{n,(1-\alpha)}$, depending from n and $\alpha$, reduces to dependence only from $\alpha$. Sample $\{\xi\}_n$ it is unpresentable, if

$$A \cdot D_n > C_{1-\alpha} \quad (8)$$

where:

$$A = \left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) \quad (9)$$

For example, at n=4 size A=2,175 and for $\alpha$=0,1 critical value $C_{1-\alpha}$=1,224, and at $\alpha$=0,05 size $C_{1-\alpha}$=1,358.

Application of a method of the decision of «a return problem» when it is in advance known, that sample $\{\xi\}_n$ it is unpresentable, has shown, that criteria (1) and (8) for values most often used in practice $\alpha$=0,05 and $\alpha$=0,1 not casual character of divergence $F_\Sigma(\xi)$ and $F_V^*(\xi)$ at small n establish only for those cases when it does not raise the doubts. For acknowledgement of this statement, we shall consider a following example. Let random numbers $\psi$ have uniform distribution $F_\Sigma(\psi)$ in an interval [0.5; 1]. Casual sample is set $\{\psi\}_n$ with n=4: {0,86346; 0,50672; 0,91424 and 0,67210}. Check up the assumption of imposing appearance of this sample for the uniform law of distribution of a random variable $\xi$ in an interval [0,1].

Results of calculations are resulted in table 1.

Table 1

Example of an estimation of imposing appearance of sample

| i | $F_\Sigma(\psi_i)$ | i/n | $D_i^+$ | $D_i^-$ | The note |
|---|---|---|---|---|---|
| 1 | 0.507 | 0.25 | -0.257 | +0.506 | $D_i^+ = 0.086$; $D_i^- = 0.506$ |
| 2 | 0.672 | 0.5 | -0.172 | +0.422 | $D_n$=0.506; $D_n < d_{4;\,0.9}$=0.565 |
| 3 | 0.863 | 0.75 | -0.113 | +0.363 | $AD_n$=1.101; |
| 4 | 0.914 | 1.00 | +0.086 | +0.164 | $AD_n < C_{0.9}$=1.224 |

As sample follows from table 1 $\{\psi\}_4$ does not contradict the assumption of imposing appearance rather $F_\Sigma(\xi)$ at $\alpha$=0,1.

These features and some assumptions of the reasons of their occurrence [5] have demanded to pass from the analysis of absolute values of the greatest divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\xi)$, to the analysis of the valid values of the greatest divergence ($St_n$). Thus under «the greatest divergence $F_\Sigma(\xi)$ and $F_V^*(\xi)$» we shall understand the greatest on the module vertical distance between $F_\Sigma(\xi)$ and $F_V^*(\xi)$ with i=1, n.

Calculations $St_n$ were spent according to the algorithm, integrated which block diagram is resulted in figure 1.

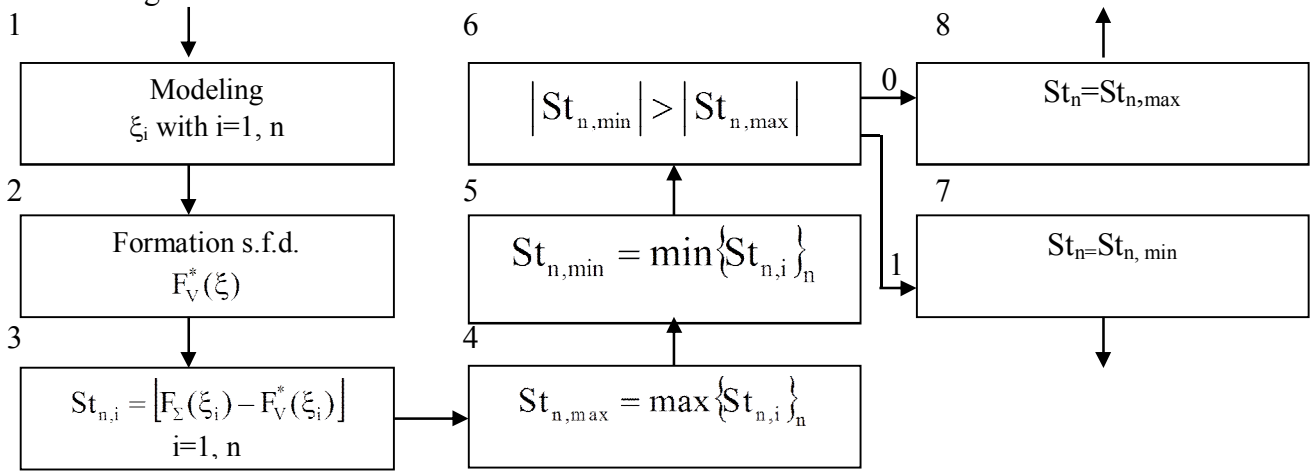



Fig.1. Block diagram of algorithm of calculation of the greatest divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\xi)$

Application of formulas of type

$$St_n = \max\left(\xi_i - \frac{i}{n}\right) \quad 1 \le i \le n \qquad (10)$$

calculation on the computer leads to erroneous results. For example, according to table 1 the maximal value among four realizations of size $D_i^+$ will, $D_i^+ = 0.086$, and the greatest vertical divergence between $F_\Sigma(\xi)$ and $F_V^*(\xi)$ it is equal $D_1^+ = -0.256$

Results of ordering of given realizations $St_n$ presented in table 2 and allow concluding:

1. Quintile distributions $F^*(St_n)=\alpha$ and $n\ge2$ are equal on size and are opposite on a sign (distinction in a sign is caused by distinction of formulas 4 and 10) quintiles distributions $F(D_n)=2\alpha$ {see tabl.16 [2]};
2. Distribution $F^*(St_n)$ is asymmetrical. In the illustrative purposes on fig. 2 are resulted s.f.d. $F^*(St_n)$ for of some n. The assumption of symmetry of distribution $F(St_n)$ it is possible to explain discrepancy of probability practically equal quintile distributions $F^*(St_n)$ and $F(D_n)$;
3. Than $\xi_n$ it is less, that negative value on sign $St_n$ on size will be more, since $St_n=(\xi_{n-1})$. On experimental data the least value $St_n$ for n=2 has appeared equal $St_n$=-0,992, and the greatest $St_n$=+0,489 at sup equal, accordingly, 1 and 0,5.

Table 2

Some results of an estimation s.f.d. $F^*(St_n)$

| $F^*(St_n)$ / n | 0,025 | 0,05 | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 | 0,95 | 0,975 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | -0.842 | -0.775 | -0.684 | -0.551 | -0.473 | -0.149 | -0.363 | -0.304 | -0.239 | -0.060 | 0.184 | 0.285 | 0.343 |
| 3 | -0.709 | -0.635 | -0.566 | -0.471 | -0.400 | -0.335 | -0.296 | -0.252 | -0.200 | -0.145 | 0.231 | 0.299 | 0.372 |
| 4 | -0.623 | -0.567 | -0.494 | -0.414 | -0.355 | -0.302 | -0.253 | -0.217 | -0.173 | 0.155 | 0.240 | 0.319 | 0.377 |
| 5 | -0.567 | -0.511 | -0.449 | -0.370 | -0.318 | -0.274 | -0.232 | -0.190 | -0.147 | 0.164 | 0.246 | 0.309 | 0.360 |
| 6 | -0.523 | -0.469 | -0.411 | -0.338 | -0.292 | -0.252 | -0.215 | -0.173 | -0.127 | 0.171 | 0.244 | 0.303 | 0.358 |
| 7 | -0.481 | -0.438 | -0.384 | -0.318 | -0.274 | -0.235 | -0.201 | -0.162 | -0.113 | 0.165 | 0.235 | 0.290 | 0.342 |
| 11 | -0.389 | -0.353 | -0.309 | -0.255 | -0.219 | -0.189 | -0.110 | -0.129 | -0.097 | 0.160 | 0.216 | 0.260 | 0.302 |
| 16 | -0.33 | -0.295 | -0.258 | -0.215 | -0.184 | -0.158 | -0.134 | -0.103 | 0.107 | 0.150 | 0.194 | 0.232 | 0.264 |
| 22 | -0.280 | -0.253 | -0.221 | -0.183 | -0.157 | -0.135 | -0.113 | -0.083 | 0.105 | 0.137 | 0.176 | 0.210 | 0.235 |
| 29 | -0.246 | -0.219 | -0.193 | -0.160 | -0.138 | -0.119 | -0.099 | -0.068 | 0.098 | 0.126 | 0.158 | 0.186 | 0.212 |
| 40 | -0.208 | -0.187 | -0.164 | -0.136 | -0.119 | -0.102 | -0.084 | -0.050 | 0.089 | 0.112 | 0.140 | 0.164 | 0.185 |
| 60 | -0.173 | -0.156 | -0.137 | -0.114 | -0.097 | -0.083 | -0.069 | 0.054 | 0.077 | 0.096 | 0.118 | 0.138 | 0.155 |
| 90 | -0.142 | -0.127 | -0.111 | -0.092 | -0.079 | -0.068 | -0.055 | 0.051 | 0.067 | 0.081 | 0.100 | 0.116 | 0.130 |
| 120 | -0.122 | -0.110 | -0.096 | -0.080 | -0.068 | -0.059 | -0.047 | 0.047 | 0.060 | 0.072 | 0.089 | 0.102 | 0.114 |
| 150 | -0.110 | -0.099 | -0.086 | -0.071 | -0.062 | -0.053 | -0.042 | 0.041 | 0.053 | 0.065 | 0.079 | 0.092 | 0.104 |

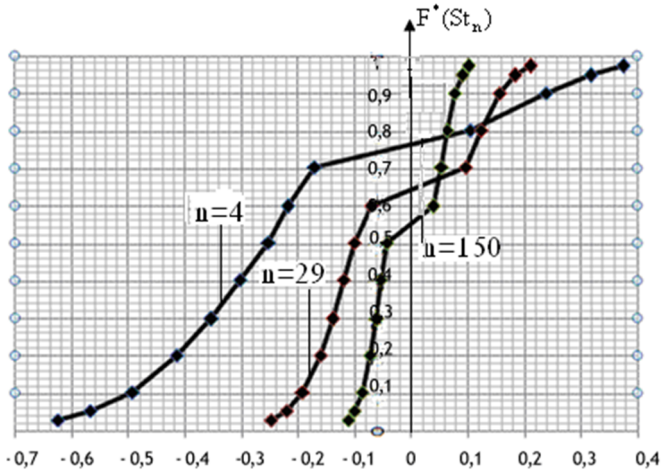



Fig.2. S.f.d. $F^*(St_n)$ for of some n

4. In distribution $F^*(St_n)$ distinguish the bottom $\underline{St_n}$ and top $\overline{St_n}$ boundary values with a significance value $\alpha$, i.e.

$$\left.\begin{aligned} F^*\left(\underline{St_n}\right) &= \alpha/2 \\ F^*\left(\overline{St_n}\right) &= (1-\alpha/2) \end{aligned}\right\} \qquad (11)$$

5. It is established, that if $0{,}25 \geq F^*(St_n) \geq 0{,}75$, i.e. if $\alpha \leq 0{,}5$

$$\overline{St_n} = -\left(\frac{1}{n} + \underline{St_n}\right) \qquad (12)$$

For example, for n=4 and $\alpha$=0.10 according to distribution $F^*(St_n)$ (see tabl.2) size $\underline{St_4} = -0.567$, and $\overline{St_4} = +0.319$. At the same time under the formula (12)

$$-(0{,}25-0{,}567=0{,}317=\overline{St_4}$$

If n=29 and $\alpha$=0,2, that $\underline{St_n} = -0.193$ and $\overline{St_n} = 0.158$. The size $\overline{St_n}$ under the formula (12) is equal - (0,034-0,193)=0,159

On fig. 3 histograms of distribution of negative and positive values $St_n$ for n=4 and n=29 are resulted.

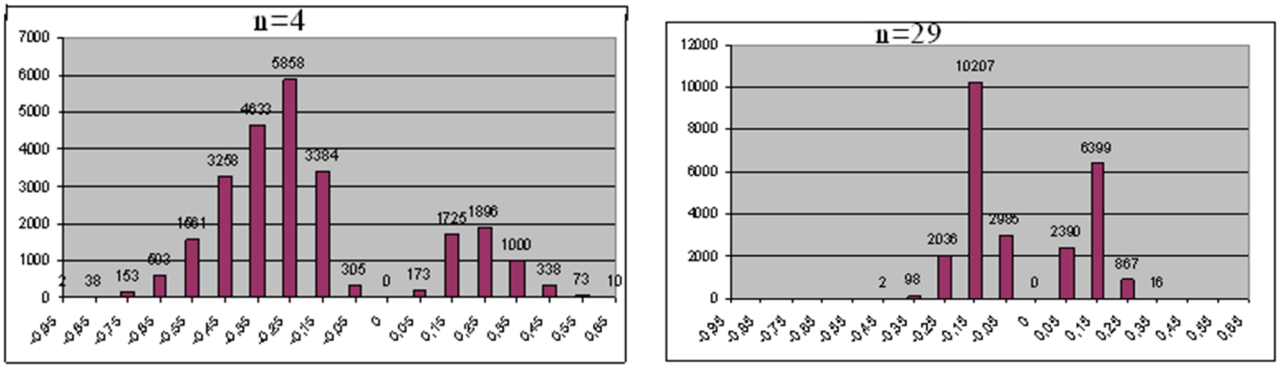



Fig.3. Histograms of distribution of the greatest divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\xi)$

As follows from fig. 3, negative values $St_n$ essentially exceed positive values $St_n$ on relative number and an interval of change. Proceeding from i. 3 it is clear, that it not casually and does not testify about unpresentable samples. With growth n the parity of negative and positive values $St_n$ decreases and aspires to unit. For n=2 negative values $St_n$ make 87,5%, and for n=29 - 61%, and for n=150 – 55%. Thus, even at n=150 quintile distributions $F^*(St_n)$ at $\alpha$=0,05 and $\alpha$==0,95 are not equal [-0.099; +0.092]. Histograms also explain laws of distribution $F^*(St_n)$ resulted on fig.2.

On fig. 4 curve changes of boundary values of statistics $St_n$ for of some values s.f.d. are resulted, $F^*(St_n)$. Criterion of the control of imposing appearance of sample $\{\xi\}_n$ with a significance value $\alpha$ thus looks like:

$$\underline{St_n} < St_n < \overline{St_n} \qquad (13)$$

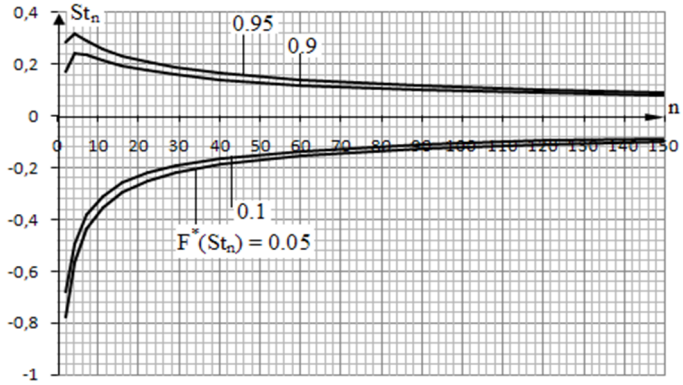



Fig.4. Laws of change of boundary values of the greatest divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\xi)$

Let's designate positive values $St_n$ through $St_n^+$, and negative values- $St_n^-$

In view of i.1. and the equations (12), sample $\{\xi\}_n$ with a significance value $\alpha \le 0{,}5$ can be accepted representative, if

$$\left.\begin{aligned} St_n^+ &< \left[ d_{n,(1-2\alpha)} - \frac{1}{n} \right] \\ \left|St_n^-\right| &< d_{n,(1-2\alpha)} \end{aligned}\right\} \qquad (14)$$

As

$$\left( St_n^+ + \frac{1}{n} \right) = \left|St_n^-\right|$$

criterion (13) for a significance value $\alpha$ can be presented, as

$$\left(\overline{St_n^+} + \frac{1}{n}\right) = \left|\underline{St_n^-}\right| = d_{n,(1-2\alpha)} \qquad (15)$$

Here it is necessary to pay attention to discrepancy of the equations of importance $St_n$ and $d_{n,(1-2\alpha)}$.

If again to address to data of table 1 it is easy to notice, that the interval criterion (13), allowing to consider a sign on the greatest divergence $St_n$, also is unable to establish unpresentable character of sample $\{\psi\}_n$.

It is known, that decrease in risk of the erroneous decision at classification of data can be reached by the account not only errors I type, but also the II types [4].

The most simple decision of this problem would be comparison $St_n$ between $F_\Sigma(\xi)$ and $F_V^*(\xi)$ with boundary values of the interval $\left[\underline{St_n}; \overline{St_n}\right]$ corresponding a significance value $\alpha$=0,5. It is that limiting case of values $\alpha$ when $St_n$=0. Thus a errors II type $\beta$=(1-$\alpha$), i.e. also it is equal 0,5. If $\alpha$ to accept it is less, than 0,5 the errors II type increases $\beta$.

In real conditions:

- configurations $F_\Sigma(\xi)$ also $F_V^*(\xi)$ are various, i.e. $St_n \neq 0$;
- for the same value $St_n$ size ($\alpha+\beta$) less or it is equal to unit;
- in process of increase $St_n$ size ($\alpha+\beta$) decreases, reaches the minimum ($St_{n,opt}$) and then increases;
- if $St_n < St_{n,opt}$, then $\alpha > \beta$, if $St_n > St_{n,opt}$, then $\alpha < \beta$;
- distinction between $\alpha$ and $\beta$ increases in process of increase in a divergence between $St_n$ and $St_{n,opt}$.

Comparison of realizations $St_n$ to boundary values $\underline{St_n}$ and $\overline{St_n}$, calculated accordingly, for $F^*(\underline{St_n}) = 0.25$ and $F^*(\overline{St_n}) = 0.75$, allows to not calculate s.f.d., which defines a errors II type $\beta$, that it is possible to carry to advantages of this way. Its lacks are necessity of increase twice numbers of modeled realizations of distribution $F_V^*(\xi)$, unjustified decrease in disorder $St_n$, the heuristic approach.

Algorithm of calculation s.f.d., describing the greatest deviation $F_\Sigma(\xi)$ and $F_V^*(\xi)$, provided that $F_V^*(\xi)$ it is unpresentable, consists of following sequence of calculations:

1. It is modeled next (from necessary N realizations) their sample n random numbers;
2. It is formed s.f.d. $F_V^*(\xi)$;
3. The greatest divergence between F is defined $F_\Sigma(\xi)$ and $F_V^*(\xi)$. Designate this size as $St_{n,e}$ where the index «e» corresponds to empirical character of sample.
   Having defined statistical characteristics of this sample $\{F_V^*(\xi)$ and $St_{n,e}\}$, start formation s.f.d. $F^*(St_n^*)$ on realizations of the greatest divergence between functions of distribution $F_\Sigma(\xi)$ and set (N) s.f.d. $F_V^*(\psi)$, modeled on s.f.d. $F_V^*(\xi)$. For what:
4. On s.f.d. $F_V^*(\xi)$ distribution is formed

$$F_V^*(\psi) = \begin{cases} 0 & \text{if } \psi \le \psi_1 \\ \dfrac{i-1}{n+1} + \dfrac{(\psi - \psi_i)}{(\psi_{i+1} - \psi_i)(n+1)} & \text{if } \psi_1 < \psi < \psi_{n+1} \\ 1 & \text{if } \psi \ge \psi_{n+1} \end{cases} \qquad (16)$$

5. Under standard program RAND the random number is modeled $\xi$ with uniform distribution in an interval [0,1];

6. On distribution (16) calculated corresponding probability ξ random number ψ. Calculations are spent under the formula

$$\psi = \psi_i + (\psi_{i+1} - \psi_i)[\xi \cdot (n+1) - (i-1)] \quad (17)$$

with i=1, (n+1)

7. Items 5 and 6 repeat n time;
8. On sample $\{\psi\}_n$ is under construction s.f.d. $F_V^*(\psi)$;
9. The greatest divergence between $F_\Sigma(\xi)$ and $F_V^*(\psi)$ is defined. Designate it through $St_n^*$;
10. Items (5÷9) will repeat N time;
11. Average value of a random variable $St_n^*$ defined. Designate it through $M^*(St_n^*)$;
12. On N to values, $St_n^*$ it is formed s.f.d. $F^*(St_n^*)$.

If to assume, that distribution $F^*(St_n^*)$ corresponds to the normal law of distribution, average value $M^*(St_n^*)$ is equal $St_{n,e}$ and corresponds $F^*(St_n^*) = \beta = 0{,}5$, for all realizations $St_{n,e}$, which probability $0.1<\alpha<0.5$, the preference should be given to assumption $H_2$. However, the assumption of the normal law of distribution of function $F^*(St_n^*)$ mismatches the validity. As an example on fig.5 the histogram of distribution of realizations $St_n^*$ for s.f.d. $F_V^*(\psi)$, resulted in table 1.

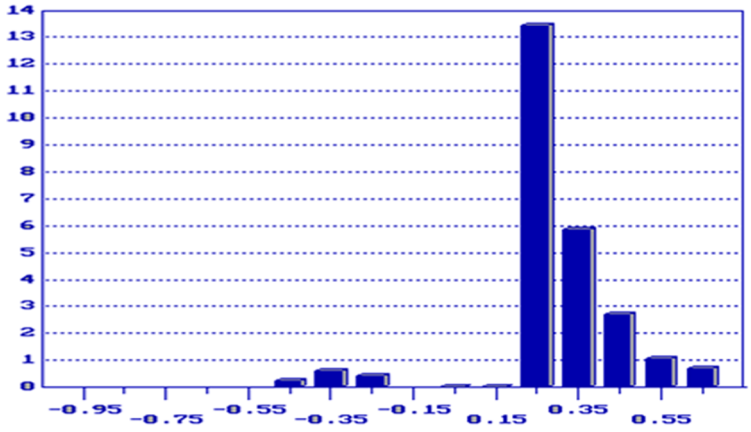



Fig.5. Histogram of realizations $St_n^*$

Let's enter into consideration two assumptions:

$H_1$ - sample $\{\psi\}_n$ reflects laws of distribution $F_\Sigma(\xi)$;

$H_2$ - sample $\{\psi\}_n$ does not reflect law of distribution $F_\Sigma(\xi)$.

The recommended algorithm of decision-making depends on a parity of average values of realizations $St_n$ and $St_n^*$. In this connection the distribution describing risk of the erroneous decision in function $St_n$ designate $Sh1(St_n)$, and in function $St_n^*$ – $Sh2(St_n)$.

At $M^*(St_n) < M^*(St_n^*)$

$$\left.\begin{aligned} Sh1(St_n) &= 1 - F^*(St_n) \\ Sh2(St_n) &= F(St_n^*) \end{aligned}\right\} \quad (18)$$

Algorithm of decision-making looks like:

$$\left.\begin{aligned} &If St_{n,э} \geq \overline{St_n}, \text{ then } H_2, \text{ else} \\ &If St_{n.э} \leq \underline{St_n^*}, \text{ then } H_1, \text{ else} \\ &If Sh1(St_n) << Sh2(St_n), \text{ then } H_2, \\ &Otherwise\ H_1 \end{aligned}\right\} \quad (19)$$

At $M^*(St_n) \geq M^*(St_n^*)$

$$\left.\begin{aligned} Sh1(St_n) &= 1 - F^*(St_n^*) \\ Sh2(St_n) &= F^*(St_n) \end{aligned}\right\} \quad (20)$$

Algorithm of decision-making looks like:

$$\left.\begin{aligned} &If St_n \geq \overline{St_n^*}, \text{ then } H_1, \text{ else} \\ &If St_{n,э} \leq \underline{St_n}, \text{ then } H_2, \text{ else} \\ &If Sh1(St_n^*) >> Sh2(St_n), \text{ then } H_2, \\ &Otherwise\ H_1 \end{aligned}\right\} \quad (21)$$

In the illustrative purposes on fig. 6 functions of distribution $Sh1(St_n)$ and $Sh2(St_n)$ are resulted. calculated according to table 1.

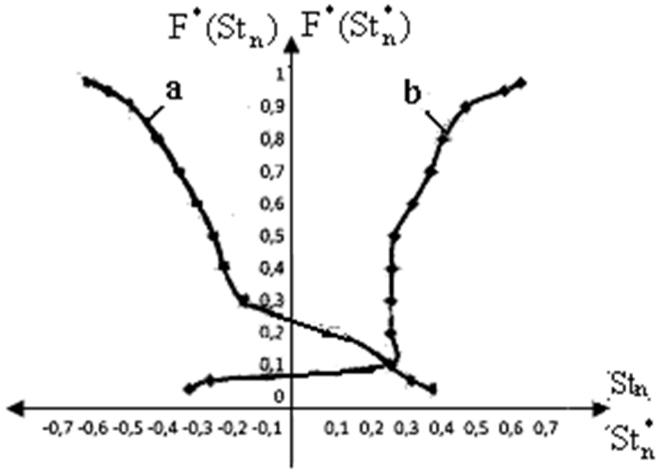



Fig. 6. Laws of change s.f.d. $F^*(St_n)$ and $F^*(St_n^*)$ for n=4: a – s.f.d. $F^*(St_n)$; b - $F^*(St_n^*)$

As $M^*(St_n)$ it has appeared less than $M^*(St_n^*)$ functions of distribution $Sh1(St_n)$. and $Sh2(St_n)$. were calculated accordingly under the formula (18).

In table 3 numerical values of the parameters defining result of the decision are systematized. As follows from tab. 3 as $Sh1(St_{n,e}) << Sh2(St_{n,e})$., the preference, according to (19) is given assumption $H_2$. In other words, attraction to the statistical analysis of size of a errors I type and errors II types, allows distinguish unpresentable samples.

Table 3

The basic parameters of calculation

| Parameter | Conditional designation | Estimation |
|---|---|---|
| 1. Number casual sample | n | 4 |
| 2. Average value of the greatest divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\xi)$ | $M^*(St_n)$ | -0,207 |
| 3. Average value of the greatest divergence of distributions $F_V^*(\psi)$ and $F_V^*(\psi)$ | $M^*(St_n^*)$ | 0,292 |
| 4. Empirical value of the greatest divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\xi)$ | $St_{n,e}$ | 0,257 |
| 5. Boundary values of an interval of change $St_n$ c $\alpha$=0.1 top | $\overline{St_n}$ | |
| bottom | $\underline{St_n}$ | 0,319 |
| 6. Boundary values of an interval of change $St_n^*$ with $\alpha$=0,01 top | | -0,567 |

| | | |
|---|---|---|
| bottom<br>7. Probability $St_{n,e}$ on s.f.d. $[1-F^*(St_n)]$<br>on s.f.d. $F^*(St_n^*)$<br>8. The assumption is accepted | $\overline{St_n^*}$<br>$\underline{St_n^*}$<br>$Sh1(St_{n,e})$<br>$Sh2(St_{n,e})$<br>H | 0,544<br>0,292<br>0,09<br>0,42<br>$H_2$ |

It is necessary to note, that attraction to an estimation of character of a divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\psi)$ distributions $F^*(St_n^*)$ for all realizations samples it is unjustified, as for of some from them, for example at $Sh1(St_n).\geq 0,5$ sample $\{\psi\}_n$ it is most truly representative, and at $Sh1(St_n)\leq 0,1$ – it is unpresentable.

There fore calculations s.f.d. $F^*(St_n^*)$ offered to spend for following conditions:

1. $M^*(St_n) < M^*(St_n^*)$

$$\left.\begin{aligned} \underline{St_{n,0.05}^*} < St_{n,э} < \overline{St_{n.0.95}} \\ \underline{St_{n,0.25}^*} \geq St_{n,э} \geq \overline{St_{n.0.75}} \end{aligned}\right\} \quad (22)$$

2. $M^*(St_n) > M^*(St_n^*)$

$$\left.\begin{aligned} \underline{St_{n,0.05}} < St_{n,э} < \overline{St_{n,0.95}^*} \\ \underline{St_{n,0.25}} \geq St_{n,э} \geq \overline{St_{n,0.75}^*} \end{aligned}\right\} \quad (23)$$

Critical values of statistics $St_n$ for $F^*(St_n)=0,25$ and average values $M^*(St_n)$ for N=25000 realizations $St_n$ and of some n are resulted in table 4.

Table 4

Bottom boundary $\left(\underline{St_n}\right)$ and average $M^*(St_n)$ values of statistics $St_n$

| N | n | $St_n$ ($F^*(St_n)=0.25$) | $M^*(St_n)$ | N | n | $St_n$ ($F^*(St_n)=0.25$) | $M^*(St_n)$ |
|---|---|---|---|---|---|---|---|
| 1 | 2 | -0.498 | -0.33 | 9 | 22 | -0.17 | -0.047 |
| 2 | 3 | -0.435 | -0.254 | 10 | 29 | -0.149 | -0.037 |
| 3 | 4 | -0.385 | -0.207 | 11 | 40 | -0.127 | -0.027 |
| 4 | 5 | -0.343 | -0.173 | 12 | 60 | -0.105 | -0.019 |
| 5 | 6 | -0.312 | -0.146 | 13 | 90 | -0.086 | -0.012 |
| 6 | 7 | --0.294 | -0.133 | 14 | 120 | -0.074 | -0.00- |
| 7 | 11 | -0.235 | -0.87 | 15 | 150 | -0.067 | -0.008 |
| 8 | 16 | -0.198 | -0.063 | | | | |

The computer technology of an estimation of parameters of individual reliability assumes automation of process of classification of multivariate data. For what, as initial data boundary values of statistics $St_n$ should entered. In this connection, by analogy to formulas (8) and (9), the opportunity of an estimation of dependence of boundary values $St_n$ from n was of interest.

The equations of regress received under the standard program of sedate transformation, are characterized by factor of determination R2: (R2> 0.999) and for of some $Sh1(St_n).=\alpha/2$ look like:

- for $Sh1\left(\overline{St_n}\right)=0,025$ $\quad \overline{St_n} = (1.23n^{0.52}-1)/n = (B_1 n^{0.52}-1)/n$ (24)

  and $Sh1\left(\underline{St_n}\right)=0,975$ $\quad \underline{St_n} = -1.23n^{-0.48} = -B_1/n^{0.48}$ (25)

- for $Sh1\left(\overline{St_n}\right)=0,05$ $\quad \overline{St_n} = (1.12n^{0.52}-1)/n = (B_2 n^{0.52}-1)/n$ (26)

and $Sh1(\underline{St_n}) = 0{,}95$ $\quad \underline{St_n} = -1.12n^{-0.48} = -B_2/n^{0.48}$ (27)

- for $Sh1(\overline{St_n}) = 0{,}1$ $\quad \overline{St_n} = (0.98n^{0.52} - 1)/n = (B_3 n^{0.52} - 1)/n$ (28)

and $Sh1(\underline{St_n}) = 0{,}9$ $\quad \underline{St_n} = -0.98n^{-0.48} = -B_3/n^{0.48}$ (29)

- for $Sh1(\overline{St_n}) = 0{,}25$ $\quad \overline{St_n} = (0.75n^{0.52} - 1)/n = (B_4 n^{0.52} - 1)/n$ (30)

and $Sh1(\underline{St_n}) = 0{,}75$ $\quad \underline{St_n} = -0.75n^{-0.48} = -B_4/n^{0.48}$ (31)

The equation of dependence of constant factors B from $\alpha$ with factor of determination R2: (R2> 0.993) looks like:

$$B = 0.652\left[Sh1\left(\overline{St_n}\right)\right]^{-0.175} \tag{32}$$

Thus, the bottom and top boundary values of statistics $St_n$ in view of the equation (12) calculated under following formulas:

$$\begin{aligned} \underline{St_n} &= -0.652\left[Sh1\left(\overline{St_n}\right)\right]^{-0.175} \cdot n^{-.048} \\ \overline{St_n} &= -\left[\underline{St_n} - \frac{1}{n}\right] \end{aligned} \tag{33}$$

For practical calculations $\underline{St_n}$ and $\overline{St_n}$ more often formulas (27) and (12) used.

## CONCLUSIONS

1. The interval nonparametric criterion of the control of conformity samples from n pseudo-random numbers is offered to the uniform law in an interval [0,1];
2. In a basis of criterion there is a distinction of distributions of positive and negative values of the greatest divergence of distributions $F_\Sigma(\xi)$ and $F_V^*(\xi)$;
3. Transition from statistics $D_n$ to statistics $St_n$ allows not only to simplify algorithm of calculation greatest divergences $F_\Sigma(\xi)$ and $F_V^*(\xi)$, but also to estimate an opportunity of use of statistics $St_n$ at an estimation of the greatest divergence s.f.d. $F_\Sigma^*(X)$ and $F_V^*(X)$, to estimate risk of the erroneous decision $Sh1(St_n)$;
4. Increase of accuracy of the control of conformity of distribution $St_n^*$ to the uniform law reached by practical realization of recommended algorithm of the decision-making considering not only a errors I type, but also the errors II type.

## REFERENCE

1. Farhadzadeh E.M., Muradaliyev A.Z., Farzaliyev Y.Z. Quantitative estimation of individual reliability of the equipment and devices of the power supply system. Journal: «Reliability: Theory&applications. R&RATA (Vol.7 No.4 (27)) 2012, December., USA, p.53-62
2. Gnedenko B.V., Beljaev J.K., Solovyov A.D. Mathematical methods in the theory of reliability. "Science", 1965, 524 p.
3. Kelton B, Law A. Imitational modeling. Classics CS. 3 СРб.: Peter, Kiev: Publishing group BHV, 2004, 847 p.
4. Ryabinin I.A. The heart of the theory and calculation of reliability of ship electro power systems. Shipbuilding. 1971, 454 p.
5. Farhadzadeh E.M. Technique of a statistical estimation of critical values of empirical distribution from theoretical. «Methodical questions of research of reliability of greater systems of power» SEI SO SA USSR, 16, Grozny, 1978, p.39-49.

ISSN 1932-2321