İstatistik, Journal of the Turkish Statistical Association
1999, Vol.2, No 2, pp. 65-69

THE SECOND RECORD TIME IN THE CASE OF ARBITRARY DISTRIBUTION

A. Stepanov
Kaliningrad State Technical University,
Department of Mathematics, Kaliningrad, Russia

Abstract

Let $X_1, X_2, ...$ be independent random variables with a common distribution function F. We study the second records time for this sequence in the case when F contains several atoms. We pay special attention to the case when the last point of increase is an atom.

Key Words: Record times, generating functions, moments.

Let $X_1, X_2, ...$ be a sequence of independent identically distributed random variables with common distribution function $F(x) = P\{X < x\}$. The sequence of record times $L(n)$ and record values $X_{(n)}$ is defined as follows:

$$L(1) = 1,\ X_{(1)} = X_1,$$

$$L(n+1) = \min\left\{j : j > L(n),\ X_j > X_{L(n)}\right\},$$

$$X_{(n)} = X_{L(n)}\ (n \geq 1).$$

The sequence of record times has been thoroughly studied in the case of continuous F. In this paper we consider $L(2)$ when F is an arbitrary distribution, i.e. $F(x)$ contains discrete, absolutely continuous and singular components simultaneously. Let $T = \{a_1, a_2, ..., a_n\}$ be the full set of atoms of the distribution $(a_1 < a_2 < ... < a_n)$.

Theorem 1. The following equality is valid

$$P\{L(2) = k\} = \frac{1}{k(k-1)} +$$

$$+ \sum_{i=1}^{n} \left(\frac{F^{k-1}(a_i) - F^{k-1}(a_i + 0)}{k-1} - \frac{F^k(a_i) - F^k(a_i + 0)}{k} \right) +$$

$$+\sum_{i=1}^{n}\left(F(a_i+0)-F(a_i)\right)F^{k-2}(a_i+0)\left(1-(a_i+0)\right) \tag{1}$$

Proof. We prove our result when the set T contains only one element a. We have

$$P\{L(2)=k\}=P\left\{L(2)=k, X_{(1)}<a\right\}+P\left\{L(2)=k, X_{(1)}=a\right\}+$$

$$+P\left\{L(2)=k, X_{(1)}>a\right\}=\int_{-\infty}^{a}F^{k-2}(x)\left(1-F(x)\right)dF(x)+$$

$$+\left(F(a+0)-F(a)\right)F^{k-2}(a+0)\left(1-F(a+0)\right)+\int_{a+0}^{\infty}F^{k-2}(x)\left(1-F(x)\right)dF(x)$$

It can be obtained from the last equation that

$$P\{L(2)=k\}=\frac{1}{k(k-1)}+\frac{F^{k-1}(a)-F^{k-1}(a+0)}{k-1}-$$

$$-\frac{F^{k}(a)-F^{k}(a+0)}{k}+\left(F(a+0)-F(a)\right)F^{k-2}(a+0)\left(1-F(a+0)\right).$$

Remark. Let γ be the very left point of $F(x)$ (if exists). Let $\gamma=a$ and $L(1)=k$ where $k=\min\{i: X_i\neq a\}$. Then the record values and times turn out to be the values and the times of one-sided successive approximations (*see* [3, 4, 5]).

Theorem 2. The generating function of $L(2)$ is

$$G(s)=Es^{L(2)}=$$

$$=s+(1-s)\log(1-s)-(1-s)\sum_{i=1}^{n}\log\left(\frac{1-sF(a_i+0)}{1-sF(a_i)}\right)+$$

$$+\sum_{i=1}^{n}\frac{s^2\left(F(a_i+0)-F(a_i)\right)\left(1-F(a_i+0)\right)}{1-sF(a_i+0)} \tag{2}$$

Proof. Here we also consider the situation when $F(x)$ has only one atom a. So

$$Es^{L(2)}=\sum_{k=2}^{\infty}\frac{s^k}{k(k-1)}+\sum_{k=2}^{\infty}\frac{s^k\left(F^{k-1}(a)-F^{k-1}(a+0)\right)}{k-1}-$$

$$-\sum_{k=2}^{\infty}\frac{s^k\left(F^{k}(a)-F^{k}(a+0)\right)}{k}+\left(F(a+0)-F(a)\right)\sum_{k=2}^{\infty}s^kF^{k-2}(a+0) \tag{3}$$

The first term of the sum corresponds to the generating function of $L(2)$ in the case of continuous distribution (*see* [1])

$$\sum_{k=2}^{\infty} \frac{s^k}{k(k-1)} = s + (1-s)\log(1-s).$$

Using the formula

$$\sum_{k=1}^{\infty} \frac{x^k}{k} = -\log(1-x)$$

we get

$$\sum_{k=2}^{\infty} \frac{s^k \left(F^{k-1}(a) - F^{k-1}(a+0)\right)}{k-1} = s\log\left(\frac{1-sF(a+0)}{1-sF(a)}\right),$$

$$\sum_{k=2}^{\infty} \frac{s^k \left(F^{k}(a) - F^{k}(a+0)\right)}{k} = s\left(F(a) - F(a+0)\right) - \log\left(\frac{1-sF(a+0)}{1-sF(a)}\right).$$

The last sum of (3) can be easily found by virtue of

$$\sum_{k=2}^{\infty} s^k F^{k-2}(a+0) = \frac{s^2}{1-sF(a+0)}$$

Corollary 1.

$$EL(2) = \log\left(\frac{1}{1-F(a_n)} \prod_{i=1}^{n-1} \frac{1-F(a_i+0)}{1-F(a_i)}\right) +$$

$$+ \sum_{i=1}^{n-1} \frac{\left(F(a_i+0) - F(a_i)\right)\left(2 - F(a_i+0)\right)}{1-F(a_i+0)}$$

if the last point of increase of $F(x)$ is an atom.

Proof. From Theorem 2 we get

$$G'(s) = \log\left(\frac{1}{1-s} \prod_{i=1}^{n} \frac{1-sF(a_i+0)}{1-sF(a_i)}\right) +$$

$$+ \sum_{i=1}^{n} \frac{(1-s)F(a_i+0)}{1-sF(a_i+0)} - \sum_{i=1}^{n} \frac{(1-s)F(a_i)}{1-sF(a_i)} +$$

$$+ s \sum_{i=1}^{n} \frac{\left(F(a_i+0) - F(a_i)\right)\left(2 - sF(a_i+0)\right)}{\left(1-sF(a_i+0)\right)^2},$$

the proof is completed if we put $F(a_n+0) = 1$ and $s \to 1$.

The existence of $EL(2)$ when the last point of increase of $F(x)$ is an atom could be mounted further.

Theorem 3. All possible moments of $L(n)$ do exist if the last point of increase of $F(x)$ is an atom.

Proof. The direct proof of the theorem for $EL(2)$ is possible by means of formula

$$E\{L(2)(L(2)-1)...(L(2)-k+1)\} = G^{(k)}(s)_{s=1} \ (k \geq 1).$$

and Theorem 2. The proof for any n for one particular case is presumably to be presented in Journal of Applied Statistical Science.

Corollary 2. The following formula is valid

$$E\left\{\frac{1}{L(2)+1}\right\} = \frac{1}{4}\sum_{i=1}^{n}\frac{(1-F(a_i))^2\left(\log(1-F(a_i))-\frac{3}{2}\right)+\frac{3}{2}-2F(a_i)}{2F^2(a_i+0)} -$$

$$-\sum_{i=1}^{n}\frac{(1-F(a_i))^2\left(\log(1-F(a_i+0))-\frac{3}{2}\right)+\frac{3}{2}-2F(a_i+0)}{2F^2(a_i+0)} +$$

$$+\sum_{i=1}^{n}\frac{(F(a_i+0)-F(a_i))(1-F(a_i))}{F^3(a_i+0)}(1-(1+F(a_i+0))^2$$

$$-2\log(1-F(a_i+0) \qquad (4)$$

Proof. We can exploit the equality

$$E\left\{\frac{1}{L(2)+1}\right\} = \left(\int_0^s G(x)dx\right)_{s=1}$$

proving (4).

Remark. Formula (4) reduces to

$$E\left\{\frac{1}{L(2)+1}\right\} = \frac{1}{4}$$

in the case of continuous distribution (*see* [2] *for comparison*).

References

[1] Renyi, A. (1976) On the extreme elements of observations. In Selected papers of Alfred Renyi., V.3, Budapest, Akademiai Kiado, pp.50-65.

[2] Nevzorov, V.B. (1987) Records. Theory Probab. Appl., V.32, pp.219-251.

[3] Stepanov, A.V. (1996) Extreme order statistics under changes of ordering relation. Theory Probab.Appl., V.41, pp.896-900.

[4] Stepanov, A.V. (1998) Limit laws for times of one sided sequential approximations. In Abs. of communications of Int. Conf. "Asymptotic methods in probab. and math. stat. in St. Petersburg university.

[5] Stepanov, A.V. (1998) Limit behavior of the times of one-sided successive approximations. Istatistic, Journal of the Turkish Statistical Association, V.1, pp.43-46.

ÖZET

$\{X_n\}_{n\geq 1}$ bağımsız ve ayni F dağılımına sahip olan rasgele değişkenler dizisi olsun. Bu dizinin ikinci recor zamanı F dağılımının sonlu sayıda atomlarının olduğu durumda incelenmiştir. Dağılım fonksiyonunun en son artma noktasının atom olduğu hal için özel irdemeler yapılmıştır.

İstatistik, Journal of the Turkish Statistical Association
1999, Vol.2, No 2, pp. 71-85

COMPUTER-AIDED DIAGNOSIS OF BREAST CANCER PART 1. MATHEMATICAL ASPECTS

Yu.I.Petunin, D.A.Klyushin
Department of Cybernetics, Kiev Shevchenko University
Kiev, Vladimirskaya, 64, DSP, 252601, Ukraine
e-mail: vm214@dcp.kiev.ua

K.P.Ganina, N.V.Boroday
R.E.Kavetsky Institute of Experimental Pathology,
Oncology and Radiobiology,
National Academy of Sciences of Ukraine
Kiev,Vasilkovskaya, 45, 252022, Ukraine

R.I.Andrushkiw
Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102, USA

Abstract

Mathematical aspects of the computer method of the diagnosis of breast cancer (CMG) and fibroadenomatosis (FAM) based on a single analysis of patient's buccal scrapes are discussed. The method of hypothesis verifing based on the simultaneous application of the quadratic, linear and order tests are developed. The algorithms of the construction of the confidence intervals and confidence ellipses is considered.

Key Words: Breast cancer, malignancy-associated changed, buccal scrapes, confidence limits, discriminant analysis, proximity measure, training samples.

1. Mathematical means of the computer diagnosis: confidence intervals

Many problems in the natural sciences and technology reduce to the following: stipulate an interval $I = (a, b)$ that contains the value x of general population G with a given probability β (for example, $\beta = 0.95$), $p\,(x \in (a, b)) = \beta$. This interval $I = (a, b)$ is called a confidence interval for the population G (or simply a confidence interval), the numbers a and b are called the lower and upper confidence limits respectively, the number β is a confidence level and the value $\alpha = 1 - \beta$ is the significance level.

The confidence interval $I = (a, b)$ can be constructed with the aid of the Chebyshev inequality [1]

$$p\,(|x - m(x)| \geq \lambda\sigma(x)) \leq \frac{1}{\lambda^2} \qquad (then\ \ \alpha \leq \frac{1}{\lambda^2}),$$

however, the confidence level is very crudely estimated. In this connection a large number of Chebyshev-type inequalities have been derived that refine the Chebyshev inequality under certain restriction on the distribution of x [1]. Nevertheless, these modifications of the Chebyshev inequality do not allow us to justify the empirical "3σ *rule*", which asserts that for distributions occurring in practice

$$p\,(|x - m(x)| \geq 3\sigma(x)) \leq 0.05.$$

where $m(x)$, $\sigma^2(x)$ are the expectation and variance of x, respectively.

Many mathematicians think that for any random variables $x \in S_n$, where

$$S_n = \left\{x : x = \sum_{i=1}^{n} y_i,\ y_i\ independent,\ F_{y_i} \equiv F(u)\right\}$$

the "3 rule" is fulfilled since the limit theorems are valid, but this is not correct. Indeed, the following assertion is proved [2]: for all $\lambda \geq 1$ and every natural number n the following inequality holds .

$$\sup_{x \in S_n} p\,(|x - m(x)| \geq \lambda\sigma(x)) \geq \frac{1 - \dfrac{1}{\lambda^2}}{\lambda^2}.$$

Thus, for every natural number n

$$\sup_{x \in S_n} p\,(|x - m(x)| \geq 3\sigma(x)) \geq \frac{8}{81} \approx 0.0987.$$

The problem of the justification of the "3σ *rule*" has been successfully solved for unimodal distribution in the way suggested by C.F.Gauss.

Recall that the random variable x is said to be unimodal if its probability density $f(u)$ has only one local maximum. More precisely, the variable x is unimodal if there exists a point a such that the distribution function of x is convex in the domain $(-\infty, a)$ and concave in the domain $(a, +\infty)$.

It turns out that the classical Gauss inequality

$$p\left(|x - m(x)| \geq \lambda\sigma(x)\right) \geq \frac{4}{9\lambda^2}.$$

which is valid for all symmetric unimodal random variables will be correct for all unimodal (not necessarily symmetric) random variables x, for all $\lambda > \sqrt{\frac{8}{3}}$[2,3]. In particular, for $\lambda = 3$

$$p\left(|x - m(x)| \geq 3\sigma(x)\right) \geq \frac{4}{81} \approx 0.49 < 0.05.$$

This refinement of the Gauss inequality is called the Vysochanskij-Petunin inequality [4,5]; at present there exist many generalizations and extensions of this inequality [4-6].

Another idea for solving the problem of construction of the confidence limits containing the bulk of the general population is based on the order statistics. Suppose G is some general population with unknown distribution function $F(u)$, $x_1, x_2, ..., x_n$ is a sample obtained from G as the result of simple random sampling. If we rearrange the sample values in increasing order $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ then we obtain the order statistics $x_{(i)}, i = 1, 2, ..., n$. It is shown [3,7] that for a population with a continuous distribution

$$p\left(x_{n+1} \in \left(x_{(i)}, x_{(j)}\right)\right) = \frac{j - i}{n + 1} \qquad (i < j), \tag{1}$$

where is a sample x_{n+1} value that does not depend on the sample $x_1, x_2, ..., x_n$ obtained by means of simple random sampling.

Consider two measurable (Borel) functions of n variables satisfying the inequality

$$f(u_1, u_2, ..., u_n) \leq g(u_1, u_2, ..., u_n) \quad (\forall (u_1, u_2, ..., u_n) \in R^n);$$

with the help of the functions $f(u_1, u_2, ..., u_n)$, $g(u_1, u_2, ..., u_n)$, and the sample $x_1, x_2, ..., x_n$ we can construct the random confidence interval

$$I = (f(x_1, x_2, ..., x_n), g(x_1, x_2, ..., x_n))$$

for the bulk of the general population G; this interval I is said to be an invariant confidence interval if

$$p\left(x_{n+1} \in I\right) = p\left(x \in \left(f\left(x_1, x_2, ..., x_n\right), g\left(x_1, x_2, ..., x_n\right)\right)\right) = const$$

for any general population G with the continuous distribution function $F(u)$.

It is shown [8] that the following statement is true: let $f\left(u_1, u_2, ..., u_n\right)$ and $g\left(u_1, u_2, ..., u_n\right)$ be continuous symmetric functions satisfying the inequality

$$f\left(u_1, u_2, ..., u_n\right) \leq g\left(u_1, u_2, ..., u_n\right) \quad \left(\forall \left(u_1, u_2, ..., u_n\right) \in R^n\right),$$

which coincide on the set from R^n with zero Lebesque measure. In order that the confidence interval $I = \left(f\left(x_1, x_2, ..., x_n\right), g\left(x_1, x_2, ..., x_n\right)\right)$ be invariant it is necessary and sufficient that $f\left(x_1, x_2, ..., x_n\right) = x_{(i)}$, $g\left(x_1, x_2, ..., x_n\right) = x_{(j)}$ $(i < j)$ where $x_{(i)}, x_{(j)}$ are some order statistics constructed with the help of sample $x_1, x_2, ..., x_n$. Therefore, a set B_n of all confidence intervals consists only of rational numbers, namely

$$B_n = \left\{0, \frac{1}{n+1}, \frac{2}{n+1}, ..., 1\right\}.$$

2. Statistical tests

The classical theory for the test of hypothesis using statistical criteria was created in the first half of XX century and was based on the Neumann-Pearson fundamental lemma [9]. This lemma allows one to obtain a powerful test in the case of two simple alternative hypotheses H and H'. We can construct this test and calculate its probability of errors of the first and second kind if we know the distribution functions $F_H\left(u_1, u_2, ..., u_n\right), F_{H'}\left(u_1, u_2, ..., u_n\right)$ corresponding to these hypotheses exactly. Unfortunately in practice of the statistical calculations we never know these distribution functions that is why it is necessary to construct the statistical criteria for the test of hypotheses which are based on training samples but not on the distribution functions $F_H\left(u_1, u_2, ..., u_n\right), F_{H'}\left(u_1, u_2, ..., u_n\right)$.

Suppose G and G' are two general populations and let $x = \left(x_1, x_2, ..., x_n\right)$ and $y = \left(y_1, y_2, ..., y_n\right)$ be samples from G and G', respectively. We shall assume for simplicity that these samples are obtained with the help of simple sampling. Let $z = \left(z_1, z_2, ..., z_k\right)$. be a sample from the general populations G or G' (we don't know where the general population are from). Denote $H = \{z \in G\}$ and $H' = \left\{z \in G'\right\}$. Suppose that (a, b) is a confidence interval for the bulk of G and $\left(a', b'\right)$ is a confidence interval for G'. We can construct these intervals with the help of the estimations of the mathematical expectation and variance in accordance with the "3σ *rule*" (or, more precisely, the "3s *rule*")

$$(a,b)=(\overline{x}-3s,\overline{x}+3s),\ \left(a^{'},b^{'}\right)=\left(\overline{x}^{'}-3s^{'},\overline{x}^{'}+3s^{'}\right),$$

where

$$\overline{x}=\frac{1}{n}\sum_{k=1}^{n}x_k,\ s^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)^2,$$
$$\overline{x}^{'}=\frac{1}{n}\sum_{k=1}^{n}x_k^{'},\ \left(s^{'}\right)^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k-\overline{x}^{'}\right)^2$$

or with the help of the order statistics $a=x_{(i)}$, $b=x_{(j)}$ $(i<j)$, $a^{'}=x_{(r)}^{'}$, $b=x_{(t)}^{'}$ $(r<t)$.

Assume for simplicity that $z=z_1$ (i.e. $k=1$). Then the statistical criterion based on the training samples has the form

1) if $z\in\left(a,a^{'}\right)$ then H is accepted;

2) if $z\in\left(b,b^{'}\right)$ then $H^{'}$ is accepted;

3) if $z\in\left(a^{'},b\right)$ then decision is not accepted.

The probability of the error of the first kind is

$$\alpha^{*}=p\left(H^{'}\,|H\right)=p\left(z\in\left(b,b^{'}\right)\middle|\,H\right)\leq p\left(z\notin(a,b)|\,H\right)=\alpha$$
$$(\alpha^{'}\approx 0.05\ \textit{for example}),$$

where is α a significance level of the confidence interval (a,b). Similarly the probability of the error of the second kind is

$$\beta^{*}=p\left(H\left|H^{'}\right.\right)=p\left(z\in\left(a,a^{'}\right)\middle|\,H^{'}\right)\leq p\left(z\notin\left(a^{'},b^{'}\right)\middle|\,H^{'}\right)=\alpha^{'}$$
$$(\alpha^{'}\approx 0.05\ \textit{for example}),$$

Now, let us consider a general case where k is any natural number . If the order statistics $x_{(1)},x_{(n)},y_{(1)},y_{(m)}$ satisfy the inequality

$$x_{(1)}\leq y_{(1)}\leq x_{(n)}\leq y_{(m)},\ .$$

then $a=x_{(1)}$, $b=x_{(n)}$, $a^{'}=y_{(1)}$, $b^{'}=y_{(m)}$.

Let ϑ be any number from $[0,1)$ and $k\geq 2$. Define a number r by the formula $r=r\left(\vartheta,k\right)=\left[\frac{k}{1+\vartheta}\right]$, where $[a]$ is an integer part of the number a. A statistical criterion T_ϑ for the test of hypothesis H and $H^{'}$ on the basis of the sample $z=(z_1,z_2,...,z_k)$ is defined in the following way [10]:

1) H is accepted if at least r sample values of the sample being investigated $z=(z_1,z_2,...,z_k)$ belong to the interval $\left(-\infty,y_{(1)}\right)$;

2) if at least r value from the sample $z=(z_1,z_2,...,z_k)$ belong to the interval $\left(x_{(n)},+\infty\right)$ then the hypothesis $H^{'}$ is accepted;

3) in any other cases the decision in not accepted.

The probability of the error of the first kind is

$$\alpha^* = p\left(H' \,|H\right) = \frac{B\left(n+k-r+1,r\right)}{B\left(k-r+1,r\right)} = \frac{(k-r+1)(k-r+2)...k}{(n+k-r+1)(n+k-r+2)...(n+k)} = o\left(\frac{1}{n^{r-1}}\right)$$

and the second kind is

$$\beta^* = p\left(H\left|H'\right.\right) = \frac{B\left(m+k-r+1,r\right)}{B\left(k-r+1,r\right)} = \frac{(k-r+1)(k-r+2)...k}{(m+k-r+1)(m+k-r+2)...(m+k)} = o\left(\frac{1}{m^{r-1}}\right)$$

where $B(a,b)$ is a beta-function.

Denote be CD a procedure of the non-acceptance decision. Assume that the simultaneous distribution functions $F_G\left(u_1,u_2,...,u_n\right)$ and $F_{G'}\left(u_1,u_2,...,u_n\right)$ of the samples $z=\left(z_1,z_2,...,z_n\right)$, when $z \in G$ and $z^{'} \in G^{'}$, are known and we construct the statistical criterion T for the test of hypothesis $H=\{z \in G\}$ and $H^{'}=\left\{z \in G^{'}\right\}$ satisfying the following conditions

$$p\left(H' \,|H\right) \leq \alpha, \;\; p\left(H\left|H'\right.\right) \leq \beta$$

and

$$p\left(CD\,|H\right) + \;\; p\left(CD\left|H'\right.\right) \rightarrow \min,$$

where α and β are fixed number. This criterion is called an optimal one.

Denote by $W, V, \overline{T} \subset R^n : W \cup V \cup \overline{T} = R^n$ the regions of the decision acceptance of the criterion T : $z \in W \Rightarrow H$ is accepted, $z \in V \Rightarrow H^{'}$ is accepted, $z \in \overline{T} \Rightarrow$ decision is not accepted. It is shown [11] that these regions are defined by the likelihood ratio

$$h\left(u\right) = \frac{f_{H'}\left(u_1,u_2,...,u_n\right)}{f_H\left(u_1,u_2,...,u_n\right)},$$

where f_H , $f_{H'}$ are probability densities.

Consider the following criterion for the test of hypothesis H about equality distribution functions $F_G\left(u\right)$ and $F_{G'}\left(u\right)$ of the general population G and $G^{'}$, respectively [12]. Let $x=\left(x_1,x_2,...,x_n\right) \in G$ and $x^{'}=\left(x_1^{'},x_2^{'},...,x_m^{'}\right) \in G^{'}$, $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$, $x_{(1)}^{'} \leq x_{(2)}^{'} \leq ... \leq x_{(m)}^{'}$ be order statistics. Suppose that $F_G\left(u\right) = F_{G'}\left(u\right)$. Then on the strength of the formulae (1)

$$p\left(A_{ij}\right) = p\left(x_k^{'} \in \left(x_{(i)},x_{(j)}\right)\right) = p_{ij} = \frac{j-i}{n+1}.$$

If we have the sample $x' = \left(x'_1, x'_2, ..., x'_m\right)$ we can find a frequency h_{ij} of the random event A_{ij} and confidence limits $p_{ij}^{(1)}$, $p_{ij}^{(2)}$ for the probability p_{ij} corresponding to given significance level β : $B = \left\{p_{ij} \in \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)\right\}$, $p(B) = 1 - \beta$. These limits have been calculated by the formulae [9]:

$$p_{ij}^{(1)} = \frac{h_{ij}m + 0.5g^2 - g\sqrt{h_{ij}(1 - h_{ij})m + 0.25g^2}}{m + g^2},$$
$$p_{ij}^{(2)} = \frac{h_{ij}m + 0.5g^2 + g\sqrt{h_{ij}(1 - h_{ij})m + 0.25g^2}}{m + g^2}, \qquad (2)$$

where g satisfies condition $\varphi(g) = 1 - \frac{\beta}{2}$, $\varphi(u)$ is a function of the normed normal distribution (if m is small then according to "3σ *rule*" $g = 3$).

It should be noted that the confidence limits (2) are not exact confidence limits corresponding to the given significance level even for the Bernoulli model but they will be the asymptotic confidence limits; in our case a calculation of the exact confidence intervals is based on the investigation of the so-called generalized Bernoulli model [13-16] (Matveychuk-Petunin model, or MP model, by terminology N.Johnson and S.Kotz [15,16]) and represents very difficult (almost hopeless) problem.

Denote by N a number of all confidence intervals $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$, $N = \frac{n(n-1)}{2}$ and L a number of those intervals I_{ij} which contain the probabilities p_{ij}; put $h = \rho(F_G, F_{G'}) = \rho\left(x, x'\right) = \frac{L}{N}$. As far as h will be a frequency of the random event $B = \{p_{ij} \in I_{ij}\}$ having the probability $p(B) = 1 - \beta$, then by setting $h_{ij} = h$, $m = N$ and $g = 3$ in formulae (2) we get a confidence interval $I = \left(p^{(1)}, p^{(2)}\right)$ for the probability $p(B)$ which have confidence level approximately equal to 0.95. A criterion for the test of hypothesis H with significance level approximately equal to 0.05 may be formulated by the following way: if the confidence interval I contains the probability $p(B)$, then hypothesis H is accepted, otherwise it is rejected. The statistics h is called p-statistics (Petunin's statistics); it is a measure of the proximity $\rho\left(x, x'\right)$ between the samples x and x'.

The last test described above is applied in diagnosis of breast cancer on the basis of the concentration of DNA in the interphase nuclei of the epithelial cells from mucous coat of the stomach and epithelium of ductus and lobes of the mammary gland under pre-tumor and tumor processes in these organs.

3. The proximity measures between a sample and a set of the training samples used in the computer diagnosis of the breast cancer

A process of the diagnosis of the cancer tumor in the mammary gland consists of three stages.

1). First, on the basis of a scrape from a mucous coat of the oral cavity we must obtain a data set investigating about 10-30 cells.

2). On the second stage, by using the Feulgan reaction we obtain the DNA (deoxyribonucleic acid) scanogram of the intrephase nuclei of these cells $x_1, x_2, ..., x_{15}$; then we are recording the main indices (all together 15) of the scanograms of the interphase nuclei of these cells: (in detail the scanograms and these indices will be described in part II of this paper. (see also [17,18]). We shall call the vector $X_C = (x_1, x_2, ..., x_{15})$ the indication vector of the cell C. There is no need to use only these indices: it may also be suggested others indices, which will be more informative and will be able to give more accurate description of the nuclei texture (see, for example, texture descriptors in [19]). Thus we get for every cell (for example, a cell number i) 15 indices $x_1^{(i)}, x_2^{(i)}, ... x_{15}^{(i)}$, where $i = 1, 2, ..., n$, $10 \leq n \leq 30$.

3). On the third stage we apply the special algorithms of the statistical and geometrical theory of the patter recognition for the detection of the changes in somatic nonmalignized cells under the breast cancer and fibroadenomatosis. These tests and algorithms will be considered in more detail in the next part; in this one we describe three stages of the above mentioned tests and the main proximity measures on which they are based.

1) At the first stage we form two groups of the patients G_1 and G_2 ; the first group G_1 consists of the patients which are suffering from the cancer of the mammary gland (CMG) and the second one G_2 consists of patients having the fibroadenoma (FAM) (the diagnoses of the patients of each group must be verified exactly!).

These are so-called training or standard groups; on the basis of these groups we will be diagnose the diseases.

2) At the second stage with the help of the p-statistic (Petunin's statistics) we calculate distances (measures of proximity) between the index of the scanograms of the patient and the corresponding indices of patients of the groups G_1 and G_2. This is made in the following way. Assume that the patient Q_i belongs to the first group $G_1 = \{Q_1, Q_2, ..., Q_n\}$. We exclude this patient from the group G_1 so that we get the group $G_1^{(i)} = \{Q_1, Q_2, ..., Q_{i-1}, Q_{i+1}, ..., Q_n\}$.

Let

$$X_{C_1}^k = \left(x_{1k}^{(1)}, x_{2k}^{(1)}, ..., x_{15k}^{(1)}\right)$$

$$X_{C_2}^k = \left(x_{1k}^{(2)}, x_{2k}^{(2)}, ..., x_{15k}^{(2)}\right)$$

.....

$$X^k_{C_{j_k}} = \left(x^{(j_k)}_{1k}, x^{(j_k)}_{2k}, ..., x^{(j_k)}_{15k}\right)$$

$(k = 1, 2, ..., n;\ 10 \leq j_k \leq 30)$ be indication vectors of the cells of the patients Q_k. Here $X^k_{C_i} = \left(x^{(i)}_{1k}, x^{(i)}_{2k}, ..., x^{(i)}_{15k}\right)$ is an indication vector of the cell C_i of the patient . Then we form the training samples for every index x_i, $i = 1, 2, ..., 15$; the first training sample for the index x_1 be

$$X^{(1)}_1 = \left(x^{(1)}_{11}, x^{(2)}_{11}, ..., x^{(j_2)}_{11}\right) \text{ (from the second patient)}$$

$$X^{(1)}_2 = \left(x^{(1)}_{12}, x^{(2)}_{12}, ..., x^{(j_2)}_{12}\right) \text{ (from the second patient)}$$

.....

$$X^{(1)}_n = \left(x^{(1)}_{1n}, x^{(2)}_{1n}, ..., x^{(j_n)}_{1n}\right) \text{ (from the n-th patient);}$$

the second training sample (for index x_2) be

$$X^{(2)}_1 = \left(x^{(1)}_{21}, x^{(2)}_{21}, ..., x^{(j_2)}_{21}\right) \text{ (from the second patient)}$$

$$X^{(2)}_2 = \left(x^{(1)}_{22}, x^{(2)}_{22}, ..., x^{(j_2)}_{22}\right) \text{ (from the second patient)}$$

.....

$$X^{(2)}_n = \left(x^{(1)}_{2n}, x^{(2)}_{2n}, ..., x^{(j_n)}_{2n}\right) \text{ (from the n-th patient);}$$

and so on. Finally, we get last training sample, i.e. training sample for the 15th index $X^{(15)}_1, X^{(15)}_2, ..., X^{(15)}_n$ (n is the number patients of the group G_1). Now we can calculate the values of the p-statistics for the samples of the i-th patient and the corresponding samples of other patients with number $k(k \neq i)$ (i is fixed!):

$$\rho^{(1)}_{ik} = \rho\left(X^{(1)}_i, X^{(1)}_k\right),\ \rho^{(2)}_{ik} = \rho\left(X^{(2)}_i, X^{(2)}_k\right),\ ...,\ \rho^{(15)}_{ik} = \rho\left(X^{(15)}_i, X^{(15)}_k\right)$$

and find the values of the averaged p-statistics

$$\rho^{(1)}_i = \frac{1}{n-1} \sum_{k=1, k\neq i}^{n} \rho\left(X^{(1)}_i, X^{(1)}_k\right),$$

$$\rho^{(2)}_i = \frac{1}{n-1} \sum_{k=1, k\neq i}^{n} \rho\left(X^{(2)}_i, X^{(2)}_k\right),$$

.....

$$\rho^{(15)}_i = \frac{1}{n-1} \sum_{k=1, k\neq i}^{n} \rho\left(X^{(15)}_i, X^{(15)}_k\right),$$

(i is fixed!) which represent the measure of the proximity between the patient Q_i (more precisely between its indices) and the group $G_1^{(i)}$, $i = 1, 2, ..., n$.

Replacing the patient Q_i by a patient $\overline{Q}_j$ from the group G_2 (remind that G_2 consists of the patients with fibroadenomatosis) we get similar averaged p-statistics for the group G_2:

$$\overline{\rho}_j^{(1)} = \frac{1}{n} \sum_{k=1}^{n} \rho\left(\overline{X}_j^{(1)}, X_k^{(1)}\right),$$
$$\overline{\rho}_j^{(2)} = \frac{1}{n} \sum_{k=1}^{n} \rho\left(\overline{X}_j^{(2)}, X_k^{(2)}\right),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\overline{\rho}_j^{(15)} = \frac{1}{n} \sum_{k=1}^{n} \rho\left(\overline{X}_j^{(15)}, X_k^{(15)}\right).$$

where $j = 1, 2, ..., m$; $m = card\, G_2$, $\overline{X}_j^{(t)}$ $(t = 1, 2, ..., 15)$ is a corresponding index of the patient $\overline{Q}_j$. If we replace the group G_1 by G_2 and perform the similar calculations, we obtain the averaged p-statistics $d_j^{(1)}, d_j^{(2)}, \ldots, d_j^{(15)}$ $(j = 1, 2, ..., m)$ and $\overline{d}_i^{(1)}, \overline{d}_i^{(2)}, \ldots, \overline{d}_i^{(15)}$ $(i = 1, 2, ..., m)$:

$$d_j^{(1)} = \frac{1}{m-1} \sum_{k=1, k\neq j}^{m} \rho\left(\overline{X}_j^{(1)}, \overline{X}_k^{(1)}\right),$$
$$d_j^{(2)} = \frac{1}{m-1} \sum_{k=1, k\neq j}^{m} \rho\left(\overline{X}_j^{(2)}, \overline{X}_k^{(2)}\right),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$d_j^{(15)} = \frac{1}{m-1} \sum_{k=1, k\neq j}^{m} \rho\left(\overline{X}_j^{(15)}, \overline{X}_k^{(15)}\right),$$

and

$$\overline{d}_i^{(1)} = \frac{1}{m} \sum_{k=1}^{m} \rho\left(\overline{X}_k^{(1)}, X_i^{(1)}\right),$$
$$\overline{d}_i^{(2)} = \frac{1}{m} \sum_{k=1}^{m} \rho\left(\overline{X}_k^{(2)}, X_i^{(2)}\right),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\overline{d}_i^{(15)} = \frac{1}{m} \sum_{k=1}^{m} \rho\left(\overline{X}_k^{(15)}, X_i^{(15)}\right).$$

3) At the third stage of the test we produce coupling of these averaged p-statistics $\left(\rho_i^{(t)}, \rho_i^{(s)}\right)$, $\left(\overline{\rho}_j^{(t)}, \overline{\rho}_j^{(s)}\right)$, $i = 1, 2, ..., n$; $j = 1, 2, ..., m$; similarly we get points $\left(d_j^{(t)}, d_j^{(s)}\right)$, $\left(\overline{d}_i^{(t)}, \overline{d}_i^{(s)}\right)$, $i = 1, 2, ..., n$; $j = 1, 2, ..., m$. After we construct so-called confidence ellipses $E_{ts}\,(t, s = 1, ., ..., 15)$ containing the averaged p-statistics $\left(\rho_i^{(t)}, \rho_i^{(s)}\right), i =$

$1, 2, ..., n$ for the group G_1, i.e. the ellipse with minimal area containing the points $\left(\rho_i^{(t)}, \rho_i^{(s)}\right)$, $i = 1, 2, ..., n$. Then we construct the confidence ellipses $\overline{E}_{ts}$ for the averaged p-statistics $\left(\overline{\rho}_j^{(t)}, \overline{\rho}_j^{(s)}\right)$, $j = 1, 2, ..., m$; and similar ellipses E_{ts}^* and $\overline{E}_{ts}^*$ by using the points $\left(d_j^{(t)}, d_j^{(s)}\right)$, $\left(\overline{d}_i^{(t)}, \overline{d}_i^{(s)}\right)$, $i = 1, 2, ..., n; j = 1, 2, ..., m$. More precisely we construct ellipses with the help of the algorithm which gives the approximate solution of this problem [20,21]. This part of the test we shall call as quadratic test because it is pased on quadratic discriminant function.

In addition, as the second part of the test, we consider a linear discriminant Fisher functions $f_{ts}(u, v)$ and $f_{ts}^*(u, v)$ separating the set $M_{ts}^{\rho} = \left\{\left(\rho_i^{(t)}, \rho_i^{(s)}\right)\right\}$, $i = 1, ..., n$ from the set $\overline{M}_{ts}^{\rho} = \left\{\left(\overline{\rho}_j^{(t)}, \overline{\rho}_j^{(s)}\right)\right\}$, $j = 1, ..., m$ and the set $\overline{M}_{ts}^{d} = \left\{\left(\overline{d}_i^{(t)}, \overline{d}_i^{(s)}\right)\right\}$, $i = 1, ..., n$ from the set $M_{ts}^{d} = \left\{\left(d_j^{(t)}, d_j^{(s)}\right)\right\}$, $j = 1, ..., m$ respectively. The function $f_{ts}(u, v)$ is constructed such way that straight line $l_{ts} = \{(u, v) : f_{ts}(u, v) = 0\}$ is perpendicular to a segment connecting the centers of the sets M_{ts}^{ρ} and $\overline{M}_{ts}^{\rho}$, and passes through the middle of this segment; similarly, $f_{ts}^*(u, v)$; in addition the centre of the set M_{ts}^{ρ} belongs to the lower halfplane π_{ts} and the center of the set $\overline{M}_{ts}^{\rho}$ belongs to the upper one λ_{ts} (similarly π_{ts}^*, λ_{ts}^*). Thus we have for the 15 indices 210 pairs of the ellipses $\left(E_{ts}, \overline{E}_{ts}\right)$ and $\left(E_{ts}^*, \overline{E}_{ts}^*\right)$, $t < s; t, s = 1, 2..., 15$ and moreover 210 pairs of the half-planes (π_{ts}, λ_{ts}) and $(\pi_{ts}^*, \lambda_{ts}^*)$, $t < s$; $t, s = 1, 2..., 15$. We shall name this test as linear test.

Let Q be a patient suffering from the cancer of the breast (hypothesis H_1) of the fibroadenomatosis (hypothesis H_2). By using the algorithms mentioned above we can calculate for this patient the averaged p-statistics $\rho_Q^{(t)}, d_Q^{(t)}$, $t = 1, 2, ..., 15$:

$$\rho_Q^{(t)} = \frac{1}{n} \sum_{k=1}^{n} \rho\left(X_Q^{(t)}, X_k^{(t)}\right), \; d_Q^{(t)} = \frac{1}{m} \sum_{k=1}^{m} \rho\left(X_Q^{(t)}, \overline{X}_k^{(t)}\right),$$

where $X_Q^{(t)}$ is a corresponding index (sample) of the patients Q and form the points $\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right)$, $\left(d_Q^{(t)}, d_Q^{(s)}\right)$, $t < s$; $t, s = 1, 2..., 15$. Consider the following random events

$$A_1 = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in E_{ts}\right\}, \qquad A_2 = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in \overline{E}_{ts}\right\},$$
$$A_3 = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in E_{ts} - \overline{E}_{ts}\right\}, \; A_4 = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in \overline{E}_{ts} - E_{ts}\right\},$$
$$A_1^* = \left\{\left(d_Q^{(t)}, d_Q^{(s)}\right) \in E_{ts}^*\right\}, \qquad A_2^* = \left\{\left(d_Q^{(t)}, d_Q^{(s)}\right) \in \overline{E}_{ts}^*\right\},$$
$$A_3^* = \left\{\left(d_Q^{(t)}, d_Q^{(s)}\right) \in E_{ts}^* - \overline{E}_{ts}^*\right\}, \; A_4^* = \left\{\left(d_Q^{(t)}, d_Q^{(s)}\right) \in \overline{E}_{ts}^* - E_{ts}^*\right\},$$
$$B_1 = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in \pi_{ts}\right\}, \qquad B_2 = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in \lambda_{ts}\right\},$$
$$B_1^* = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in \pi_{ts}^*\right\}, \qquad B_2 = \left\{\left(\rho_Q^{(t)}, \rho_Q^{(s)}\right) \in \lambda_{ts}^*\right\}, \; t < s$$

and

$$C_1 = A_3 \cup A_4, \; C_2 = A_4 \cup A_3^*, C_3 = A_1 \cup A_2^*,$$
$$C_4 = A_2 \cup A_1^* \, , C_5 = B_1 \cup B_2^*, \; C_6 = B_1 \cup B_1^* \, .$$

Denote by $h_i = h(C_i), i = 1, 2, ..., 6$ a frequency of the event C_i under 210 tests (experiments) when $t, s = 1, 2, ..., 15; t < s$. By using the formulae (2) we can get the asymptotic confidence limits corresponding to the given significance level for the probability $p_i = p(C_i)$ on the basis of the frequency (in this connection we must take $h_{ij} = h_j, m = 210$); these limits will be apparently too wide. We shall call the frequency h_1 the index of CMG (briefly CMG) and h_3 - the total CMG (briefly TCMG) as far as these indices are the proximity measures between the scanograms of the interphase nuclei of the cells of the examined patient Q and the corresponding scanograms of the patients which are suffering from carcinoma of the mammary gland; by virtue of the similar arguments we shall denote the frequencies h_2 and h_4 as the FAM and the total FAM (TFAM) index, respectively. By analogy we shall call frequencies h_5 and h_6 the linear CMG (LCMG) and the linear FAM (LFAM) index.

The third part of our test is so-called order test. The first stage of this test is the same as the previous tests. Let $X^k_{C_i} = \left(x^{(i)}_{1k}, x^{(i)}_{2k}, ..., x^{(i)}_{15k}\right), i = 1, 2, ..., j_k; k = 1, 2, ..., n$ be an indication vector of the cell C_i of the patient Q from the group G_1 and $Y^k_{D_i} = \left(y^{(i)}_{1k}, y^{(i)}_{2k}, ..., y^{(i)}_{15k}\right), i = 1, 2, ..., l_k; k = 1, 2, ..., m$ be the corresponding indication vector of the cell D_i of the patient $\overline{Q}_k \in G_2$.

At the second stage the averaged indication vector

$$\widetilde{X}^{(k)} = \frac{1}{j_k} \sum_{t=1}^{j_k} X^k_{C_t} = (\widetilde{x}_{1k}, \widetilde{x}_{2k}, ..., \widetilde{x}_{15k})$$

for every patient $Q_k \in G_1$ is calculated; similarly for every $\overline{Q}_k \in G_2$ the averaged indication vector $\widetilde{Y}^{(k)}$ has the form

$$\widetilde{Y}^{(k)} = \frac{1}{l_k} \sum_{t=1}^{l_k} Y^k_{D_t} = (\widetilde{y}_{1k}, \widetilde{y}_{2k}, ..., \widetilde{y}_{15k}).$$

Put

$$x^{\min}_{tk} = \min\left(x^{(1)}_{tk}, x^{(2)}_{tk}, ..., x^{(j_k)}_{tk}\right), k = 1, 2, ..., n; t = 1, 2, ..., 15;$$
$$x^{\max}_{tk} = \max\left(x^{(1)}_{tk}, x^{(2)}_{tk}, ..., x^{(j_k)}_{tk}\right), k = 1, 2, ..., n; t = 1, 2, ..., 15;$$
$$y^{\min}_{tk} = \min\left(y^{(1)}_{tk}, y^{(2)}_{tk}, ..., y^{(j_k)}_{tk}\right), k = 1, 2, ..., m; t = 1, 2, ..., 15;$$
$$y^{\max}_{tk} = \max\left(y^{(1)}_{tk}, y^{(2)}_{tk}, ..., y^{(j_k)}_{tk}\right), k = 1, 2, ..., m; t = 1, 2, ..., 15;$$
$$a^{\min}_t = \min\left(x^{\min}_{t1}, x^{\min}_{t2}, ..., x^{\min}_{tk}\right), a^{\max}_t = \max\left(x^{\min}_{t1}, x^{\min}_{t2}, ..., x^{\min}_{tk}\right),$$
$$k = 1, 2, ..., n; t = 1, 2, ..., 15;$$
$$b^{\min}_t = \min\left(x^{\max}_{t1}, x^{\max}_{t2}, ..., x^{\max}_{tk}\right), b^{\max}_t = \max\left(x^{\max}_{t1}, x^{\max}_{t2}, ..., x^{\max}_{tk}\right),$$
$$k = 1, 2, ..., n; t = 1, 2, ..., 15;$$
$$\overline{a}^{\min}_t = \min\left(y^{\min}_{t1}, y^{\min}_{t2}, ..., y^{\min}_{tk}\right), \overline{a}^{\max}_t = \max\left(y^{\min}_{t1}, y^{\min}_{t2}, ..., y^{\min}_{tk}\right),$$
$$k = 1, 2, ..., m; t = 1, 2, ..., 15;$$
$$\overline{b}^{\min}_t = \min\left(y^{\max}_{t1}, y^{\max}_{t2}, ..., y^{\max}_{tk}\right), \overline{b}^{\max}_t = \max\left(y^{\max}_{t1}, y^{\max}_{t2}, ..., y^{\max}_{tk}\right),$$

$k = 1, 2, ..., m;\ t = 1, 2, ..., 15;$
$c_t^{\min} = \min\limits_{k=1,...,n} \widetilde{x}_{tk}, c_t^{\max} = \max\limits_{k=1,...,n} \widetilde{x}_{tk}, \overline{c}_t^{\min} = \min\limits_{k=1,...,m} \widetilde{y}_{tk},$
$\overline{c}_t^{\max} = \max\limits_{k=1,...,m} \widetilde{y}_{tk},\ t = 1, 2, ..., 15.$

Then $a_t^{\min}$, $a_t^{\max}, \overline{a}_t^{\min}, \overline{a}_t^{\max}$, $b_t^{\min}$, $b_t^{\max}, \overline{b}_t^{\min}, \overline{b}_t^{\max}, c_t^{\min}, c_t^{\max}, \overline{c}_t^{\min}, \overline{c}_t^{\max}$ will be minimal and maximal order statistics, respectively. By means of these order statistics the confidence intervals $\alpha_t = \left(a_t^{\min}, a_t^{\max}\right)$, $\beta_t = \left(b_t^{\min}, b_t^{\max}\right)$, $\gamma_t = \left(c_t^{\min}, b_t^{\max}\right)$, $\overline{\alpha}_t = \left(\overline{a}_t^{\min}, \overline{a}_t^{\max}\right)$, $\overline{\beta}_t = \left(\overline{b}_t^{\min}, \overline{b}_t^{\max}\right)$, $\overline{\gamma}_t = \left(\overline{c}_t^{\min}, \overline{c}_t^{\max}\right)$ are formed.

Let Q be an examined patient and $X_{C_i} = \left(x_1^{(i)}, x_2^{(i)}, ..., x_{15}^{(i)}\right)$, $i = 1, 2, ..., j;$, i=1,2,...,j be indication vectors of this patient. At the third stage of the order test we calculate the averaged indication vector of the Q:

$$\widetilde{X} = \frac{1}{j} \sum_{t=1}^{j} X_{C_t} = (\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_{15})$$

and indices $x_t^{\min} = \min\limits_{i=1,...,j} x_t^{(i)}$, $x_t^{\max} = \max\limits_{i=1,...,j} x_t^{(i)}$, $t = 1, 2, ..., 15$ and then we define the indicators of the falling outside the limits $I_t^{\min}$, I_t, $I_t^{\max}$ $(t = 1, 2, ..., 15)$:

$$I_t^{\min} = \begin{cases} 1, if\ x_t^{\min} \notin \alpha_t, \\ 0, if\ x_t^{\min} \in \alpha_t, \end{cases}, \quad I_t = \begin{cases} 1, if\ \widetilde{x}_t \notin \gamma_t, \\ 0, if\ \widetilde{x}_t \in \gamma_t, \end{cases}, \quad I_t^{\max} = \begin{cases} 1, if\ x_t^{\max} \notin \beta_t, \\ 0, if\ x_t^{\max} \in \beta_t, \end{cases}$$

Similarly the indicators $\overline{I}_t^{\min}$, $\overline{I}_t$, $\overline{I}_t^{\max}$ are defined. Then we evaluate the indices

$$\alpha_1 = \sum_{t=1}^{15} \left(I_t^{\min} + I_t + I_t^{\max}\right), \alpha_2 = \sum_{t=1}^{15} \left(\overline{I}_t^{\min} + \overline{I}_t + \overline{I}_t^{\max}\right).$$

These indices are also the proximity measures between the scanogramms of the interphase nuclei of the cells of the examined patient Q and the corresponding scanogramms of the patients who are suffering from breast cancer and fibroadenoma of the mammary gland, respectively.

These proximity measures permit us to obtain algorithms and test for recognition of the differential diagnosis for breast cancer (the main hypothesis H) and fibroadenomatosis (the alternative hypothesis H'). The test of the acceptance of the hypothesis H has the form: if the training samples of patients suffering from breast cancer and fibroadenomatosis have equal size (approximately equal to 25 patients) then hypothesis H is accepted if $h_3 > h_4$ or $\alpha_1 \leq \alpha_2$. Investigations of the scanograms of 104 patients suffering from breast cancer and fibroadenomatosis show that the probability of the first kind of error is approximately 0.06, the probability of non-acceptance of decision is approximately 0.06 provided the hypothesis H is true, and the probability of the second kind of error is practically zero. Details of the statistical tests of the proposed method will be described in the Part 2.

References

[1] Kendall M.G. and Stuart A. (1958) *The advanced theory of statistics.* Vol.1. Distribution theory.- 2nd ed., Hafner., New York

[2] Vysochanskij D.F. and Petunin Yu.I. (1980) Justification of the 3 rule for unimodal distribution. *Theor. Probability and Mathem. Statistics.* 21. - P.25-36 (In Russian).

[3] Petunin Yu.I. (1981) *Use of the Theory of Random Processes in Biology and Medicine.*- Naukowa Dumka, Kiev. (in Russian).

[4] Pukelsheim F. (1994) The three sigma rule. *Amer. Statist.* 48. -P.88-91.

[5] Sellke Th. (1994) *Generalized Gauss-Chebyshev inequalities for unimodal random variables.* Technical Report #94-17. Department of Statistics. Purdue University. P.1-11.

[6] Dharmadhikari S. and Joag-dev K. (1988) *Unimodality, Convexity, and Applications.* - New York: Academic Press.

[7] Madreimov I. and Petunin Yu.I. (1982) Characterization of the uniform distribution with the help of the order statistics. *Theory Probab. and Math. Statist.* 27. - P.96-102.

[8] Bairamov I. and Petunin Yu.I. (1991) Structure of the invariant confidence intervals containing the underlying main distributed mass. *Theor. Probability and its Applications.* 35, N 1. - P.15-26.

[9] Van der Waerden B.L. (1969) *Mathematische Statistic.* 2nd ed. Springer-Verlag, Berlin and New York.

[10] Bairamov I. and Petunin Yu.I. (1991) Statistical tests based on training samples. *Cybernetics.* N 3. - P.408-413.

[11] Petunin Yu.I., Matveychuk S.A. (1994) Test of hypothesis with the help of the statistical criteria using a procedure of the non-acceptance decision. *Proc. of Acad. Sci.* 336, N 3. - P.301-303 (In Russian).

[12] Petunin Yu.I., Timoshenko Ja.G. and Petunina M.Yu. (1984) Test for the identification of the general population for the finite class of the alternative hypothesises. *Proceeding of Ukrainian Academy of Sciences*, Ser. "A". N 6. - pp.29-32 (in Russian).

[13] Matveychuk S.A. and Petunin Yu.I. (1990) A generalization of the Bernoulli model arising in variation statistics. I. *Ukr. Math. J.* 42, N 4. - P.518-528 (In Russian).

[14] Matveychuk S.A. and Petunin Yu.I. (1991) A generalization of the Bernoulli model arising in variation statistics. II. *Ukr. Math. J.* 43, N 6. - P.779-785 (In Russian).

[15] Johnson N.L. and Kotz S. (1991) Some generalizations of Bernoulli and Polya-Eggenberger contagion models. *Statist. Paper.* 32. - P.1-17.

[16] Johnson N.L., Kotz S. (1992) Further comments on Matveychuk and Petinin's generalized Bernoulli model, and nonparametric tests of homogeneity (pre-print).

[17] Petunin Yu.I., Kljushin D.A., Andrushkiw R.I., Ganina K.P., Boroday N.V. (1995) Computer diagnosis of the breast cancer // Abstracts of the 14th IMACS World Congress on Computational and Applied Mathematics (Atlanta, USA, July 11-15)- P.879-881.

[18] Ganina K.P., Boroday N.V., Petunin Yu.I., Klyushin D.A. (1998) Quantitative estimations of malignancy-associated changes in buccal epithelium under breast cancer and fibroadenomatosis // Experimental Oncology. 20, N 2. - P.130-134 (In Russian).

[19] Wein B. et al. (1998) Automated breast tumor diagnosis and grading based on wavelet chromatin texture description // Cytometry. 33. - P.32-40.

[20] Petunin Yu.I., Rublev B.V. (1996) Pattern recognition with the help of quadratic discriminant functions // Computational and Applied Mathematics. TViMS, Kiev. N 80. - P. 90-105 (In Russian).

[21] Petunin Yu.I., Kljushin D.A., Andrushkiw R.I. (1997) Nonlinear algorithm of pattern recognition for computer-aided diagnosis of breast cancer // Nonlinear Analisys, Theory, Methods & Applications. Elsevier Science Ltd. Vol.30. - No. 8. - P.5431-5436.

ÖZET

Göğüs kanseri teşhisi (CMG) ve FAM için bilgisayara dayalı yöntemlerin matematiksel yönleri incelenmiştir. Karesel, doğrusal ve sıra istatistikleri testlerinin eşanlı uygulamaları hakkında bir hipotez testi yöntemi geliştirilmiştir. Güven aralıkları ve güven elipslerinin kurulması için algoritmalar ele alınmıştır.

İstatistik, Journal of the Turkish Statistical Association
1999, Vol.2, No 2, pp. 87-105

COMPUTER-AIDED DIAGNOSIS OF BREAST CANCER PART 2. TESTS AND EXPERIMENTS

Yu.I.Petunin, D.A.Klyushin
Department of Cybernetics, Kiev Shevchenko University
Kiev, Vladimirskaya, 64, DSP, 252601, Ukraine
e-mail: vm214@dcp.kiev.ua

K.P.Ganina, N.V.Boroday
R.E.Kavetsky Institute for Experimental Pathology,
Oncology and Radiobiology,
National Academy of Sciences of Ukraine
Kiev,Vasilkovskaya, 45, 252022, Ukraine

R.I.Andrushkiw
Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102, USA

Abstract

A computer method of the diagnosis of breast cancer (CMG) and fibroadenomatosis (FAM) is developed, based on a single analysis of patient's buccal scrapes. The probability of error in the diagnosis of CMG and the probability of non-acceptance of decision do not exceed 6%. For FAM the probability of error in the diagnosis is practically zero, and the probability of acceptance of decision is 43%. The computer method of diagnosis is supplementary and can be applied to mass screening of patients for early of breast cancer and to the selection of patients who are in a high risk category.

Key Words: Breast cancer, malignancy-associated changed, buccal scrapes, confidence limits, discriminant analysis, proximity measure, training samples.

1. Biomedical background

It is well known that an organism has a potential capability to respond to changes of the external/internal media, which act on the organism in the whole or on some of it's internal system (nervous, endocrine, immune, etc.). This capability is called the reactivity of the organism [1]. The reactivity is generally different for different types of systems, organs, or cells of the organism. By biological reactivity we shall mean the changes arising in the cells in response to diverse factors. The problem of cytological reactivity during the formation of a malignant or benign tumor in the organism is rather complicated and many questions remain unanswered despite recent attempts to study the problem. The buccal epithelium is an indicator of common somatic diseases. There are many papers devoted to studing of morphological and functional states of the epithelium, however, in these investigations the subjective factor is predominant, while the quantitative methods can produce more objective estimation of the epithelium state. The DNA content in the nuclei of the cells is such index. H.Nieburgs, Ogden G.R., Cowpe, Green M.W. et al. (see detail bibliography in [1]). Investigated characteristic changes in the cells of the buccal epithelium of the patient with tumors localized out of the oral cavity and they named these changes as malignancy associated changes (MAC). These changes are characterized by increasing of the epitheliocytes nuclei size, discontinuous nuclear membrane, increasing of the zone sizes of the connected chromatin surrounded by the light areas. We investigate MAC as a basis of the computer-aided cytogenetical disgnosis of the breast cancer in women.

2. Material and method

For purposes of our investigation we considered women patients from 25 to 53 years old, who were suffering from breast cancer (second and third stage) and fibroadenomatosis (altogether 103 patients) were taken. Scrapes from various depths of the spinous layer were obtained (conventionally they were denoted as median and deep), after gargling and removing the superficial cell layer of buccal mucous. The smears were dried out under room temperature and fixed during 30 min in Nikiforov mixture. Then, the Feulgen reaction was made with cold hydrolysis in 5 n. HCl for 15 min, under the temperature t=21-22 ^{0}C. Optical density of the nuclei was registrated by a cytospectrophotometer using the scanning method with wave length 575 nm and probe diameter 0.05 mcm. We investigated from 10 to 30 nuclei in each preparation. The DNA- fuchsine content in the nuclei of the epitheliocytes was defined as a product of the optical density on area. Based on the investigation of the interphase nucleus we obtained a scanogram of the DNA distribution which is represented by a rectangular matrix $R = \|r_{ij}\|_{i=\overline{1,m}}^{j=\overline{1,n}}$, where r_{ij} characterizes the DNA content in the

grid cell with index (i, j), m and n are the rows and columns of the matrix R. In the formation of the training samples to validate their statistical homogeneity this size varied in sufficiently narrow limits from 56 to 81.

3. Morpho- and densitometric indecies of nuclei

The scanogram, or a numerical portrait of a cell, is given by the matrix $R = \|r_{ij}\|_{i=\overline{1,m}}^{j=\overline{1,n}}$, where r_{ij} are values of pointwise optical density of chromatin in interphase nuclei of the buccal cell, expressed in terms of conventional unit of measure; n and m are numbers of points in the scanogram along vertical and horizontal lines, respectively.

On the basis of these cytophotospectrometric indecies we calculate the following morpho- and densitometric indecies that characterize structural and textural peculiarities of chromatin [2-5].

1. Area of nuclei x_1 is a number of elements of R where $r_{ij} \geq 0.08$.
2. Area of condensed chromatin x_2 is a number of elements of R where $r_{ij} \geq 0.35$.
3. Area of decondensed chromatin x_3 is a number of elements of R where $0.08 \leq r_{ij} < 0.35$.
4. Area of strongly decondenced chromatin x_4 is a number of elements of R where $0.08 \leq r_{ij} < 0.15$.
5. Specific area of condensed chromatin:

$$x_5 = \frac{x_2}{x_1}$$

6. Specific area of decondensed chromatin:

$$x_6 = \frac{x_3}{x_1}$$

7. Integral density:

$$x_7 = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij},$$

where the summation is taken over indecies i and j for which $r_{ij} \geq 0.08$.

8. Mean density:

$$x_8 = \frac{x_1}{nm - p},$$

where p is a number of the elements $r_{ij} < 0.08$.

9. Averaged sum of overfalls:

$$x_9 = \frac{1}{q} \sum_{k=1}^{q} v_k,$$

where q is the number of the elements such that
$\min(r_{ij}, r_{i+1j}, r_{ij+1}, r_{i+1j+1}) \geq 0.08$;
$v_k = \max(r_{ij}, r_{i+1j}, r_{ij+1}, r_{i+1j+1}) - \min(r_{ij}, r_{i+1j}, r_{ij+1}, r_{i+1j+1})$, $k = \overline{1,q}$.
(The summation is taken over elements mentioned above).

10. General cluster index:

$$x_{10} = \frac{1}{q} \sum_{k=1}^{q} v_k^2.$$

11. Dispersion coefficient:

$$x_{11} = \sqrt{\left(\frac{1}{q-1} \sum_{k=1}^{q} (v_k - x_9)^2 \right)}.$$

12. Index of overfall variation:

$$x_{12} = x_9 + x_{11}.$$

13. Relief index:

$$x_{13} = \frac{\sum_{i=2}^{m} \sum_{j=1}^{n} |r_{ij} - r_{i-1j}|}{(2mn - m + n - q)}$$

where q is a number of the points (i,j)such that $\max(r_{ij}, r_{i-1j}) < 0.08$.
14. Textural coefficient:

$$x_{14} = \frac{x_{13}}{\varepsilon}, \; \varepsilon = \frac{1}{mn - p} \sum_{i=1}^{m} \sum_{j=1}^{n} (r_{ij} - x_7),$$

where p is defined as for x_8 and the summation is taken over indecies i and j for which $r_{ij} \geq 0.08$.
15. Coefficient of mutual disposition:

$$x_{15} = \frac{a}{bx_8^2},$$

where

$$a = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\sum_{k=1}^{m}\sum_{l=j+1}^{n}\frac{r_{ij}r_{kl}}{(k-i)^2+(l-j)^2}+\sum_{k=i+1}^{m}\sum_{l=1}^{n}\frac{r_{ij}r_{kl}}{(k-i)^2+(l-j)^2}\right),$$

$$b = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\sum_{k=1}^{m}\sum_{l=j+1}^{n}\frac{1}{(k-i)^2+(l-j)^2}+\sum_{k=i+1}^{m}\sum_{l=1}^{n}\frac{1}{(k-i)^2+(l-j)^2}\right),$$

moreover, the summation both for a and for b is taken over elements such that $\min(r_{ij}, r_{kl}) > 0.875 \max\limits_{i=1,2,...,n;j=1,2,...,m;r_{ij}\geq 0.08} \min r_{ij}$.

4. Calibration of training samples

The proposed diagnosis of oncological diseases is based on the tests and algorithms of the statistical and geometric theory of pattern recognition.

On the first stage we form two groups of patient's scanogram $A = \{X_i\}$, $i = 1, 2, ..., N$ and $B = \{Y_j\}$, $j = 1, 2, ..., M$ whose diagnosis must be verified exactly. Below, for definiteness, we shall suppose that the group A (or B) contains the scanograms of the patients suffering from the cancer of mammary gland - CMG (or the fibroadenomatosis - FAM). After the procedures of the registration and measurement of the morpho - and densitometric indecies, we obtain so-called training samples for every index x_k ($k = 1, 2, ..., 15$): $G_A^{(1)}, G_A^{(2)}, ..., G_A^{(15)}$ for the patients of the group A (CMG-samples) and $G_B^{(1)}, G_B^{(2)}, ..., G_B^{(15)}$ for the patients of the group B (FAM-samples).

Consider the problem of determining what should be the number of training samples in the groups A and B to insure sufficiently high level of reliability of the diagnosis.

At the beginning it is naturally to suppose that the number of samples in the groups A and B must be equal. To accept or reject this hypothesis we have proposed a procedure of calibration of training samples, which consists of the following stages.

1. Exclude patient X_i, $i = 1, 2, ..., N$ (or Y_j, $j = 1, 2, ..., M$) from the set $A \cup B$.
2. On the basis of the set of samples $\{A \cup B\} \setminus X_i$ (or $\{A \cup B\} \setminus Y_j$) construct the tests using pairs of ellipses $\left(E_{ts}, \overline{E}_{ts}\right)$, $\left(E_{ts}^*, \overline{E}_{ts}^*\right)$ and half-planes (π_{ts}, λ_{ts}), $(\pi_{ts}^*, \lambda_{ts}^*)$.
3. Calculate statistics for patient X_i, $i = 1, 2, ..., N$ (or Y_j, $j = 1, 2, ..., M$).
4. Return patient X_i, $i = 1, 2, ..., N$ (or Y_j, $j = 1, 2, ..., M$) in the set $A \cup B$ and repeat this procedure for the next patient.

The results of calibration in the case when the set A consists of 25 scanograms of patients suffering from CMG (so-called CMG-patients) and the set B consists of 25 scanograms of patients suffering from FAM (FAM-patients), are given in Tables 1 and 2.

Now let us consider the following criteria of diagnostics

1) quadratic: $h_3 > h_4 \Rightarrow$ CMG; $h_3 \leq h_4 \Rightarrow$FAM;

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)
1	0.00952	0.00476	0.98571	0.98095	0.38095	0.61905
2	0.05238	0.17143	0.75238	0.87143	0.22857	0.77143
3	0.02857	0.00000	0.99524	0.96667	0.74762	0.25238
4	0.00952	0.33810	0.65714	0.98571	0.44286	0.55714
5	0.03333	0.18095	0.81429	0.96190	0.23810	0.76190
6	0.00952	0.04762	0.93333	0.97143	0.42857	0.57143
7	0.02857	0.00476	0.99524	0.97143	0.70000	0.30000
8	0.04762	0.13810	0.82857	0.91905	0.60476	0.39524
9	0.01429	0.08095	0.91905	0.98571	0.69524	0.30476
10	0.04762	0.07619	0.90476	0.93333	0.78571	0.21429
11	0.06667	0.10476	0.87619	0.91429	0.64286	0.35714
12	0.08095	0.09524	0.86667	0.88095	0.70952	0.29048
13	0.02381	0.01905	0.96190	0.95714	0.34286	0.65714
14	0.00000	0.01905	0.98095	1.00000	0.86190	0.13810
15	0.08571	0.23333	0.70476	0.85238	0.72857	0.27143
16	0.07143	0.07143	0.81905	0.81905	0.66190	0.33810
17	0.00476	0.00000	1.00000	0.99524	0.49048	0.50952
18	0.00476	0.03333	0.96190	0.99048	0.34762	0.65238
19	0.02381	0.01905	0.98095	0.97619	0.83333	0.16667
20	0.00000	0.00000	1.00000	1.00000	0.42381	0.57619
21	0.00476	0.03810	0.95238	0.98571	0.54762	0.45238
22	0.10476	0.04286	0.93333	0.87143	0.71905	0.28095
23	0.04762	0.25714	0.61905	0.82857	0.40952	0.59048
24	0.01905	0.08571	0.88571	0.95238	0.58095	0.41905
25	0.00000	0.01905	0.98095	1.00000	0.40952	0.59048

Table 1: Values of the statistics $h_k = h(C_k)$, $k = 1, 2, ..., 6$ for the CMG-patient's scanograms under calibration of the training samples (24 CMG and 25 FAM)

2) linear: $h_5 > h_6 \Rightarrow$CMG; $h_5 \leq h_6 \Rightarrow$FAM.

Denote by D_1 the diagnosis of "CMG" and by D_2 the diagnose of "FAM". Let ν_{11} be the frequency of the event D_1 for the CMG-samples, $\nu_{21}$the frequency of the event D_2 for the CMG-samples, ν_{12} the frequency of D_1 for the FAM-samples, and ν_{22} the frequency of D_2 for the FAM-samples.

Analysis of the results of calibration of the samples from groups A and B of equal size (Table 3) allow us to make the following inference:

1. In overwhelming majority of cases we observe the predominance of the statistics h_4 (total FAM) over h_3 (total CMG), and statistics h_2 (FAM) over h_1 (CMG) (we shall call this phenomenon *the effect of stable predominance*). However, for the group A we do not detect this effect.

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)
1	0.00000	0.03810	0.96190	1.00000	0.25238	0.74762
2	0.03333	0.35238	0.56667	0.88571	0.33810	0.66190
3	0.00000	0.31429	0.68095	0.99524	0.21429	0.78571
4	0.03333	0.08095	0.89524	0.94286	0.83333	0.16667
5	0.00952	0.01905	0.98095	0.99048	0.50952	0.49048
6	0.00476	0.02381	0.95714	0.97619	0.70476	0.29524
7	0.00000	0.03810	0.96190	1.00000	0.25714	0.74286
8	0.03810	0.20952	0.75238	0.92381	0.22857	0.77143
9	0.04286	0.24762	0.56190	0.76667	0.30476	0.69524
10	0.05238	0.19048	0.43333	0.57143	0.20476	0.79524
11	0.04286	0.03810	0.90952	0.90476	0.59048	0.40952
12	0.03810	0.05714	0.91905	0.93810	0.80952	0.19048
13	0.00476	0.05238	0.94762	0.99524	0.86667	0.13333
14	0.10000	0.10952	0.87143	0.88095	0.58095	0.41905
15	0.01905	0.21429	0.78571	0.98095	0.42857	0.57143
16	0.04286	0.00952	0.98571	0.95238	0.68095	0.31905
17	0.06190	0.00952	0.97143	0.91905	0.79524	0.20476
18	0.01429	0.03333	0.96667	0.98571	0.91905	0.08095
19	0.01429	0.02381	0.96667	0.97619	0.53333	0.46667
20	0.07143	0.13333	0.70952	0.77143	0.45238	0.54762
21	0.03810	0.00952	0.96667	0.93810	0.63810	0.36190
22	0.04286	0.10000	0.88571	0.94286	0.20476	0.79524
23	0.16190	0.14762	0.60000	0.58571	0.92857	0.07143
24	0.05238	0.29048	0.42381	0.66190	0.27143	0.72857
25	0.00952	0.11429	0.88571	0.99048	0.24762	0.75238

Table 2: Values of the statistics for the FAM-patient's scanograms under calibration of the training samples (24 CMG and 25 FAM)

Criteria	ν_{11}	ν_{21}	ν_{22}	ν_{12}
Quadratic	0.28	0.72	0.80	0.20
Linear	0.56	0.44	0.48	0.52
Combined	0.72	0.28	0.80	0.20

Table 3: Frequencies of the random events D_k, $k = 1, 2, ..., 6$ under calibration of the training samples (24 CMG and 25 FAM)

2. In the case of the linear criterion, the events D_1 and D_2 are nearly equiprobable both for group A (training samples of the CMG-patients) and for group B (training samples of the FAM-patients). Therefore, this criterion is unfit for the differential diagnostics of CMG from FAM.

3. The quadratic criterion for group B gives much better results, i.e., in 80% of the cases we obtain correct diagnosis (event D_2 occurs) and in 20% of the cases the diagnosis is incorrect (event occurs). However, for group A the results are reversed, i.e., in 28% of the cases we obtain correct diagnosis and in 72% incorrect diagnosis. Therefore, this criterion is also unfit for the differential diagnostics of CMG from FAM.

Since using groups of training samples A and B of equal size with only linear or quadratic criteria did not produce acceptable results, we calibrated the training samples for the case when the group A (25 scanograms of the CMG-patients) was approximately twice as large group B (12 scanograms of the FAM-patients). This selection of sizes had to provide predominance of the statistics h_3 (total CMG) over statistics h_4 (total FAM) and also h_1 (CMG) over h_2 (FAM). The results of calibration of these samples are shown in the Tables 4-6.

Based on the analysis of these results we can conclude that:

1. In the overwhelming majority of cases for group A we observe the predominance of the statistics h_3 (total CMG) over h_4 (total FAM) and also h_1 (CMG) over h_2 (FAM), i.e. the effect of stable predominance occurs. For the group B this effect does not occur.

2. For the linear criterion the events D_1 and D_2 are practically equiprobable therefore it is unfit for the differential diagnostics of the CMG from the FAM.

3. The quadratic criterion for the group A provides good results (in 90% of the cases we obtain the correct diagnosis, i.e. the event D_1 appears and in 8% of the cases the incorrect diagnosis is detected, i.e. the event D_2 occurs). However, for the group B we determine the correct diagnosis in 59% of the cases, and in 44% of the cases the computer diagnosis is not correct. Therefore, this criterion is also unfit for the differential diagnostics.

It should be noted that the effect of stable predominance of the statistics h_3 over h_4 for the group A, mentioned above, is observed only where the areas of the scanogram registration field vary in a rather narrow range (in the above case - from 56 to 81). If this principle is violated, then the statistically non-homogeneous sample is formed and the effect of stable predominating becomes slightly marked.

In summary, we must established that for the samples A and B such that size ratio is 2:1 (more exactly, such that $\left[\frac{1}{2} card\, A\right] \approx card\, B$, where $[x]$ denotes the integer part of the number x), the use of both quadratic and linear criteria alone does not permit to obtain acceptable results. Nevertheless, the above mentioned effect of stable

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)
1	0.16667	0.01905	0.97143	0.82381	0.42381	0.57619
2	0.14286	0.08095	0.81429	0.75238	0.27619	0.72381
3	0.07143	0.00000	0.99524	0.92381	0.80952	0.19048
4	0.12857	0.25714	0.69524	0.82381	0.41905	0.58095
5	0.23810	0.11905	0.77143	0.65238	0.26190	0.73810
6	0.05714	0.01905	0.93810	0.90000	0.55238	0.44762
7	0.34286	0.00952	0.90095	0.64762	0.80952	0.19048
8	0.20476	0.04286	0.82381	0.66190	0.60952	0.39048
9	0.18095	0.04762	0.91429	0.78095	0.85238	0.14762
10	0.26190	0.06190	0.89048	0.69048	0.83333	0.16667
11	0.27143	0.02381	0.85238	0.60476	0.68095	0.31905
12	0.35714	0.04286	0.88571	0.57143	0.75714	0.24286
13	0.23333	0.01905	0.95714	0.74286	0.33810	0.66190
14	0.29524	0.00000	0.97619	0.68095	0.91429	0.08571
15	0.32381	0.07619	0.70952	0.46190	0.79524	0.20476
16	0.21905	0.00476	0.80952	0.59524	0.65238	0.34762
17	0.15238	0.00000	1.00000	0.84762	0.51905	0.48095
18	0.11905	0.00476	0.98095	0.86667	0.40000	0.60000
19	0.13810	0.00476	0.99048	0.85714	0.83810	0.16190
20	0.00952	0.00000	1.00000	0.99048	0.66667	0.33333
21	0.16667	0.01905	0.97143	0.82381	0.61905	0.38095
22	0.17143	0.03333	0.92381	0.78571	0.71905	0.28095
23	0.15714	0.22381	0.67619	0.74286	0.38095	0.61905
24	0.20000	0.04286	0.89048	0.73333	0.60000	0.40000
25	0.08571	0.00952	0.97143	0.89524	0.41905	0.58095

Table 4: Values of the statistics $h_k = h(C_k)$, $k = 1, 2, ..., 6$ for the CMG-patient's scanograms under calibration of the training samples (25 CMG and 12 FAM)

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)
1	0.04762	0.06190	0.92857	0.94286	0.34286	0.65714
2	0.10000	0.25714	0.58571	0.74286	0.31429	0.68571
3	0.02857	0.29048	0.69048	0.95238	0.21429	0.78571
4	0.20476	0.02857	0.91429	0.73810	0.86667	0.13333
5	0.05238	0.01905	0.97143	0.93810	0.71429	0.28571
6	0.11905	0.00952	0.93333	0.82381	0.80476	0.13333
7	0.05238	0.07619	0.91905	0.94286	0.34762	0.65238
8	0.09524	0.20952	0.72857	0.84286	0.27143	0.72857
9	0.14286	0.04286	0.50000	0.40000	0.37143	0.62857
10	0.09524	0.04286	0.43810	0.38571	0.13810	0.86190
11	0.24286	0.01905	0.92857	0.70476	0.55238	0.44762
12	0.21905	0.01429	0.95714	0.75238	0.85714	0.14286

Table 5: Values of the statistics $h_k = h(C_k)$, $k = 1, 2, ..., 6$ for the CMG-patient's scanograms under calibration of the training samples (25 CMG and 12 FAM)

Criteria	ν_{11}	ν_{21}	ν_{22}	ν_{12}
Quadratic	0.92	0.08	0.42	0.56
Linear	0.68	0.32	0.58	0.42
Combined	0.92	0.08	0.08	0.42

Table 6: Frequencies of the random events D_k, $k = 1, 2, ..., 6$ under calibration of the training samples (25 CMG and 12 FAM)

predominance that is observed for training samples of equal size ($card\ A \approx card\ B$), and for training samples such that $\left[\frac{1}{2} card A\right] \approx card\ B$, allows us to formulate the so-called filtering criterion. The principal idea of this criterion consists of the following.

Consider first the calibration results for the training samples of equal size. As was shown in this case, for the group B (FAM-patients) we have the effect of stable predominance of the statistics h_4 (total FAM) over h_3 (total CMG), whereas for the group A (CMG-patients) the effect is missing. Let Q be a patient whose disease we try to diagnose. On the basis of the groups (training samples) A and B, we compute the values of the statistics $h_3(Q)$ and $h_4(Q)$ for this patient. Suppose that $h_3(Q) \geq h_4(Q)$. Which hypothesis (D_1 or D_2) is in better agreement with experimental results? As far as the effect of stable predominance $h_3(Q) < h_4(Q)$ for the group B is observed, the probability of the event "patient Q is suffering from the same disease as patients in the group B (i.e. FAM)" will be small. Hence, it is more probable that this patient is suffering from CMG. So, the hypothesis D_2 will be in better accord with the experimental results. But if for this patient $h_3(Q) < h_4(Q)$, then we cannot accept any decision about the patient's disease on the basis of the group A and B of equal size (i.e. we cannot diagnose the disease), because such data may be inherent both in the CMG-patients and the FAM-patients.

Now, let us define the so-called A-filter, which can be used to diagnose CMG for examined patients in some cases. We shall say that a scanogram passes through the -filter if and fails to pass if $h_3(Q) < h_4(Q)$.

Next, let us describe the B-filter. To this end we use another pair of training samples A and B of size $\left[\frac{1}{2} card A\right] \approx card\ B$, i.e. the size of the A twice as large as that of B. Calibration results in this case show that the reverse effect of stable predominance $h_3(Q) > h_4(Q)$ for the group A is achieved. Hence, values of the statistics $h_3^*(Q)$ and $h_4^*(Q)$ satisfying the inequality $h_3^*(Q) \leq h_4^*(Q)$, obtained for the examined patient with the help of this pair of the training samples A and B, indicate higher probability for the diagnosis of FAM than CMG. Finally, if $h_3^*(Q) > h_4^*(Q)$ than we cannot diagnose the disease (non-acceptance decision). Thus, we have described the second part of the filtering criterion (so-called B-filter), which allows us in some cases to diagnose FAM in the patients. We shall say that the scanograms of the examined patient pass through the B-filter if $h_3^*(Q) \leq h_4^*(Q)$ and faile to do so if $h_3^*(Q) > h_4^*(Q)$.

In view of the above, the quadratic filtering criterion may be described as follows:

1. Form two pairs of the training samples A and B with the sizes $card\ A \approx card\ B$ and $\left[\frac{1}{2} card A\right] \approx card\ B$, respectively. The first pair is used in the construction of the A-filter, the second one in the construction of the B-filter.

2. Perform the above-mentioned process of filtration of the patient's scanogram through the A-filter and B-filter. If the scanograms pass through the A-filter, than the

diagnosis of CMG ais indicated. If they pass through B-filter, then FAM is indicated. Otherwise, if neither filter is passed, the diagnosis is not made (non-acceptance of decision).

The quadratic filtering criterion is interesting, but it cannot be considered acceptable for clinical medicine, since the probability of the non-acceptance of decision is too high. This brings us to idea of using a combined filtering criteria (quadratic and order), which can be described as follows.

The combined filtering criterion also consists of two filters: A-filter and B-filter. Let h_i, $i = 1, 2, ..., 6$ and α_1, α_2 be the statistics of the examined patient Q obtained with the help of the training samples A and B of equal size ($card\ A \approx card\ B$) and denote by h_i^*, $i = 1, 2, ..., 6$ and α_1^*, α_2^* the corresponding statistics obtained with the help of the training samples A and B, such that $cardA = 25$, $cardB = 12$. We shall say that the scanograms of the patient Q pass through the A-filter if at least one of the following inequalities is true:

$$h_3 > h_4;\ \ \alpha_1 \leq \alpha_2$$

(the proposition $h_3 > h_4 \,\&\, \alpha_1 \leq \alpha_2$ is true), and that they pass through the B-filter if at least one of the following inequalities is true:

$$h_3^* \leq h_4^*;\ \ \alpha_1^* > \alpha_2^*$$

(the proposition $h_3^* \leq h_4^* \,\&\, \alpha_1^* > \alpha_2^*$ is true).

The combined filtering criterion is formed in the following way: if the scanograms of the examined patient pass through the A-filter, then the diagnosis is CMG. If they pass through the B-filter, then the diagnosis is FAM. Otherwise, we cannot diagnose disease (non-acceptance of decision).

For the experimental testing of the quality of the proposed criteria we have selected 17 CMG-patients and 7 FAM-patients. All these patients did not belong neither to the A-group or the B-group. The A-filter have been constructed on the basis of the 24 CMG-scanograms (the group A) and 25 FAM-scanograms (the group B), and B-filter have been constructed on the basis of the 25 CMG-scanograms (the group A) and 12 FAM-scanograms (the group B). The results of testing for both the filters are shown in Tables 7-11.

Analysis of the experimental results show that in the case of combined filtering criterion we can have three possible decisions: 1) to diagnose FAM in the examined patient; 2) to diagnose CMG and 3) fail to diagnose any disease (non-acceptance of decision). If we obtaine a diagnosis FAM for a patient who is suffering from CMG, then this produces a so-called error of the first kind. If a diagnosis of CMG is obtained for a patient suffering from FAM, then this produces an error of the second kind. On the basis of the experimental results we can conclude the following statements (see Tables 7 and 8): the probability of the first kind of error is approximately 6% and the

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)	α_1
1	0.01905	0.01429	0.54286	0.54286	0.42381	0.45714	5
2	0.00000	0.01429	1.00000	0.64762	0.27619	0.35238	1
3	0.02857	0.07143	0.71429	0.40000	0.80952	0.60000	6
4	0.02857	0.14762	0.72381	0.27143	0.41905	0.72857	6
5	0.06667	0.20000	0.52857	0.23333	0.26190	0.76667	13
6	0.01905	0.10952	0.91429	0.32857	0.55238	0.67143	7
7	0.02381	0.06190	0.68571	0.66667	0.80952	0.33333	7
8	0.02857	0.01905	0.97143	0.77143	0.60952	0.22857	1
9	0.06190	0.22381	0.47619	0.24762	0.85238	0.75238	14
10	0.02381	0.00476	9.97143	0.44762	0.83333	0.55238	0
11	0.07143	0.01905	0.83333	0.31905	0.68095	0.68095	2
12	0.00952	0.25238	0.61429	0.34286	0.75714	0.65714	13
13	0.03810	0.09048	0.77619	0.80000	0.33810	0.20000	5
14	0.01905	0.12857	0.65714	0.51429	0.91429	0.48571	9
15	0.07619	0.05238	0.60952	0.41905	0.79524	0.58095	10
16	0.00952	0.12381	0.80000	0.25238	0.65238	0.74762	8
17	0.02857	0.11905	0.69524	0.24286	0.51905	0.75714	11

Table 7: Values of the statistics $h_k = h(C_k)$, $k = 1, 2, ..., 6$ for the CMG-patient's scanograms under testing by A-filter (24 CMG and 25 FAM)

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)	α_1
1	0.00000	0.04762	0.95238	1.00000	0.59524	0.40476	1
2	0.03333	0.37619	0.28571	0.62857	0.17143	0.82857	19
3	0.01905	0.21905	0.56667	0.76667	0.26667	0.73333	16
4	0.00476	0.08571	0.91429	0.99524	0.17143	0.82857	11
5	0.00000	0.39524	0.60476	1.00000	0.88571	0.11429	1
6	0.00000	0.53810	0.45714	0.99524	0.16667	0.83333	5
7	0.00000	0.10952	0.89048	1.00000	0.18571	0.81429	3

Table 8: Values of the statistics $h_k = h(C_k)$, $k = 1, 2, ..., 6$ for the FAM-patient's scanograms under testing by A-filter (24 CMG and 25 FAM)

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)	α_1
1	0.18095	0.00000	0.80952	0.62857	0.35238	0.44762	5
2	0.08095	0.01429	0.98571	0.91905	0.71429	0.28571	1
3	0.17143	0.04286	0.66190	0.53333	0.40476	0.59524	6
4	0.12381	0.10476	0.60000	0.58095	0.30476	0.69524	6
5	0.21905	0.12381	0.41905	0.32381	0.28095	0.71905	13
6	0.20952	0.10952	0.79524	0.69524	0.34286	0.65714	7
7	0.17143	0.04762	0.60952	0.48571	0.67619	0.32381	7
8	0.20476	0.00000	0.97143	0.76667	0.87143	0.12857	1
9	0.17143	0.10476	0.35714	0.29048	0.27619	0.72381	14
10	0.14286	0.00000	0.98571	0.84286	0.50000	0.50000	0
11	0.20952	0.00476	0.89048	0.68571	0.35238	0.64762	2
12	0.12857	0.16667	0.36190	0.40000	0.37143	0.62857	13
13	0.28095	0.02857	0.72381	0.47143	0.77143	0.22857	5
14	0.10476	0.05714	0.52381	0.47619	0.55238	0.44762	9
15	0.19048	0.01905	0.60476	0.43333	0.46667	0.53333	10
16	0.19524	0.07143	0.70476	0.58095	0.30000	0.70000	8
17	0.23333	0.07619	0.63810	0.48095	0.29048	0.70952	11

Table 9: Values of the statistics $h_k = h(C_k)$, $k = 1, 2, ..., 6$ for the CMG-patient's scanograms under testing by B-filter (24 CMG and 25 FAM)

N	h_1 (CMG)	h_2 (FAM)	h_3 (TCMG)	h_4 (TFAM)	h_5 (LCMG)	h_6 (LFAM)	α_1
1	0.10476	0.02381	0.95714	0.87619	0.69048	0.30952	1
2	0.11429	0.29524	0.32381	0.50476	0.22381	0.77619	19
3	0.17143	0.18571	0.56667	0.58095	0.29524	0.70476	16
4	0.06190	0.07143	0.86190	0.87143	0.18095	0.81905	11
5	0.35714	0.08571	0.59524	0.32381	0.93810	0.06190	1
6	0.09524	0.24286	0.40476	0.55238	0.20476	0.79524	5
7	0.20476	0.09048	0.86190	0.74762	0.29048	0.70952	3

Table 10: Values of the statistics $h_k = h(C_k)$, $k = 1, 2, ..., 6$ for the FAM-patient's scanograms under testing by B-filter (24 CMG and 25 FAM)

Criteria	ν_{11}	ν_{21}	ν_{22}	ν_{12}
Quadratic	0.29	0.06	0.22	0.11
Linear	0.35	0.06	0.67	0.11
Combined	0.94	0.06	0.43	0.00

Table 11: Frequencies of the random events D_k, $k = 1, 2, ..., 6$ under testing of the patient's scanograms by A- and B-filters

probability of the second kind of error is practically equal to 0%. This means that the probability (more exactly, frequency) of the FAM-diagnosis for the CMG-patients is approximately equal to 0.06 and probability of CMG-diagnosis for FAM-patients is equal to 0.00. In addition, the probability of diagnosing the disease (acceptance of decision) is equal to 94% for CMG-patients and 43% for FAM-patients. Thus, based on the analysis of the above process, we can diagnose cancer of the mammary gland with high probability after single analysis, however for more accurate determination of fibroadenomatosis we must repeat the process.

5. Repeated analysis

The repeated analysis is produced in the following cases:

1) Sequential analysis. If after first analysis the decision is not accepted we repeat taking scrape, construction of scanogram, calculation of its indices and test procedure.

2) Multiple analysis. To increase the accuracy of diagnosis (i.e., for decreasing of the probability of the false-positive and false-negative diagnosis) and to decrease the probability of non-acceptance of decision we take several scrapes at once, construct the corresponding svanograms and so on.

First, consider how we can to exclude completely the non-acceptance of decision and to produce definite diagnosis (CMG or FAM) by sequential analysis. Hereinafter we assume, that size n of the group of examined patients so large, that the frequency h and corresponding probability p of the event (correct diagnosis, misdiagnosis and non-acceptance of decision) are practically coincide. Suppose, that after the first analysis the probability (frequency) of misdiagnosis equals to α, the probability of non-acceptance of decision equals to β, and the probability of correct diagnosis is equal to γ, so that $\alpha + \beta + \gamma = 1$. Let n be a size of group of examined patients, i_k the number of misdiagnoses, m_k the number of non-acceptance of decision and l_k the number of correct diagnoses obtained under the k-th repeatition of the analysis. We suppose that under repeating of analysis the frequences α, β, γ are constant. Then $i_1 = \alpha n$, $m_1 = \beta n$, $l_1 = \gamma n$, $i_2 = \alpha\beta n$, $m_2 = \beta^2 n$, $l_2 = \gamma\beta n$, $i_3 = \alpha\beta^2 n$, $m_3 = \beta^3 n$, $l_3 = \gamma\beta^2 n$, ... , $i_k = \alpha\beta^{k-1} n$, $m_k = \beta^k n$, $l_k = \gamma\beta^{k-1} n$ as far as the repeated analyses are used only for the patients with indefinite diagnoses (non-acceptance of decision). After N repetitions of analysis the number of misdiagnoses $i\,(N)$ has the form

$$i\,(N) = \sum_{k=1}^{N} i_k = \alpha n \sum_{k=1}^{N} \beta^{k-1} = \alpha n \frac{1-\beta^{N-1}}{1-\beta} \approx \frac{\alpha n}{1-\beta},$$

since under large N the number β^{N-1} is negligible, and the number of non-acceptance of decision $m(N) = m_N = \beta^N n \approx 0$.

Thus, under the large N the probability of misdiagnosys α_N satisfies the inequality

$$\alpha_N = \frac{i(N)}{n} = \alpha \frac{1-\beta^{N-1}}{1-\beta} < \frac{\alpha}{1-\beta},$$

and the probability of non-acceptance of decision β_N is given by

$$\beta_N = \frac{m(n)}{n} = \beta^N,$$

so that, under $\beta < \frac{1}{2}$, the probability of misdiagnosys increases less than twice, and the probability of non-acceptance of decision is negligible (since β^N very quickly tends to zero, then under $N = 5$ as a rule $\beta^N \approx 0$). Hence, the probability of the correct diagnosis $\gamma_N = 1 - \alpha_N - \beta_N$.

For example, according the results obtained for the patients suffering from CMG we have $\alpha = 0.06$, $\beta = 0.0$, $\gamma = 0.94$ (see Table 11), so for the sequentional two-step analysis the theoretical estimations of the probabilities are invariant. But, for the patients suffering from FAM $\alpha = 0.0$, $\beta = 0.57$, $\gamma = 0.43$ (see Table 11), and

$$\begin{aligned}
&\alpha_2 = 0.0,\ \beta_2 = 0.3249,\ \gamma_2 = 0.6751,\\
&\alpha_3 = 0.0,\ \beta_3 = 0.1852,\ \gamma_3 = 0.8148,\\
&\alpha_4 = 0.0,\ \beta_4 = 0.1056,\ \gamma_2 = 0.8944,\\
&\alpha_5 = 0.0,\ \beta_6 = 0.0602,\ \gamma_2 = 0.9398,\\
&\alpha_2 = 0.0,\ \beta_2 = 0.0343,\ \gamma_2 = 0.9657.
\end{aligned}$$

An estimate for the number of repeatition of the analysis depending on the significance level β^* for the probability of the non-acceptance of decision, has the form:

$$N = \left[\frac{\ln \beta^*}{\ln \beta}\right] + 1.$$

Consider now the repeated analysis in the case of multiple analysis. It should be stressed, that in this case, in contrast to sequential analysis, the repeated analysis is obtained for all patients at once. Let A_1 denote a misdiagnosis under the first analysis ("the first" and "the second" refers only to the order number in series scrapes for multiple analysis), let A_2 denote a non-acceptance of decision (indefinite diagnosis), and A_3 be a correct diagnosis. In accordance with the above notations the probability of these events is: $p(A_1) = \alpha$, $p(A_2) = \beta$, $p(A_3) = \gamma$. As a result of multiple analysis can be occured 9 events: $A_i^{(1)} A_j^{(2)}$, $i, j = 1, 2, 3$ where the product of events means that they occure simultaneously, the upper indecies mean the order number of analysis.

Suppose, that patient P suffers from CMG. Then the event A_1means the misdiagnosis "FAM", and A_3 is the correct diagnosis "CMG". After series from two analysis the misdiagnosis "FAM" is made in the following cases: $A_1^{(1)} A_1^{(2)}$, $A_1^{(1)} A_2^{(1)}$, $A_2^{(1)} A_1^{(2)}$. That is why the probability of the misdiagnosis is equal to

$p_1 = p\left(A_1^{(1)}A_1^{(2)} + A_1^{(1)}A_2^{(1)} + A_2^{(1)}A_1^{(2)}\right) = p\left(A_1^{(1)}A_1^{(2)}\right)+p\left(A_1^{(1)}A_2^{(1)}\right)+p\left(A_2^{(1)}A_1^{(2)}\right) = \left(p\left(A_1^{(1)}\right)\right)^2 + 2p\left(A_1^{(1)}\right)p\left(A_2^{(1)}\right) = \alpha^2 + 2\alpha\beta,$

provided that the results of the repeated analyses are independent events (as above, we suppose that under repeating of analysis the frequences α, β, γ are constant).

The non-acceptance of decision arises in cases $A_2^{(1)}A_2^{(2)}$, $A_1^{(1)}A_3^{(2)}$, $A_3^{(1)}A_1^{(2)}$, and the probability of this event is given by

$p_2 = p\left(A_2^{(1)}A_2^{(2)} + A_1^{(3)}A_3^{(2)} + A_3^{(1)}A_1^{(2)}\right) = \left(p\left(A_2^{(1)}\right)\right)^2 + 2p\left(A_1^{(1)}\right)p\left(A_3^{(2)}\right) = \beta + 2\alpha\gamma.$

At least, the correct diagnosis "CMG" is made in the cases $A_3^{(1)}A_3^{(2)}$, $A_3^{(1)}A_2^{(2)}$, $A_2^{(1)}A_3^{(2)}$, and the probability p_3 of the correct diagnosis is:

$p_3 = \gamma^2 + 2\beta\gamma.$

If

$$\alpha + 2\beta < 1,\ 2\alpha\gamma < \beta\,(1-\beta)\,, \tag{1}$$

then $p_1 = \alpha^2 + 2\alpha\beta < \alpha$, $p_2 = \beta + 2\alpha\gamma < \beta$, $p_3 = \gamma^2 + 2\beta\gamma > \gamma$. So, the probabilities of the misdiagnosis and non-acceptance of decision decrease. Note, that nonequalities (1) are always true provided that $0 < \delta < \beta < \frac{1}{2}$ and α is near zero. The similar situation arises for the patients suffering from FAM.

Thus, according the above-mentioned results for the patients suffering from CMG we have $\alpha = 0.06$, $\beta = 0.0$, $\gamma = 0.94$ (see Table 11), so for double analysis the theoretical estimates of the probabilities are the follows:

$$p_1 = 0.0036,\ p_2 = 0.1128,\ p_3 = 0.8836.$$

For the patients suffering from FAM $\alpha = 0.0$, $\beta = 0.57$, $\gamma = 0.43$ (see Table 11), and

$$p_1 = 0.0,\ p_2 = 0.3249,\ p_3 = 0.6751.$$

Calculation of the changes of the probabilities α, β, and γ under multiple analysis is produced according to above mentioned scheme.

CONCLUSIONS

In summary, the above investigations have been shown that the proposed computer method for the diagnosis of breast cancer (CMG) and fibroadenomatosis (FAM) allows us to identify with high probability the diagnosis of breast cancer, based on a single analysis of patient's buccal scrapes (the probability of error in the diagnosis and the probability of non-acceptance of decision do not exceed 6%). In the case of

patients suffering from FAM, the probability of error in the diagnosis is practically zero, however the probability of non-acceptance decision based on a single analysis of buccal scrapes is 43%.

If decision is not accepted, we must repeat the analysis by taking more trials (buccal scrapes) If the results of the analysis are similar after n trials, then the probability of non-acceptance of decision is approximately equal to $\left(\frac{1}{2}\right)^n$ provided that the results were obtained independently (so-called independent trials). If it is known that the patient is suffering only from one disease, then the value $\left(\frac{1}{2}\right)^n$ quickly tends to zero and, as a rule after 5-6 trials (buccal scrapes), we can diagnose FAM.

The computer method of diagnosis is a supplementary method that can be used only in conjuction with other methods of clinical examination of patients (mammography, ultrasound examination, etc). The proposed method can be applied also to mass screening of patients for the early detection of breast cancer, or to the selection of patients who are in the high risk category.

References

[1] Ganina K.P., Polischuk L.Z., Boroday N.V., Naleskina L.A., Isakova L.M., Nesina I.P., Buchinskaya L.G. (1995) *Cytological reactivity of the oncological patient.* Naukova Dumka, Kiev. (In Russian).

[2] Kapantsyan A.L., Karalova E.M., Magakyan Yu.A. (1988) The coefficient of mutual location of particles and its use in the quantitative analysis of material in the cells. *Cytology.* 30 361-366. (In Russian).

[3] Papayan G.V., Magakyan Yu.A., Agroskin L.S., Karalova E.M. (1982) The use of scanning cytometric data for the analysis of the absorbent distribution in the cells. *Cytology.* 24 1359-1366. (In Russian).

[4] Petunin Yu.I., Boiko Yu.V., Ganina K.P., Litvinko P.G. (1990) A clustering index of chromatin structure in interphase nuclei. *Proceeding of Academy of Science of UkSSR.* N 5. - P.75-78. (In Russian).

[5] Magakyan Yu.A., Karalova E.M. (1989) *Cytophotometria of DNA.* Erevan: Public House of the Academy of Sciences of Armenia, 203 P.

ÖZET

Göğüs kanseri teşhisi (CMG) ve FAM için bilgisayara dayalı yöntemlerin matematiksel yönleri incelenmiştir. CMG'nin tespiti ve kabul etmeme kararı hata olasılıklarının %6'dan büyük olmadığı gösterilmiştir. FAM için teşhis hatası olasılığının pratik olarak sıfır, kararın kabul edilmesi olasılığının %43 olduğu belirlenmiştir. Teşhis için bilgisayar yöntemi sunulmuştur. Bu yöntem yüksek risk sınıfında olan hastaların erken göğüs kanseri teşhisi amacıyla uygulanabilmektedir.

İstatistik, Journal of the Turkish Statistical Association
1999, Vol.2, No 2, pp. 107-121

COMPARISON OF THE BANDWIDTH SELECTION METHODS FOR KERNEL ESTIMATION OF PROBABILITY DENSITY FUNCTION

Öniz Toktamış and Serpil Gökçe Cula
Hacattepe University Faculty of Science,
Department of Statistics, Beytepe-Ankara

Serdar Kurt
Dokuz Eylül University Faculty of Art and Sciences,
Department of Statistics, Alsancak-İzmir

Abstract

In this study the bandwidth selection methods for the kernel estimation of the probability density function are discussed. Least-squares cross-validation method, biased cross-validation method and bootstrap method are reviewed, compared and their applications are presented.

Key Words: Kernel estimation, bandwidth, cross-validation, biased cross-validation, bootstrap.

1. Introduction

Let $X_1, X_2, ..., X_n$ be a random sample from an unknown absolutely continuous distribution with probability density function f. The kernel estimate derived from this sample is

$$\hat{f}(x,h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

Here, K is the kernel function such that it is usually a probability density function unimodal and symmetric around zero. h is the window width which is called

the bandwidth or also called smoothing parameter because it controls the degree of smoothing in data for the kernel estimation.

The kernel estimation at a given point is weighted mean which is calculated by overlapping the mean point of the kernel function with the given point and taking account the other observations with certain weights which are obtained according to the kernel function and bandwidth (Toktamış, 1995).

Kernel estimation method is one of the non-parametric methods to estimate probability density function and it was suggested first in 1956 by Rosenblatt and theoretical properties were investigated by Parzen in 1962 (Rosenblatt, 1956; Parzen, 1962).

In application, K and h are selected by the users. For different choices of K and h, the estimates of the density function differ. The way of choosing K and h has been the interest of many studies. The choice of the kernel function K was studied first by Epanechnikov in 1969. Epanechnikov showed that there exists an optimal kernel in some sense,but there are other kernels which give almost optimal results (Epanechnikov, 1969). It is quite satisfactory to choose a kernel for computational convenience or differentiability properties. For this reason, the choice of kernel function is not as important as the choice of bandwidth in application (Silverman, 1986).

The choice of bandwidth has a very important place in the kernel density estimation. Boneva and his colleagues showed that small changes in bandwidth could change estimates on large scale (Silverman, 1978). A lot of methods were suggested and investigated to select the bandwidth. But there has been no commonly acceptable method up to now. In this study, the most commonly used methods will be investigated, compared and some applications will be presented.

2. Bandwidth selection methods

To examine the performance of kernel estimator, several criterions related to the deviation of $\hat{f}$ from the real probability density function f were considered. Commonly used one of these criterions was suggested by Rosenblatt and it is known as the mean integrated squared error (MISE). MISE is a preferable criterion because it is mathematically simple. It is defined as follows:

$$MSE\left\{\hat{f}(x,h)\right\} = \int_{-\infty}^{\infty} E\left\{\hat{f}(x,h) - f(x)\right\}^2 dx$$

$$= (nh)^{-1} \int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4}h^4 \left[\int_{-\infty}^{\infty} u^2 K(u) du\right]^2 \int_{-\infty}^{\infty} f^{\prime\prime}(x)^2 dx + o\left\{(nh)^{-1} + h^4\right\} \quad (2)$$

And, the asymptotic mean integrated squared error (AMISE) is defined as follows (Wand and Jones, 1995):

$$AMISE\left\{\hat{f}(x,h)\right\} = (nh)^{-1}\int_{-\infty}^{\infty} K(u)^2du + \frac{1}{4}h^4\left[\int_{-\infty}^{\infty} u^2K(u)du\right]^2\int_{-\infty}^{\infty} f^{\text{II}}(x)^2dx. \quad (3)$$

The most appropriate way to select bandwidth is to find h value which minimizes MISE. Optimal h value obtained from (2) is given as, h_{opt} ,

$$h_{opt} \cong \left[\int_{-\infty}^{\infty} u^2K(u)du\right]^{-2/5}\left\{\int_{-\infty}^{\infty} K(u)^2du\right\}^{1/5}\left\{\int_{-\infty}^{\infty} f^{\text{II}}(x)^2dx\right\}^{-1/5} n^{-1/5} \quad (4)$$

As it can be seen from this equation, the optimal h value depends on the second derivative of the unknown density function f. For this reason different methods are suggested to select h value. Some of these methods will be given in the following sections.

2.1. Least squares cross-validation

As a data method, the least-squares cross-validation (LSCV) was suggested by Rudemo (1982) and Bowman (1984), independently each other. Let $\hat{f}$ be a kernel estimator of a probability density function f. Then MISE can be written as follows:

$$MSE\left\{\hat{f}(x,h)\right\} = E\int \hat{f}(x,h)^2dx - 2E\int \hat{f}(x,h)f(x)dx + \int f(x)^2dx. \quad (5)$$

Here the aim is to find h value which minimizes MISE. Selection of h value which minimizes MISE is equivalent to the h value which minimizes the following expression

$$MSE\left\{\hat{f}(x,h)\right\} - \int f(x)^2dx = E\int \hat{f}(x,h)^2dx - 2E\int \hat{f}(x,h)f(x)dx. \quad (6)$$

The right-hand side of the (6) depends on f , it is not known. But it can be shown that an unbiased estimator for the right-hand side is,

$$LSCV(h) = \int \hat{f}(x,h)^2dx - 2n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}(X_i,h)\,. \quad (7)$$

Here

$$\hat{f}_{-i}(X_i,h) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{X_i - X_j}{h}\right)$$

and it is a kernel estimate obtained by using all observations with X_i deleted. This is the reason for the term "cross-validation" which refers to the use of part of a sample to obtain information about another part. (7) is called least-squares cross-validation function and because of (7) gives an unbiased estimator of (6) it is also called unbiased cross-validation function (Cula, 1998).

2.3. The bootstrap

The bootstrap (B) is based on the following base : Main simple was assumed to be a population. Several samples were drawn from main sample with replacement. For each of the drawn sample, related estimators are calculated. If the bootstrap is used to estimate MISE, then bias component can not be calculated. For this reason, the procedure to choice of the bandwidth h is different than the ordinary bootstrap and it is called as smoothed bootstrap. In this method the bandwidth h for sample $\{X_i\}$ by using one of the preceding methods used to find the kernel estimate $\hat{f}$. The bootstrap sample $\{X_i^*\}$ is chosen from density $\hat{f}$ using the following algorithm (Faraway and Jhun, 1990).

Step 1 : An integer j is chosen from $\{1, 2, ..., n\}$ with equal probability.

Step 2 : A random variable Φ is derived from the probability density function K.

Step 3 : Set $X_j^* \rightarrow X_j^* + h\Phi$

Then many samples are obtained by repeating the preceding procedure and taking expected value MISE is calculated. Taylor said that if the standard probability density function is taken as the kernel function, the for bootstrap estimate of MISE it is not necessary to take a sample over again. If the standard normal probability density function is used for the kernel function, then bootstrap estimate of MISE is,

$$B(h) = \frac{1}{2n^2h\sqrt{2\pi}}\left[\sum_{i,j}\exp\left\{-\frac{(x_j-x_i)^2}{8h^2}\right\} - \frac{4}{\sqrt{3}}\sum_{i,j}\exp\left\{-\frac{(x_j-x_i)^2}{6h^2}\right\}\right.$$

$$\left. + \sqrt{2}\sum_{i,j}\exp\left\{-\frac{(x_j-x_i)^2}{4h^2}\right\} + n\sqrt{2}\right] \qquad (10)$$

(Taylor, 1989). It can be seen that if observed values are replaced in (10), then only a bootstrap function depending on h is obtained. The bandwidth which minimizes this function is found and this bandwidth is shown as $\hat{h}_B$.

2.4. Comparison of bandwidth selection methods

In the theoretical point of view, various estimators are compared according to the rate of convergence of some non-random error criterion, such as MISE, to zero. The concept of rate of convergence is an asymptotic concept. For this reason, the concept of rate of convergence is used for large sample sizes. For small samples, comparisons

The basic principle of least-squares cross-validation is to find the kernel estimates from the data for various h values and to select h value which minimizes (7). Bandwidth obtained by this strategy will be shown as $\hat{h}_{LSCV}$.

The least-squares cross-validation function can have more than one local minimum. Researchers show that in this case it is appropriate to take the largest local minimizer of LSCV. Because the largest local minimizer is the nearest bandwidth to the optimal bandwidth obtained from MISE (Wand and Jones,1995).

2.2. Biased cross-validation

AMISE, which is a simple formula with respect to MISE, also depend on $\int_{-\infty}^{\infty} f^{\prime\prime}(x)^2 dx$ like MISE. Scoot, Tapia and Thompson have taken for $\int_{-\infty}^{\infty} f^{\prime\prime}(x)^2 dx$ the integral $\int_{-\infty}^{\infty} \hat{f}^{\prime\prime}(x)^2 dx$ by using kernel estimator $\hat{f}$ in order to obtain the bandwidth. Scott and Terrell said that this estimator is deficient asymptotically and it is appropriate to use

$$\int_{-\infty}^{\infty} \tilde{f}^{\prime\prime}(x)^2 dx = \int_{-\infty}^{\infty} \hat{f}^{\prime\prime}(x)^2 dx - \frac{1}{nh^5} \int_{-\infty}^{\infty} K^{\prime\prime}(x)^2 dx \tag{8}$$

for the estimate of $\int_{-\infty}^{\infty} f^{\prime\prime}(x)^2 dx$ (Scott and Terrell, 1987). (8) is the adjusted value of $\int_{-\infty}^{\infty} \hat{f}^{\prime\prime}(x)^2 dx$. Then biased cross-validation (BCV) function is obtained by substituting (8) into the asymptotic expression and is given as follows:

$$BCV(h) = \frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4} h^4 \left[\int_{-\infty}^{\infty} u^2 K(u) du \right]^2 \int_{-\infty}^{\infty} \tilde{f}^{\prime\prime}(x)^2 dx \tag{9}$$

In this study, the bandwidth value which minimizes the function given by (9) will be shown as $\hat{h}_{BCV}$.

are made with respect to simulation studies. The commonly used method to choose a bandwidth was LSCV criterion between the years 1982-1988. General performance of this method is not satisfactory very well. The rate of convergence of the estimator of LSCV is $O(n^{-1/10})$ (Wand and Jones, 1995). To obtain optimal bandwidth, very large sample size is needed (Chiu, 1992).

An unpleasant aspect of LSCV and BCV, which has been noticed in simulation studies and in applications to real data sets, is that the LSCV function and BCV function often have more than one local minima. BCV selection method has also the same rate of convergence of order $n^{-1/10}$ like LSCV method (Hall and Marron, 1991). The studies show that the bandwidth which is obtained by BCV is larger than the bandwidth which is obtained by LSCV. One advantage of BCV method is to have small variance with respect to LSCV. Scoot and Terrell showed that an attractive property of $\hat{h}_{BCV}$ with respect to $\hat{h}_{LSCV}$ is that $\hat{h}_{BCV}$ has minimum asymptotic variance. In this case, $\hat{h}_{BCV}$ is more stable than $\hat{h}_{LSCV}$ (Chao et al., 1994). From the simulation studies, it was seen that these selectors are to be directed to select small bandwidth with respect to asymptotic theorems. Density estimate which is obtained by using small bandwidths show spurious structure (Chiu, 1991). Researchers show that these selectors have an unsatisfied performance with theoretical and practical points of view. Performance of B method is better than the performance of LSCV and BCV for some distributions because of the standard deviation of bootstrap's bandwidth selectors is small (Faraway and Jhun, 1990).

The bandwidth value which is obtained by B is larger than the bandwidth values which are obtained by LSCV and BCV. The bootstrap bandwidth has smaller variance but computational cost of bootstrap is higher than LSCV and BCV criterion.

3. Application

The optimal bandwidth, h_{opt} , which minimizes (4) is

$$h_{opt} \cong \left[\int_{-\infty}^{\infty} u^2 K(u)du\right]^{-2/5} \left\{\int_{-\infty}^{\infty} K(u)^2 du\right\}^{1/5} \left\{\int_{-\infty}^{\infty} f''(x)^2 dx\right\}^{-1/5} n^{-1/5}$$

If it is assumed that probability density function is known as a normal distribution with sample mean 100 and variance 4 and kernel function is assumed as a standard

normal distribution, then optimal bandwidths for sample sizes $n = 50, 100, 250$ and 500 can be obtained as follows by using the formula above:

$$h_{opt} = 0.969240 \ \ for \ n = 50$$

$$h_{opt} = 0.843773 \ \ for \ n = 100$$

$$h_{opt} = 0.702486 \ \ for \ n = 250$$

$$h_{opt} = 0.611549 \ \ for \ n = 500$$

For example, if the probability density function is an exponential distribution with parameter $\lambda = 1$, then for $n = 50, 100, 250$ and 500 optimal bandwidths will be obtained as follows :

$$h_{opt} = 0.407868 \ \ for \ n = 50$$

$$h_{opt} = 0.355070 \ \ for \ n = 100$$

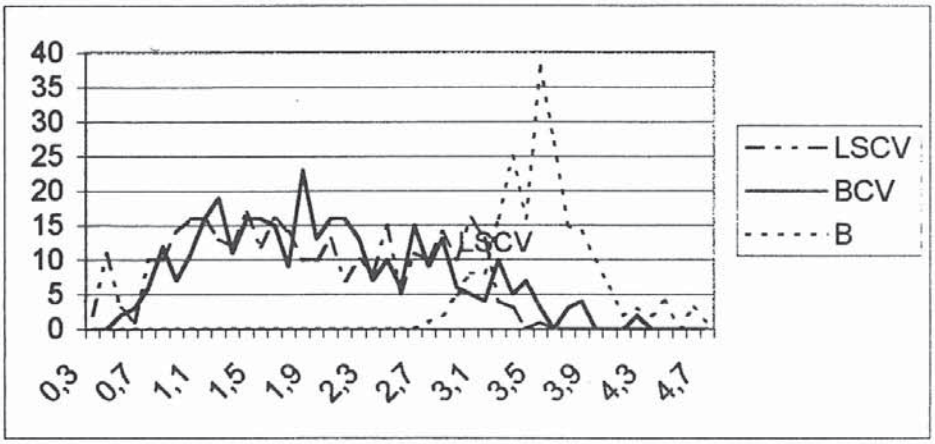
$$h_{opt} = 0.295615 \ \ for \ n = 250$$

$$h_{opt} = 0.257348 \ \ for \ n = 500$$

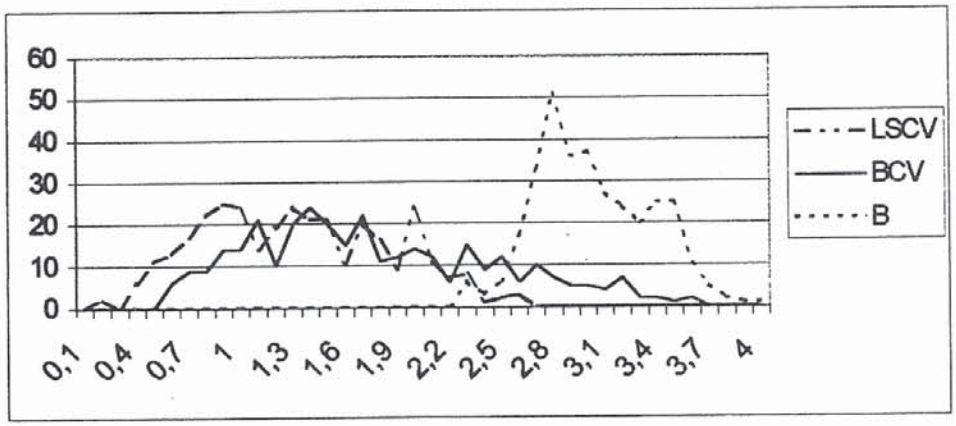
In fact the kernel estimation was used when the sample was taken from an unknown distribution. In this study, samples with several sizes were drawn from known distribution and optimal bandwidths were obtained using (4). The bandwidths which were obtained by using the methods of LSCV, BCV, B were both compared with each other and investigated the closeness of them to optimal bandwidth.

First 350 samples were drawn from the normal distribution with mean 100 and variance 4 for each sizes $n = 50, 100, 250$ and 500. By taking standard normal distribution as kernel function, for each 350 samples with $n = 50$ bandwidths were obtained by using the methods of cross-validation, biased cross-validation and bootstrap and their distributions were found. These procedures were repeated for $n = 100, 250$ and 500. The distributions of bandwidth are given in the following Figure 3.1.

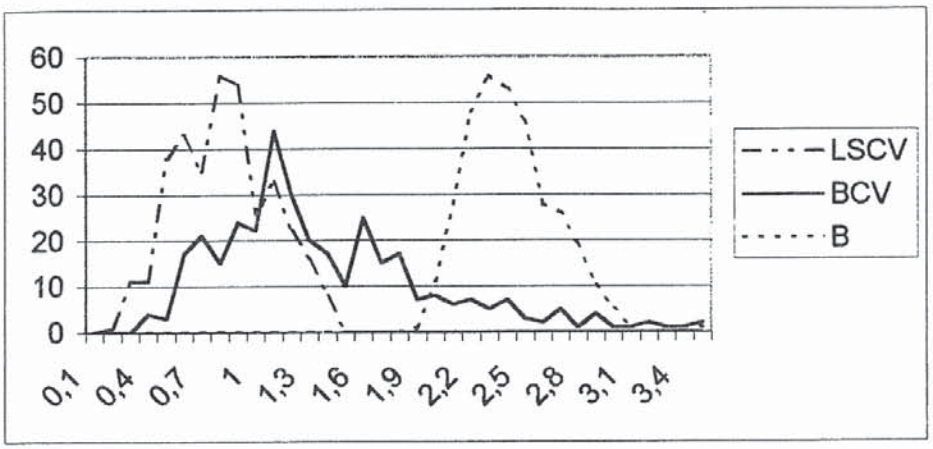
$n = 50$


40
35
30
25
20
15
10
5
0
0,3
0,7
1,1
1,5
1,9
2,3
2,7
3,1
3,5
3,9
4,3
4,7
LSCV
LSCV
BCV
B

$n = 100$


60
50
40
30
20
10
0
0,1
0,4
0,7
1
1,3
1,6
1,9
2,2
2,5
2,8
3,1
3,4
3,7
4
LSCV
BCV
B

n=250



n=500

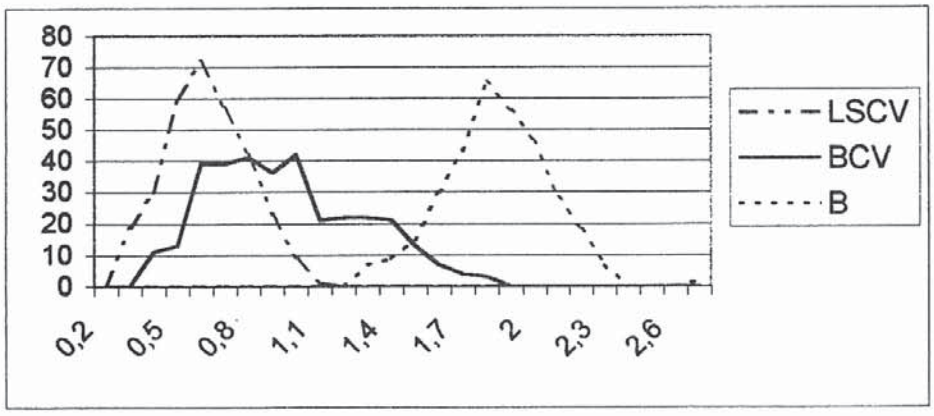


Figure 3.1. The distributions of bandwidths for each 350 samples from the normal distribution with sample sizes $n = 50, 100, 250$, and 500 by using the methods of LSCV, BCV, and B.

For each bandwidth distribution, mean and variance are obtained. These values are given in following Table 3.1.

Table 3.1. Means and variances of the bandwidth distributions which are found for samples from normal distribution with various methods.

	LSCV	BCV	B
n=50	$\hat{h} = 1{,}821407$ $Var(\hat{h}) = 0{,}649172$	$\hat{h} = 1{,}943505$ $Var(\hat{h}) = 0{,}721859$	$\hat{h} = 3{,}529612$ $Var(\hat{h}) = 0{,}127460$
n=100	$\hat{h} = 1{,}328485$ $Var(\hat{h}) = 0{,}282590$	$\hat{h} = 1{,}765337$ $Var(\hat{h}) = 0{,}495502$	$\hat{h} = 2{,}981763$ $Var(\hat{h}) = 0{,}383692$
n=250	$\hat{h} = 0{,}822881$ $Var(\hat{h}) = 0{,}071172$	$\hat{h} = 1{,}387032$ $Var(\hat{h}) = 0{,}376796$	$\hat{h} = 2{,}430861$ $Var(\hat{h}) = 0{,}06767$
n=500	$\hat{h} = 0{,}623642$ $Var(\hat{h}) = 0{,}02995$	$\hat{h} = 0{,}953251$ $Var(\hat{h}) = 0{,}106302$	$\hat{h} = 1{,}837037$ $Var(\hat{h}) = 0{,}04729$

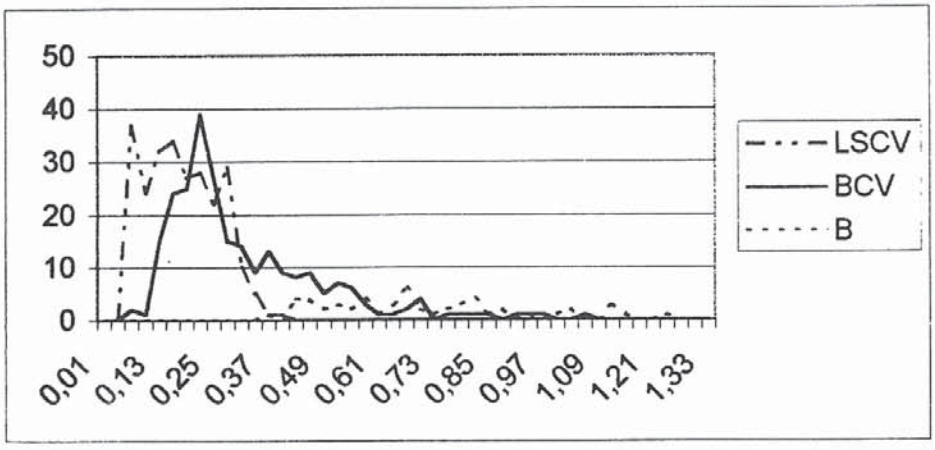
From these figures, it can be seen that when sample size increases, then the bandwidths which are obtained from each of 350 samples have more smooth distribution. It is seen that bandwidths which are obtained by using BCV criterion are greater than the bandwidths which are obtained by using LSCV criterion and the bandwidths which are obtained by using B criterion are greater than the bandwidth which are obtained by using BCV criterion. From these figures and table it can be seen that the variance of the bandwidths of B is small. This shows that the estimation is more stable. When the sample size is increasing, then the variance of the bandwidth is decreasing for all methods. But when $n = 50$, in samples out off 350 samples bandwidth can not be obtained with bootstrap method. For the other samples, bandwidth can not be obtained with bootstrap method. For this reason when $n = 50$, the variance of the distribution of bandwidths is smaller than others in the table and this is a fallacy.

As seen from the table the method which gives the faraway bandwidth from the optimal bandwidth is the bootstrap method. For example , for $n = 100$ $\hat{h}_{opt} = 0.843773, \hat{h}_{LSCV} = 1.328485, \hat{h}_{BCV} = 1.765337$, and $\hat{h}_B = 2.981763$. When the sample size is small, then the variance of LSCV selectors is greater than the variance of BCV estimators. But when the sample size is increasing LSCV selectors is closer to the optimal bandwidth and have small variance with respect to BCV selectors.

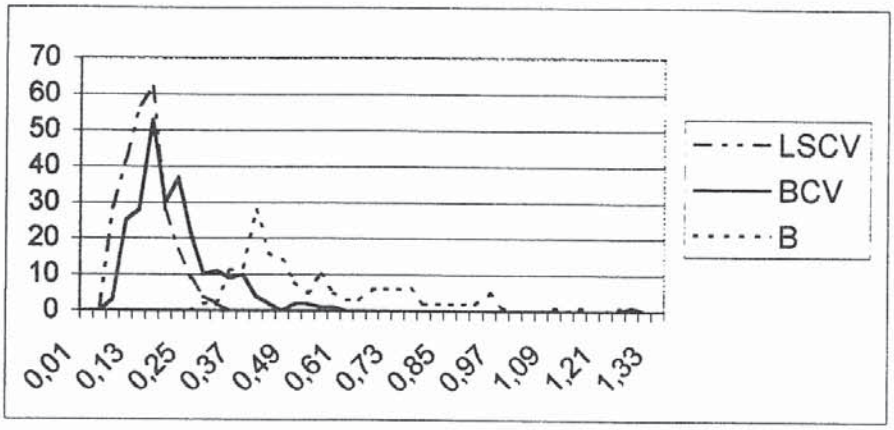
If the probability density function shows a symmetric and smooth distribution as in the Figure 3.1, it can be said that LSCV's bandwidth in large sample is closer to the optimal bandwidth.

Secondly, 250 samples were drawn from a non-symmetric distribution with parameter $\lambda = 1$ for each sample sizes $n = 50, 100, 250$, and 500. By taking standard normal distribution as kernel function, for each 250 samples by using the methods of cross-validation, biased cross validation , and bootstrap, the bandwidths were obtained and their distributions were found. These distributions are given in the following Table 3.2.

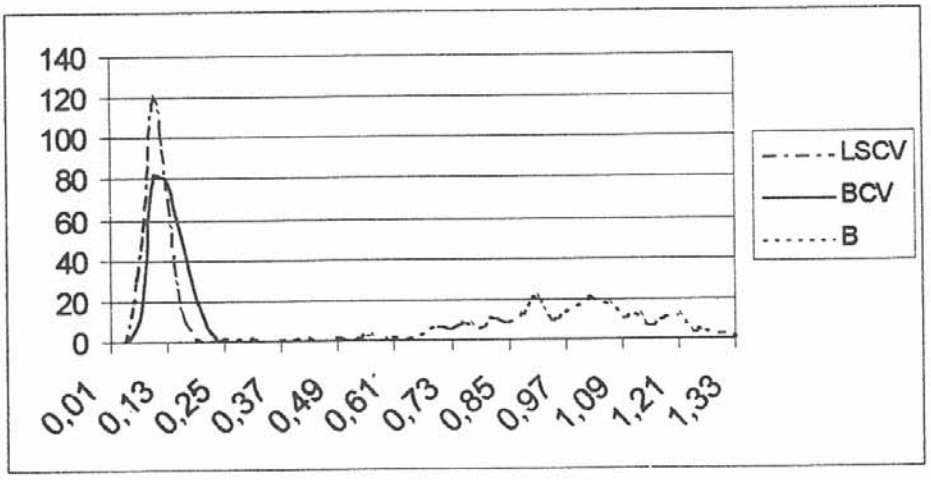
$n = 50$



$n = 100$


70
60
50
40
30
20
10
0
0,01
0,13
0,25
0,37
0,49
0,61
0,73
0,85
0,97
1,09
1,21
1,33
LSCV
BCV
B

$n = 250$


140
120
100
80
60
40
20
0
0,01
0,13
0,25
0,37
0,49
0,61
0,73
0,85
0,97
1,09
1,21
1,33
LSCV
BCV
B

$n = 500$

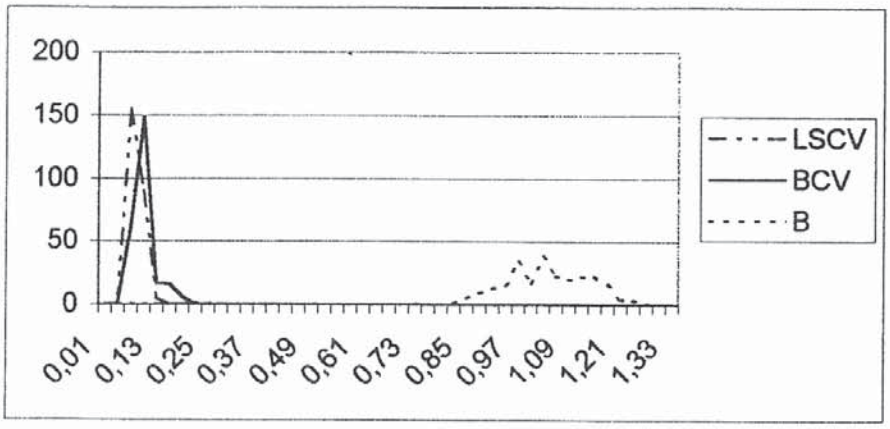


Figure 3.2. The distributions of bandwidths for each 250 random samples from the exponential distribution with sample sizes $n = 50, 100, 250$, and 500 by using the methods of CV, BCV, and B.

For each bandwidth distribution the mean and variances are obtained. These values are given in Table 3.2.

Table 3.2. : Means and variances of the bandwidth distributions which are found for samples from exponential distribution

	LSCV	BCV	B
n=50	$\hat{h} = 0{,}18028$ $Var(\hat{h}) = 0{,}006111$	$\hat{h} = 0{,}313036$ $Var(\hat{h}) = 0{,}029497$	$\hat{h} = 0{,}705556$ $Var(\hat{h}) = 0{,}047312$
n=100	$\hat{h} = 0{,}14644$ $Var(\hat{h}) = 0{,}002624$	$\hat{h} = 0{,}210843$ $Var(\hat{h}) = 0{,}008596$	$\hat{h} = 0{,}539259$ $Var(\hat{h}) = 0{,}036186$
n=250	$\hat{h} = 0{,}10672$ $Var(\hat{h}) = 0{,}000632$	$\hat{h} = 0{,}130600$ $Var(\hat{h}) = 0{,}001119$	$\hat{h} = 0{,}953080$ $Var(\hat{h}) = 0{,}022534$
n=500	$\hat{h} = 0{,}0820$ $Var(\hat{h}) = 0{,}000255$	$\hat{h} = 0{,}10060$ $Var(\hat{h}) = 0{,}000712$	$\hat{h} = 1{,}03492$ $Var(\hat{h}) = 0{,}014567$

In this case it can be seen that biased cross-validation's bandwidths are larger than cross-validation's bandwidths, and bootstrap's bandwidths are larger than biased cross-validation's bandwidths. From the table and figures, we can see that the variance of cross-validation's bandwidth are smaller than the variances of the other method's bandwidth distributions. But the method which gives the nearest bandwidth to the optimal bandwidth is biased cross-validation method. For example, for $n = 100, \hat{h}_{opt} = 0.35507, \hat{h}_{LSCV} = 0.14644, \hat{h}_{BCV} = 0.210843$, and $\hat{h}_B = 0.539259$.

If the probability density function shows a non-symmetric distribution, for large samples BCV method's bandwidth gets closer to the optimal bandwidth.

References

[1]Bowman, A. W., 1984, An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, 71, 2, 353-360.

[2]Cao, R., Cuevas, A. and Manteiga, G. W., 1994, A comparative study of several smoothing methods in density estimation, *Comp. Statist.and Data Anal.*, 17, 153-176.

[3]Chiu, S. T., 1991, The effect of discretization error on bandwidth selection for kernel density estimation, *Biometrika*, 78, 436-441.

[4]Chiu, S. T., 1992, An Automatic Bandwidth Selector for Kernle Density Estimation, *Biometrika*, 79, 4, 771-782.

[5]Cula, S.G., 1998, Çok Değişkenli Olasılık Yoğunluk Fonksiyonunun Çekirdek Fonksiyonları ile Kestirimi, *Doktora Tezi, H.Ü., Fen Fakültesi*, Ankara, 1212.

[6]Epanechnikov, V. A., 1969, Nonparametric estimation of a multivariate probability density, *Theory of Probability and It's Applications*, 14, 153-158.

[7]Faraway, J. J. and Jhun, M., 1990, Bootstrap choice of bandwidth for density estimation, *Journal of the American Statistical Assoc.* Vol. 85, No.412, 1119-1122.

[8]Hall, P., and Marron, J. S., 1991, Local Minima in Cross-validation Functions, *J.R.Statis. Soc.* B, 53, No. 1, 245-252.

[9]Park, B. U. and Marron, J. S.,1990, Comparison of data-driven bandwidth selectors, *Journal of the American Statistical Association Methods*, Vol.85, No.409, 66-72.

[10]Parzen, E., 1962, On the estimation of a probability density function and the mode, *Ann. Math Statistics*, 33, 1065-1076. *Royal Stat. Soc.* Ser B 34, 385-392.

[11]Rosenblatt, M., 1956, Remarks on some non-parametric estimates a density function, *Ann. Math.Statist.*, 27, 832-837.

[12]Rudemo, M., 1982, Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* 9, 65-78.

[13]Scott, D. W. and Terrell, G. R.,1987, Biased and unbiased cross-validation in density estimation, *Journal of the American Statisticsl Association*, Vol.82, No.400, 1131-1146.

[14]Sheather, S. J., Jones, M. C., 1991, A reliable data-based bandwidth selection methods for kernel density estimation, *J. R. Statist. Soc.* Ser B 53, No.3, 683-690.

[15]Silverman, B. W., 1978, Choosing the window width when estimating a density, *Biometrika*, 65, 1, 1-11p.

[16]Silverman, B. W., 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[17]Simonoff, J. S., 1996, *Smoothing Methods in Statistics*, Verlag, New York.

[18]Taylor, C. C., 1989, Bootstrap choice of the smoothing paramater in kernel density estimation, *Biometrika*, 76,4, 705-712.

[19]Toktamış, Ö., 1995, Olasılık yoğunluk fonksiyonunun çekirdek kestirimi üzerine bir çalışma *Hacettepe Fen ve Mühendislik Bilimleri Dergisi*, 16, 145-163.

[20]Wand, M. P. and Jones, M. C., 1995, *Kernel Smoothing*, Chapman and Hall, New York.

ÖZET

Bu çalışmada olasılık yoğunluk fonksiyonlarının çekirdek tahmin edicileri için bant genişliği seçimine yönelik yöntemler ele alınmıştır. En küçük kareler, çapraz doğrulama ve bootstrap yöntemleri tanıtılmış ve karşılaştırılarak uygulamalar yapılmıştır.

İstatistik, Journal of the Turkish Statistical Association
1999, Vol.2, No 2, pp. 123-133

ON THE NUMBER OF PRODUCTIVE ANCESTORS IN LARGE POPULATIONS

I. Rahimov
KFUPM Box 1339, Dhahran 31261, Saudi Arabia

H. Hasan
Department of Mathematics, USM, Penang 11800, Malaysia

Abstract

We consider a population of n individuals. Each of these individuals generates a discrete time branching stochastic process. We study the number of ancestors $S(n,t)$ whose offspring at time t exceeds level $\theta(t)$, where $\theta(t)$ is some positive valued function. It is proved that $S(n,t)$ may be approximated as $t \to \infty$ by some stochastic processes with independent increments if $n \to \infty$ depending on the time of observation.

Key Words: Population, ancestor, branching process, Poisson process, Brownian motion, Binomial process, exceedance

1. Introduction

We consider a population containing n individuals of the same type at time zero. Each of these individuals (ancestors) initiates a discrete time branching population process. Let $\theta(t), t \in \mathbb{N}_0 = \{0, 1, \ldots\}$ be a positive valued function and $S(n,t)$ be the number of ancestors having more than $\theta(t)$ descendants at time t.

Branching processes started by the initial ancestors may be considered as population processes describing population growth in different regions of an area R. Then it is easy to see that $S(n,t)$ is the number of regions of R whose population at time t exceeds level $\theta(t)$. Process $S(n,t)$ can be associated with a problem on the number of vertexes of rooted random trees as well. In fact each realization of the scheme under the consideration can be interpreted as a forest containing n rooted trees. Consequently a realization of $S(n,t)$ is the number of trees in the forest having more than $\theta(t)$ vertexes of the level t.

We note here the rise of interest in recent years to problems concerning extrema in branching stochastic processes. For example the recent publications in this direction have been devoted to the asymptotic behaviour of the expectation of the maxima of branching processes (Borovkov, Vatutin (1996), Pakes (1998)), to the limit distribution

for the maximum family size (Arnold, Villasenor (1996), Rahimov, Yanev (1999)) and to other problems. Limit distributions for the index of the first process in a sequence of branching processes exceeding some fixed or increasing levels were obtained in Rahimov, Hasan (1998). Thus the study of $S(n,t)$ can be considered as a contribution to this program of investigation of the extrema in population processes.

It follows from well-known properties of branching processes (see Athrey, Ney (1972), for example) that if n is fixed and the process is critical or subcritical, then $S(n,t)$ in the long run equals to zero with probability 1, for any level function $\theta(t)$.

What happens if the size of the initial population is large? In other words what is the asymptotic behaviour of $S(n,t)$ if the number of initial ancestors increases depending on the time of observation? To answer these questions we consider family of stochastic process $y(x,t) = S([m(t)x],t)$, where $x \in [0,\infty)$ and $m(t) \to \infty$ as $t \to \infty$. We approximate $y(x,t)$ by some known processes with independent increments. Behaviour of the parameter $m(t)$ and the form of limit processes naturally depend on criticality of the initial branching process. It turns out that, if the process is supercritical, then $y(x,t)$ may be approximated by either a "binomial process" (process with independent and binomially distributed increments) or by the Brownian motion depending on the behaviour of $m(t)$. If the process is subcritical or critical, then the approximating process is either a Poisson process or the Brownian motion.

Now we give a rigorous definition of the process $S(n,t)$. Let $\mathcal{A}_t^i$ be the random population at time t generated by i-th initial ancestor, $i = 1,2,\ldots,n$. For any positive valued function $\theta(t)$ functional $S(n,t) = S(n,t)[\theta]$ can be defined as following

$$S(n,t) = \#\{i : \text{ card } \mathcal{A}_t^i > \theta(t)\}.$$

Let $X_i(t) = \text{card } \mathcal{A}_t^i$ be i-th branching process and $X(t)$ be a branching process such that $X(t) \stackrel{d}{=} X_i(t)$ for all $i \geq 1$. We denote $\{P_k, k \geq 0\}$ the offspring distribution of $X(t)$ and put

$$\begin{aligned} f(S) &= \sum_{k=0}^{\infty} P_k S^k, \quad R(x,t) = P\{X(t) > x\}, Q(t) = R(0,t), \\ A &= \sum_{k=1}^{\infty} kP_k, \quad \sigma^2 = \sum_{k=1}^{\infty} k(k-1)P_k. \end{aligned}$$

2. Critical processes

First we consider the critical case, i.e., the case of $A = 1$, $0 < \sigma^2 < \infty$. We assume that there exits the following

$$\lim_{t\to\infty} \frac{\theta(t)}{t} = \theta \in [0,\infty] \tag{1}$$

and consider $y(x,t) = S([tx],t)$, i.e., $m(t) = t$.

Theorem 1. *If $A = 1$, $0 < \sigma^2 < \infty$ and (1) is satisfied, then $y(x,t) \xrightarrow{\mathcal{D}} y(x)$ as $t \to \infty$, where $\mathcal{D}$ means convergence in the weak sense and $y(x)$ is the Poisson process with $Ey(x) = 2x\exp\{-2\theta/\sigma^2\}/\sigma^2$ for $\theta \in [0,\infty)$ and it is a "zero process" (i.e., $y(x) \equiv 0$ with probability 1 for all $x \in [0,\infty)$) for $\theta = \infty$.*

Theorem 1 gives an approximation of $S(n,t)$ for the case when $n = o(t)$ or $n \asymp t$ as $t \to \infty$. Now we consider the case when $n/t \to \infty$, $t \to \infty$. More precisely we put $m(t) = a(t)t$, where $a(t) \to \infty$. We define the stochastic process $W_t^{(1)}(x)$ as follows

$$W_t^{(1)}(x) = \frac{S([ta(t)x],t) - [ta(t)x]R(\theta(t),t)}{\sqrt{a(t)}},$$

where $R(\theta(t),t) = P\{X(t) > \theta(t)\}, \quad x \in [0,\infty)$.

Theorem 2. *If $A = 1$, $0 < \sigma^2 < \infty$ and (1) is satisfied, then $W_t^{(1)}(x) \xrightarrow{\mathcal{D}} W^{(1)}(x)$ as $t \to \infty$, where $\mathcal{D}$ means, as before, convergence in the weak sense and $W^{(1)}(x)$ is the Brownian motion with zero shift and with the diffusion parameter $2\sigma^{-2}\exp\{-2\theta/\sigma^2\}$ for $\theta \in [0,\infty)$ and it is a zero process for $\theta = \infty$.*

Proof of Theorem 1. Since the lives of individuals are independent and identically distributed we obtain that

$$ES^{S(n,t)} = [1 - (1-S)R(\theta(t),t)]^n. \tag{2}$$

We use the following well known results for critical branching processes (see Harris (1963), pp. 19–22). If $A = 1, 0 < \sigma^2 < \infty$, then for any fixed $x > 0$

$$P\{Q(t)X(t) > x | X(t) > 0\} \sim e^{-x}, \quad Q(t) \sim 2/\sigma^2 t, \qquad \text{as } t \to \infty. \tag{3}$$

It follows from (3) that under the condition (1)

$$R(\theta(t),t) \sim \frac{2}{\sigma^2 t}\exp\left\{-\frac{2\theta}{\sigma^2}\right\} t \to \infty. \tag{4}$$

Therefore

$$\begin{aligned}\lim_{t\to\inf} \ln ES^{y(x,t)} &= -\lim_{t\to\infty}[xt]R(\theta(t),t)(1-s) \\ &= -\frac{2x}{\sigma^2}e^{-2\theta/\sigma^2}(1-s) \end{aligned} \tag{5}$$

for $\theta \in [0,\infty)$. Consequently the generating function of $y(x,t)$ tends as $t \to \infty$ to

$$\exp\left\{\frac{2x}{\sigma^2}e^{-2\theta/\sigma^2}(s-1)\right\}$$

which is the generating function of the one dimensional distribution for Poisson process $y(x)$.

Now we consider

$$P\{y(x_i,t) = k_i,\ i = 0,1,\ldots,r\},$$

where $0 = x_0 < x_1 < \cdots < x_r < \infty$, $r = 1,2,\ldots$ First we prove that

$$\lim_{t\to\infty} ES^{y(x_2,t)-y(x_1,t)} = \exp\left\{\frac{2(x_2-x_1)}{\sigma^2}e^{-2\theta/\sigma^2}(s-1)\right\}. \tag{6}$$

In fact, since

$$y(x_2,t)-y(x_1,t)=\sum_{i=[x_1t]+1}^{[x_2t]}\varepsilon_i, \tag{7}$$

where

$$\varepsilon_i=\varepsilon_i(t)=\begin{cases}1, & \text{if } X_i(t)>\theta(t)\\ 0, & \text{if } X_i(t)\le\theta(t)\end{cases}$$

and $X_i(t)$ is the process generated by i-th ancestor, we have

$$\lim_{t\to\infty}\ln ES^{y(x_2,t)-y(x_1,t)}=\lim_{t\to\infty}([x_2t]-[x_1t])R(\theta(t),t)(s-1).$$

Thus again due to the limit theorem for critical processes we obtain (6) from the last relation.

It follows from (6) and (7) that

$$\lim_{t\to\infty}E\left[\prod_{i=1}^{r}S_i^{y(x_i,t)-y(x_{i-1},t)}\right]=\exp\left\{\sum_{i=1}^{r}\frac{2(x_i-x_{i-1})}{\sigma^2}e^{-2\theta/\sigma^2}(s_i-1)\right\}.$$

Since the last limit is the generating function of $(y(x_i)-y(x_{i-1}), i=1,\dots,r)$, we conclude that joint distributions of increments of $y(x,t)$ tend to ones of $y(x)$. According to Corollary 1 to Theorem 5 in Billingsley (1968) (see Billingsley (1968), p. 31) from convergence of increments we obtain that

$$(y(x_1,t),\dots,y(x_r,t))\to(y(x_1),\dots,y(x_r)),$$

as $t\to\infty$ in distribution for any $r\ge 1$ and $\theta\in[0,\infty)$.

Thus theorem is proved for $\theta\in[0,\infty)$.

The proof for $\theta=\infty$ follows from the fact that in this case the limit on the right side of (5) is zero.

Proof of Theorem 2. It follows from (2) and definition of the process $W_t^{(1)}(x)$ that

$$Ee^{i\lambda W_t^{(1)}(x)}=\exp\left\{-\frac{i\lambda nR(\theta(t),t)}{\sqrt{a(t)}}\right\}(1-(1-s)R(\theta(t),t))^n,$$

where $n=[ta(t)x]$, $s=e^{i\lambda/\sqrt{a(t)}}$. If we use the following Taylor expansions

$$\ln(1-x) = -x+O(x^2), \quad x\to 0 \tag{8}$$

$$e^{\alpha i} = 1+i\alpha-\frac{\alpha^2}{2}+o(\alpha^2), \quad \alpha\to\infty, \tag{9}$$

we obtain

$$\ln Ee^{i\lambda W_t^{(1)}(x)}=-[ta(t)x]\frac{\lambda^2}{2a(t)}R(\theta(t),t)+o(tR(\theta(t),t)).$$

Taking into account relation (4) we conclude that

$$\lim_{t\to\infty}Ee^{i\lambda W_t^{(1)}(x)}=\exp\left\{-\frac{\lambda^2x}{\sigma^2}e^{-2\theta/\sigma^2}\right\},$$

which is the characteristic function (Fourier transform) of the one dimensional distribution of the Brownian motion $W^{(1)}(x)$.

Let $0 = x_0 < x_1 < \cdots < x_r < \infty$. To prove convergence of finite dimensional distributions we first show that the distribution of

$$(W_t^{(1)}(x_j) - W_t^{(1)}(x_{j-1}), \quad j = 1, \ldots, r)$$

as $t \to \infty$ converges to the distribution of $(W^{(1)}(x_j) - W^{(1)}(x_{j-1}), \quad j = 1, \ldots, r)$.

Again using (2), definition of $W_t^{(1)}(x)$ and Taylor expansions (8), (9) we obtain that

$$\lim_{t\to\infty} E \exp\left\{ i \sum_{j=1}^{r} \lambda_j \left[W_t^{(1)}(x_j) - W_t^{(1)}(x_{j-1}) \right] \right\}$$

$$= \exp\left\{ -\sigma^{-2} e^{-2\theta/\sigma^2} \sum_{j=1}^{r} \lambda_j^2 (x_j - x_{j-1}) \right\}, \tag{10}$$

for any $r \geq 1$, where $\lambda_j \in R, \quad j = 1, \ldots, r$. Since the last limit is the Fourier transform of the distribution of $(W^{(1)}(x_j) - W^{(1)}(x_{j-1}), j = 1, \ldots, r)$ we conclude from here that the joint distribution of increments of the process $W_t^{(1)}(x)$ converges as $t \to \infty$ to the joint distribution of Brownian motion's increments.

Hence due to the mentioned above Corollary 1 in Billingsley (1968, p. 31) the finite dimensional distributions of $W_t^{(1)}(x)$ converges as $t \to \infty$ to ones of the Brownian motion $W^{(1)}(x)$ with zero shift and with diffusion parameter $2\sigma^{-1} \exp\{-2\theta/\sigma^2\}$. The theorem is proved for $\theta \in [0, \infty)$. The proof for $\theta = \infty$ follows from the same arguments, if we take into account that in this case the limit on the right side of (1) is 1. Theorem 2 is proved.

3. Supercritical processes

Now we consider the case of supercritical processes. It is known (Athreya, Ney (1972)) that if $A > 1$, $EX(1) \ln X(1) < \infty$, then $X(t)A^{-1}$ converges with probability one to a random variable W and the Laplace transform $\varphi(\lambda)$ of W satisfies the following equation

$$\varphi(\lambda) = f\left(\varphi\left(\frac{\lambda}{A}\right)\right).$$

It is also known that the distribution function $\pi(x)$ of W is absolute continuous for $x > 0$ and has an atom of the mass q at $x = 0$. Here q is the extinction probability.

We assume that there exists

$$\lim_{t\to\infty} \theta(t) A^{-1} = \theta \in [0, \infty] \tag{11}$$

and

$$\sum_{k=2}^{\infty} k P_k \ln k < \infty. \tag{12}$$

and consider "discrete time" process $S(n, t), n = 0, 1, \ldots$ for $t \in \mathbb{N}_0$. Note that here n is the time parameter.

Theorem 3. *If $A > 1$ and conditions (11) and (12) are satisfied, then $S(n,t) \xrightarrow{\mathcal{D}} \xi(n)$, $n \in \mathbb{N}_0$ as $t \to \infty$, where $\xi(n)$ is a stochastic process with independent and binomially distributed increments such that*

$$P\{\xi(n_i) - \xi(n_{i-1}) = k\} = \begin{pmatrix} n_i - n_{i-1} \\ k \end{pmatrix} [1 - \pi(\theta)]^k \pi(\theta)^{n_i - n_{i-1} - k}$$

for any $0 \le n_{i-1} < n_i < \infty$, $n_i \in \mathbb{N}_0$, for $\theta \in [0, \infty)$ and it is a zero process for $\theta = \infty$.

Example 1. Let the offspring distribution be the positive geometric distribution, i.e. $P_k = \alpha(1-\alpha)^{k-1}, k \ge 1$ and $P_0 = 0$. In this case the offspring generating function has the form $f(s) = \alpha s(1 - \beta s)^{-1}, \beta = 1 - \alpha$ and $A = \alpha^{-1}$ and the equation for the Laplace transform is:

$$\varphi\left(\frac{\lambda}{\alpha}\right) = \frac{\alpha\varphi(\lambda)}{1 - \beta\varphi(\lambda)}.$$

Now it is not difficult to check that the Laplace transform $\varphi(\lambda) = \alpha(\alpha + \lambda)^{-1}$ satisfies the above equation. Hence the limit distribution $\pi(x)$ is exponential with the density function $\alpha e^{-\alpha x}$ and Theorem 3 gives the following result.

Corollary. If conditions of Theorem 3 are satified and the offspring distribution is the positive geometric of the parameter $o < \alpha < 1$, then for $\theta \in [0, \infty)$ the limit process $\xi(n)$ in Theorem 3 is binomial such that

$$P\{\xi(n_i) - \xi(n_{i-1}) = k\} = \begin{pmatrix} n_i - n_{i-1} \\ k \end{pmatrix} e^{-\alpha\theta k} \left[1 - e^{-\alpha\theta}\right]^{n_i - n_{i-1} - k}.$$

Proof of Theorem 3. Let $n_0, n_1, \ldots, n_r$ be such number that $0 = n_0 < n_1 < \cdots < n_r < \infty$ and $n_i \in \mathbb{N}_0$, $0 \le i \le r$. First we prove that for $1 \le i \le r$

$$\lim_{t \to \infty} E S^{S(n_i,t) - S(n_{i-1},t)} = (\hat{\pi}(\theta) S_i + \pi(\theta))^{n_i - n_{i-1}}, \tag{13}$$

where $\hat{\pi}(\theta) = 1 - \pi(\theta)$. It follows from representation (7) that

$$E S_i^{S(n_i,t) - S(n_{i-1},t)} = (R(\theta(t), t) S_i + 1 - R(\theta(t), t))^{n_i - n_{i-1}}. \tag{14}$$

Now we consider the estimate

$$|P\{X(t) \le \theta(t)\} - \pi(\theta)| \le \sup_x |P\{X(t)A^{-1} \le x\} - \pi(x)| + |\pi(\theta(t)A^{-t}) - \pi(\theta)|. \tag{15}$$

First term on the right side of (15) tends to zero as $t \to \infty$ due to the limit theorem for supercritical processes. It follows from condition (11) and continuity of $\pi(x)$ that the limit of the second term is also zero. Thus

$$R(\theta(t), t) \to 1 - \pi(\theta), \tag{16}$$

as $t \to \infty$. From relations (14) and (16) we obtain (13).

Using independence of increments of $S(n,t)$ from relation (13) we have

$$\lim_{t\to\infty} E\left[\prod_{i=1}^{r} S^{S(n_i,t)-S(n_{i-1},t)}\right] = \prod_{i=1}^{r}\{\hat{\pi}(\theta)S_i + \pi(\theta)\}^{n_i-n_{i-1}},$$

which proves the theorem for $\theta \in [0,\infty)$. In the case $\theta = \infty$ the limit on the right side of (13) equals 1 and the "limit process" $\xi(n)$ equals zero for all n. The theorem is proved.

Theorem 3 shows that stochastic process $S(n,t)$ for fixed $n \in \mathbb{N}_0$ can be approximated as $t \to \infty$ by a binomial process. Now we consider the case when $n \to \infty$ together with t. Let $a(t)$ be a positive function such that $a(t) \to \infty$ as $t \to \infty$. We consider the following stochastic process

$$W_t^{(2)}(x) = \frac{S([a(t)x],t) - [a(t)x]R(\theta(t),t)}{\sqrt{a(t)}},$$

where $x \in [0,\infty)$.

Theorem 4. *If $A > 1$ and conditions (11) and (12) are satisfied, then $W_t^{(2)}(x) \xrightarrow{\mathcal{D}} W^{(2)}(x)$ as $t \to \infty$, where $W^{(2)}(x)$ is the Brownian motion with zero shift and with diffusion parameter $\pi(\theta)(1-\pi(\theta))$ for $\theta \in [0,\infty)$ and it is a zero process for $\theta = \infty$.*

Example 2. If, as in Example 1,the offspring distribution is the positive geometric of the parameter $o < \alpha < 1$, then it is not difficult to see that the Brownian motion in Theorem 4 has the diffusion parameter $e^{-\alpha\theta}\left(1 - e^{-\alpha\theta}\right)$.

Proof. First we prove the convergence of the one dimensional distribution. Let

$$A(\lambda) = Ee^{i\lambda W_t^{(2)}(x)}, B(\lambda) = \exp\left\{-\frac{\lambda^2}{2a(t)}\sum_{j=1}^{[a(t)x]} \text{var } \varepsilon_j\right\},$$

where $\varepsilon_j = \varepsilon_j(t), j = 1,2,\ldots$ are the same as in the representation (7). Note that it follows from the definition of $W_t^{(2)}(x)$ and (7) that

$$A(\lambda) = \prod_{j=1}^{[a(t)x]} Ee^{i\lambda(\varepsilon_j - R(t))/\sqrt{a(t)}}, \tag{17}$$

where $R(t) = R(\theta(t),t)$.

Using inequality $\left|\prod_j a_j - \prod_j b_j\right| \le \sum_j |a_j - b_j|$, $|a_j| \le 1$, $|b_j| \le 1$, and taking into account the fact that $E(\varepsilon_j - R(t)) = 0$ for $j = 1,2,\ldots$, we have the following estimate

$$|A(\lambda) - B(\lambda)| \le T_1 + T_2, \tag{18}$$

where with $\alpha_j(t) = \lambda^2 \text{ var } (\varepsilon_j)/2a(t)$

$$T_1 = \sum_{j=1}^{[a(t)x]} \left| e^{i\lambda(\varepsilon_j - R(t))/\sqrt{a(t)}} - 1 - \frac{\varepsilon_j - R(t)}{\sqrt{a(t)}} i\lambda + \alpha_j(t)\right|,$$

$$T_2 = \sum_{j=1}^{[a(t)x]} E\left|1 - e^{-\alpha_j(t)} - \alpha_j(t)\right|.$$

Using the inequality

$$\left|e^{i\alpha} - 1 - i\alpha + \frac{\alpha^2}{2}\right| \leq \frac{1}{3}|\alpha|^3,$$

we obtain that

$$T_1 \leq \frac{[a(t)x]}{a^{3/2}(t)} E|\varepsilon_1 - R(t)|^3.$$

Here it is easy to see that $E|\varepsilon_1 - R(t)|^3 \leq 2$. Thus we have $T_1 \to 0$ as $t \to \infty$.

Taking into account that var $\varepsilon_j =$ var $\varepsilon_1, j = 1, \ldots$ is bounded and using the Taylor expansion $e^{-x} = 1 - x + o(x), x \to 0$, we obtain

$$T_2 = [a(t)x]|\alpha_1(t) + o(\alpha_1(t)) - \alpha_1(t)| = o(1), \quad t \to \infty.$$

¿From these estimates and from (18) we conclude that functions $A(\lambda)$ and $B(\lambda)$ have the same limit as $t \to \infty$. On the other hand, since var $\varepsilon_j = R(\theta(t), t)(1 - R(\theta(t), t))$, the function $B(\lambda)$ tends as $t \to \infty$ to

$$\exp\left\{-\frac{\lambda^2}{2} x\pi(\theta)(1 - \pi(\theta))\right\},$$

which is the Fourier transform of the one dimensional distribution of the Brownian motion $W^{(2)}(x)$.

Let $0 = x_0 < x_1 < \cdots < x_r < \infty, r \geq 1$. It is not difficult to see that, if we repeat the above arguments, we obtain that the characteristic function $Ee^{i\lambda(W_t^{(2)}(x_j) - W_t^{(2)}(x_{j-1}))}$ tends as $t \to \infty$ to

$$\exp\left\{-\pi(\theta)(1 - \pi(\theta))\frac{\lambda^2}{2}(x_j - x_{j-1})\right\},$$

for any $j = 1, 2, \ldots, r$ and $\theta \in [0, \infty)$. Therefore the limit of the characteristic function

$$E \exp\left\{i \sum_{j=1}^{k} \lambda_j \left[W_t^{(2)}(x_j) - W_t^{(2)}(x_{j-1}\right]\right\},$$

as $t \to \infty$ equals to

$$\exp\left\{-\pi(\theta)(1 - \pi(\theta)) \sum_{j=1}^{r} \frac{\lambda_j^2}{2}(x_j - x_{j-1})\right\}.$$

Since the last limit is the Fourier transform of the joint distribution of increments of the Brownian motion $W^{(2)}(x)$, we obtain convergence of finite dimensional distributions from the mentioned above (see proof of Theorem 1) Corollary 1 Billingsley (1968). Theorem 4 is proved.

4. Subcritical processes

Let now $A < 1$, i.e., the initial process is subcritical. In this case we use the following limit theorem for subcritical processes (Sevastyanov (1971), p. 29). If $A < 1$, there exist

$$\lim_{t\to\infty} P\{X(t) = j | X(t) > 0\} = P_j^*; \quad j \geq 1, \tag{19}$$

and the generating function $F^*(s)$ of $P_j^*, \quad j \geq 1$ satisfies the equation

$$1 - F^*(s)) = A(1 - F^*(s)). \tag{20}$$

It is also known that, if $A \leq 1$, then $Q(t) = R(0,t) \to 0$ as $t \to \infty$. If $A < 1$ and in addition $EX(1)\ln X(1) < \infty$, then we have the following asymptotics for $Q(t)$ (see Sevastyonov (1971), p. 56)

$$Q(t) \sim KA^t, 0 < K = \prod_{m=0}^{\infty} B(P\{X(m) = 0\}) < \infty, \tag{21}$$

where $B(s) = (1 - f(s))/(A(1 - s))$.

Let $y(x,t) = S([xA^{-t}], t)$.

Theorem 5. *If $A < 1$ and (12) is satisfied, then $y(x,t) \xrightarrow{\mathcal{D}} y(x)$ as $t \to \infty$, where $y(x)$ is the Poisson process with $Ey(x) = Kx\sum_{j>\theta} P_j^*$ for $\theta(t) \equiv \theta \in \mathbb{N}_0$ and it is a zero process if $\theta(t) \to \infty$.*

Proof. We use again (2) with $n = [xA^{-t}]$. In this case it follows from the above mentioned limit theorem for subcritical processes that under the condition (12)

$$R(\theta(t), t) \sim KA^t \sum_{j>\theta} P_j^*, \quad t \to \infty, \tag{22}$$

for $\theta \in \mathbb{N}_0$. Since

$$\lim_{t\to\infty} \ln ES^{y(x,t)} = (s-1) \lim_{t\to\infty} [xA^{-t}] R(\theta(t), t)$$

we obtain that the limit of $ES^{y(x,t)}$ as $t \to \infty$ is $\exp\left\{Kx\sum_{j>\theta} P_j^*(s-1)\right\}$.

To prove convergence of finite dimensional distribution it is sufficient to show that

$$\lim_{t\to\infty} ES^{y(x_2,t)-y(x_1,t)} = \exp\left\{K(x_2 - x_1)\sum_{j>\theta} P_j^*(s-1)\right\}, \tag{23}$$

for any $0 < x_1 < x_2 < \infty$ and $0 < s < 1$. It follows from representation (7) and (2) that in this case

$$\lim_{t\to\infty} \ln ES^{y(x_2,t)-y(x_1,t)} = \lim_{t\to\infty} n(t) R(\theta(t), t)(s-1),$$

where $n(t) = [x_2 A^{-t}] - [x_1 A^{-t}]$. We obtain (23) from here taking into account relation (22). Thus due to (23)

$$\lim_{t\to\infty} E\prod_{j=1}^{r} S^{y(x_j,t)-y(x_{j-1},t)} = \exp\left\{K(s-1)\sum_{i>\theta} P_i^* \sum_{j=1}^{r}(x_j - x_{j-1})\right\},$$

i.e., $(y(x_j,t) - y(x_{j-1},t), \quad i = 1,2,\dots r)$ converges in distribution to $(y(x_j) - y(x_{j-1}),$ $j = 1,2,\dots,r$.

We obtain convergence of finite dimensional distributions from convergence of increments by the Corollary 1 in Billingsley (1968) (see Billingsley (1969, p. 31) as in the proof of previous theorems. Theorem 5 is proved.

Now we consider the case $nA^t \to \infty$. Let, as before, $a(t)$ be a positive valued function such that $a(t) \to \infty$ as $t \to \infty$. We define process $W_t^{(3)}(x)$ by the relation

$$W_t^{(3)}(x) = \frac{1}{\sqrt{a(t)}}\left\{S([xA^{-t}a(t)],t) - [xA^{-t}a(t)]R(\theta(t),t)\right\},$$

where $x \in [0,\infty)$.

Theorem 6. *If $A < 1$ and (12) is satisfied, then $W_t^{(3)}(x) \xrightarrow{\mathcal{D}} W^{(3)}(x)$ as $t \to \infty$, where $W^{(3)}(x)$ is the Brownian motion with zero shift and with diffusion parameter $K\sum_{i>\theta} P_i^*$ for $\theta(t) = \theta \in \mathbb{N}_0$ and it is a zero process if $\theta(t) \to \infty$.*

Proof. Let $0 \le x_0 < x_1 < \infty$. It follows from definition of $W_t^3(x)$ and representation (7) that

$$A(\lambda) = Ee^{i\lambda(W_t^{(3)}(x_1) - W_t^{(3)}(x_0))} = \exp\left\{-\frac{i\lambda n R(\theta(t),t)}{\sqrt{a(t)}}\right\}(1-(1-s)R(\theta(t),t))^n, \quad (24)$$

where $n = [x_1 A^{-t}a(t)] - [x_0 A^{-t}a(t)]$ and $S = e^{i\lambda/\sqrt{a(t)}}$. If we use Taylor expansions (8) an (9), we have

$$\ln A(\lambda) = -\frac{\lambda^2}{2a(t)} nR(\theta,t) + o(A^{-t}R(\theta,t)), \quad t \to \infty.$$

Taking into account relation (22) we obtain from here that

$$A(\lambda) = \exp\left\{-\frac{\lambda^2}{2}(x_1 - x_0)K\sum_{j>\theta} P_j^*\right\} + o(1), \quad t \to \infty. \quad (25)$$

Let now $0 = x_0 < x_1 < \cdots < x_r < \infty, \quad r \ge 1$. Since lives of different individuals are independent, the increments $W_t^{(3)}(x_j) - W_t^{(3)}(x_{j-1}), \quad j = 1,\dots,r$, are independent. Therefore it follows from (25) that the joint distribution of these increments tends as $t \to \infty$ to the joint distribution of increments $W^{(3)}(x_j) - W^{(3)}(x_{j-1}), \quad j = 1,\dots,r$. To obtain from here convergence of finite dimensional distributions we again appeal to the mentioned above Corollary 1 from Billingsley (1968). Theorem 6 is proved.

References

1. Arnold, B.C., Villasenor, J.A. (1996). The tallest man in the world. *In Statistical Theory and Applications:* Papers in honor of Herbert A. David. ed. H.N. Nagaraja et al., Springer, Berlin, pp. 81–88.

2. Athreya, K., Ney, P. (1972). *Branching Processes*, Springer-Verlag.

3. Billingsley, P.(1968) *Convergence of Probability Measures.* New York; Wiley.

4. Borovkov, K.A., Vatutin, V.A. (1996). On distribution tails and expectations of maxima in critical branching processes, *J. Appl. Probab.*, 33, 614–622.

5. Pakes, A. (1998). A limit theorem for the maxima of the paracritical branching process, *Adv. Apl. Probab.* 30, 740–756.

6. Rahimov, I., Hasan, H. (1998). Limit theorems for exceedances of a sequence of branching processes, *Bull. Malaysian Math. Soc.* (Second Series) V. 21, 37–46.

7. Rahimov I., Yanev, G. (1999). On the maximum family size in branching processes, *J. Appl. Probab.* 36, No. 3.

8. Sevastyanov, B.A. (1971). *Branching Processes*, Nauka, Moscow.

ÖZET

n bireyden alınan bir kitlede bireylerin kesikli zamanlı dallanma süreçleri yarattığı düşünülmüş, atalar ve çocukların t zamanı bakımından sayıları ele alınmıştır. Ataların sayısı olan $S(n,t)$'nin, t sonsuza giderken artımları bağımsız olan bir sürece yakınsadığı ispatlanmıştır.

İstatistik, Journal of the Turkish Statistical Association
1999, Vol.2, No 2, pp. 135-144

CONCOMITANT OF ORDER STATISTICS IN FGM TYPE BIVARIATE UNIFORM DISTRIBUTIONS

Ismihan G. Bairamov and Muhammet Bekçi
Ankara University, Faculty of Science,
Department of Statistics,
06100, Tandoğan, Ankara, Turkey

Abstract

We consider the bivariate FGM distributions with uniform marginals. The distribution of the concomitant of the r th order statistic of one of the component is obtained. Recurrence relations between moments of concomitants are given.

Key Words: Farlie-Gumbel-Morgenstern distributions, order statistics, norm, concomitants, recurrence relations, product moments.

1. Introduction

The class of bivariate distributions originally proposed by Morgenstern (1956) having a natural form

$$F_{X,Y}(x,y) = F_X(x) F_Y(y) \{1 + \alpha [1 - F_X(x)] [1 - F_Y(y)]\} \tag{1.1}$$

is a flexible family useful in applications provided the correlation between the variables is not too large. It can be utilized for arbitrary continuous marginals. This structure was studied by Farlie (1960) in the form (FGM) of

$$F_{X,Y}(x,y) = F_X(x) F_Y(y) \{1 + \alpha A(F_X(x)) B(F_Y(y))\} \tag{1.2}$$

where $A(x)$ and $B(y)$ satisfy certain regularity conditions ensuring that (1.2) is a distribution function with absolutely continuous marginals $F_X(x)$ and $F_Y(y)$.

Further generalizations of (1.1) to distributions with more than two variables and a stronger correlation structure can be found in Johnson and Kotz (1975, 1977), Kotz and Johnson (1977) and Huang and Kotz (1984). Recent results dealing with this family of distributions are due to Huang and Kotz (1998) who introduce an additional parameter to increase the dependence between the underlying variables. Bairamov and Kotz (1999) present several theorems characterizing symmetry and dependence properties of FGM and Huang-Kotz FGM distributions and provide a modification of Huang-Kotz FGM distributions with large correlation between the components.

Let (X_i, Y_i) , $i = 1, 2, ..., n$ be a random sample from an absolutely continuous bivariate population (X, Y) with distribution function (d.f.) $F_{X,Y}(x, y)$. Let $X_{r:n}$ denote the r th order statistics of the X sample values. Denote by $Y_{[r:n]}$ the Y values associated with $X_{r:n}$. We call $Y_{[r:n]}$ the concomitant of the r th order statistic. Concomitants are used, for example, in selection procedures. Recently Balasubramanian and Beg (1998) have studied concomitants in Gumbel's bivariate exponential distribution. For more details we refer to the review articles of Bhattacharya (1984) and David (1993). Denote probability density function (p.d.f.) of $Y_{[r:n]}$ by $g_{[r:n]}(y)$. It is known that

$$g_{[r:n]}(y) = \int_{-\infty}^{+\infty} f(y \mid x) f_{r:n}(x) \, dx \ , \tag{1.3}$$

where $f(y \mid x)$ is the condition density function of Y, given X and $f_{r:n}(x)$ is the p.d.f. of $X_{r:n}$ (see David (1981), p.110.).

In this paper we shall consider the classical Morgenstern distribution (1.1) with uniform marginals and investigate the distributional and moment properties of concomitants of order statistics.

2. Concomitants in Morgenstern type bivariate distributions

Consider (1.1) for $F_X(x) = x$, $F_Y(y) = y$; $0 < x, y < 1$.

$$F(x, y) = xy\{1 + \alpha(1 - x)(1 - y)\} \ , \ -1 \leq \alpha \leq 1. \tag{2.1}$$

and

$$f(x, y) = 1 + \alpha(1 - 2x)(1 - 2y) \ , \ 0 \leq x, y \leq 1. \tag{2.2}$$

The d.f. of $Y_{[r:n]}$ is given by

$$G_{[r:n]}(y) = \int_{-\infty}^{+\infty} F(y \mid x) f_{r:n}(x) \, dx, \tag{2.3}$$

where $F(y \mid x)$ is a conditional d.f. of Y given X, and

$$f_{r:n}(x) = \frac{n!}{(r-1)!\,(n-r)!}\,[F(x)]^{r-1}\,[1-F(x)]^{n-r}\,f(x)$$

is the p.d.f. of r th order statistic. Since the marginals of (2.1) are Uniform [0,1] we arrive at

$$F(y \mid x) = y\,\{1+\alpha\,(1-2x)\,(1-y)\} \text{ and } f(y \mid x) = 1+\alpha\,(1-2x)\,(1-2y) \quad (2.4)$$

Using (2.3) and (2.4) we derive the d.f. of $Y_{[r:n]}$:

$$G_{[r:n]}(y) = \int_0^1 y\,\{1+\alpha\,(1-2x)\,(1-y)\}\,\frac{n!}{(r-1)!\,(n-r)!}x^{r-1}\,(1-x)^{n-r}\,dx =$$

$$= y\left\{1+\alpha\left[1-2\frac{r}{n+1}\right](1-y)\right\} \quad (2.5)$$

and the corresponding p.d.f. is

$$g_{[r:n]}(y) = 1+\alpha\left[1-2\frac{r}{n+1}\right](1-2y) \quad (2.6)$$

Consider the moments of $Y_{[r:n]}$. From (2.6), the k th moment of $Y_{[r:n]}$ is given by

$$\mu_{r:n}^{(k)} = E\left\{Y_{[r:n]}^{k}\right\} = \int_0^1 y^k\left\{1+\alpha\left[1-2\frac{r}{n+1}\right](1-2y)\right\}dy =$$

$$= \frac{1}{k+1}\left\{1-\alpha\left[1-2\frac{r}{n+1}\right]\left(\frac{k}{k+2}\right)\right\}\ ,\ k=0,1,2,\ldots \quad (2.7)$$

Consequently the expected value and the variance of $Y_{[r:n]}$ can be obtained from (2.7) as follows:

$$E\left\{Y_{[r:n]}\right\} = \frac{1}{2}\left\{1-\frac{\alpha}{3}\left[1-2\frac{r}{n+1}\right]\right\} \quad (2.7a)$$

and

$$Var\left\{Y_{[r:n]}\right\} = \frac{1}{12}\left\{1-\frac{\alpha^2}{3}\left[1-4\frac{r}{n+1}\left(1-\frac{r}{n+1}\right)\right]\right\}. \quad (2.7b)$$

The moment generating function (m.g.f.) of $Y_{[r:n]}$ is given by

$$M_{[r:n]}(t) = E\left\{e^{tY_{[r:n]}}\right\} = \int_0^1 e^{ty}\left\{1+\alpha\left[1-2\frac{r}{n+1}\right](1-2y)\right\}dy$$

$$= \frac{e^t-1}{t}\left\{1+\alpha\left[1-2\frac{r}{n+1}\right]\left[1+2\left(\frac{1}{t}-\frac{e^t}{e^t-1}\right)\right]\right\} \quad (2.8)$$

3. Recurrence relation between moments of concomitants

From (2.6) one observes $g_{[r:n-1]}(y) = 1 + \alpha\left[1 - 2\frac{r}{n}\right](1-2y)$ and

$$g_{[r:n]}(y) - g_{[r:n-1]}(y) = \alpha\frac{2r}{n(n+1)}(1-2y) \tag{3.1}$$

Also

$$g_{[r-1:n]}(y) = 1 + \alpha\left[1 - 2\frac{r-1}{n+1}\right](1-2y)$$

and

$$g_{[r:n]}(y) - g_{[r-1:n]}(y) = -\alpha\frac{2}{n+1}(1-2y). \tag{3.2}$$

Relation (3.1) and (3.2) can be extended:

$$g_{[r:n]}(y) - g_{[r:n-i]}(y) = \alpha\frac{2ri}{(n+1)(n-i+1)}(1-2y)\ ,\ 1 \le i \le n-r \tag{3.3}$$

and

$$g_{[r:n]}(y) - g_{[r-j:n]}(y) = -\alpha\frac{2j}{(n+1)}(1-2y)\ ,\ 1 \le j \le r-1 \tag{3.4}$$

Moreover the following equalities are valid

$$g_{[r-j:n-i]}(y) = 1 + \alpha\left[1 - 2\frac{r-j}{n-i+1}\right](1-2y)\ ,\ 1 \le i \le n-r\ ;\ 1 \le j \le r-1$$

and

$$g_{[r:n]}(y) - g_{[r-j:n-i]}(y) = \alpha\frac{2\left[ri - j(n+1)\right]}{(n+1)(n-i+1)}(1-2y)\ ,\ 1 \le i \le n-r\ ;\ 1 \le j \le r-1 \tag{3.5}$$

Let $1 \le i_1 < i_2 \le n-r$ and $1 \le j_1 < j_2 \le r-1$. Then

$$g_{[r-j_1:n-i_1]}(y) - g_{[r-j_2:n-i_2]}(y) = 2\alpha\left[\frac{r-j_2}{n-i_2+1} - \frac{r-j_1}{n-i_1+1}\right](1-2y) \tag{3.6}$$

Using (3.6) one obtains the following general recurrence relation between the moments of concomitants:

$$\mu^{(k)}_{[r-j_1:n-i_1]} - \mu^{(k)}_{[r-j_2:n-i_2]} = 2\alpha\left[\frac{r-j_1}{n-i_1+1} - \frac{r-j_2}{n-i_2+1}\right]\frac{k}{(k+1)(k+2)} \tag{3.7}$$

In particular the following relations are valid

$$\mu^{(k)}_{[r:n]} - \mu^{(k)}_{[r:n-i]} = \alpha\frac{2ri}{(n+1)(n-i+1)}\frac{-k}{(k+1)(k+2)} \tag{3.8}$$

$$\mu^{(k)}_{[r:n]} - \mu^{(k)}_{[r-j:n]} = \alpha \frac{2j}{(n+1)(k+1)(k+2)} \tag{3.9}$$

$$\mu^{(k)}_{[r:n]} - \mu^{(k)}_{[r-j:n-i]} = \alpha \frac{2[ri - j(n+1)]}{(n+1)(n-i+1)} \frac{-k}{(k+1)(k+2)} \tag{3.10}$$

From (2.8) clearly follows the following recurrence relation for m.g.f. of concomitants:

$$M_{[r-j_1:n-i_1]}(t) - M_{[r-j_2:n-i_2]}(t) = 2\alpha \left[\frac{r-j_2}{n-i_2+1} - \frac{r-j_1}{n-i_1+1} \right] \times$$

$$\times \frac{e^t - 1}{t} \left[1 + 2\left(\frac{1}{t} - \frac{e^t}{e^t - 1} \right) \right], \; 1 \le i_1 < i_2 \le n-r \; ; \; 1 \le j_1 < j_2 \le r-1. \tag{3.11}$$

In particular, one can obtain the following relations between m.g.f. $Y_{[r:n]}$ and $Y_{[r-j:n-i]}$:

$$M_{[r:n]} - M_{[r:n-i]} = \alpha \frac{2ri}{(n+1)(n-i+1)} \frac{e^t - 1}{t} \left[1 + 2\left(\frac{1}{t} - \frac{e^t}{e^t - 1} \right) \right] \tag{3.12}$$

$$M_{[r:n]} - M_{[r-j:n]} = -\alpha \frac{2j}{(n+1)} \frac{e^t - 1}{t} \left[1 + 2\left(\frac{1}{t} - \frac{e^t}{e^t - 1} \right) \right] \tag{3.13}$$

$$M_{[r:n]} - M_{[r-j:n-i]} = \alpha \frac{2[ri - j(n+1)]}{(n+1)(n-i+1)} \frac{e^t - 1}{t} \left[1 + 2\left(\frac{1}{t} - \frac{e^t}{e^t - 1} \right) \right] \tag{3.14}$$

4. Joint distribution of concomitants

Let $Y_{[r_1:n]}$, $Y_{[r_2:n]}, ..., Y_{[r_k:n]}$ be the concomitants of $X_{r_1:n}$, $X_{r_2:n}, ..., X_{r_k:n}$, respectively, where $1 \le r_1 < r_2 < ... < r_k \le n$. The joint probability density function of $\left(Y_{[r_1:n]}, Y_{[r_2:n]}, ..., Y_{[r_k:n]}\right)$ is

$$g_{[r_1,r_2,...,r_k:n]}(y_1, y_2, .., y_k) =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{x_k} ... \int_{-\infty}^{x_2} f(y_1 \mid x_1) f(y_2 \mid x_2) ... f(y_k \mid x_k) f_{r_1,r_2,...,r_k:n}(x_1, x_2, .., x_k) dx_1 ... dx_k \tag{4.1}$$

where $f_{r_1,r_2,...,r_k:n}(x_1, x_2, .., x_k)$ is the joint p.d.f. of $(X_{r_1:n}, X_{r_2:n}, ..., X_{r_k:n})$.

The joint p.d.f. of two concomitants $\left(Y_{[r:n]}, Y_{[s:n]}\right)$ $(1 \le r < s \le n)$ for (2.1) can be calculated by using (4.1) as follows

$$g_{Y_{[r:n]},Y_{[s:n]}}(y_1, y_2) = 1 + \alpha \left(1 - 2\frac{r}{n+1} \right)(1 - 2y_1) + \alpha \left(1 - 2\frac{s}{n+1} \right)(1 - 2y_2)$$

$$+\alpha^2\left(1-2\frac{r+s}{n+1}+4\frac{r\,(s+1)}{(n+1)\,(n+2)}\right)(1-2y_1)\,(1-2y_2) \tag{4.2}$$

For $k=3$, the joint p.d.f. of $\left(Y_{[r_1:n]},Y_{[r_2:n]},Y_{[r_3:n]}\right)$ $(1\leq r_1<r_2<r_3\leq n)$ is

$$g_{[r_1,r_2,r_3]}(y_1,y_2,y_3)=1+\alpha\left(1-2\frac{r_1}{n+1}\right)(1-2y_1)+$$

$$+\alpha\left(1-2\frac{r_2}{n+1}\right)(1-2y_2)+\alpha\left(1-2\frac{r_3}{n+1}\right)(1-2y_3)+$$

$$+\alpha^2\left(1-2\frac{r_1+r_2}{n+1}+4\frac{r_1\,(r_2+1)}{(n+1)\,(n+2)}\right)(1-2y_1)\,(1-2y_2)+$$

$$+\alpha^2\left(1-2\frac{r_1+r_3}{n+1}+4\frac{r_1\,(r_3+1)}{(n+1)\,(n+2)}\right)(1-2y_1)\,(1-2y_3)+$$

$$+\alpha^2\left(1-2\frac{r_2+r_3}{n+1}+4\frac{r_2\,(r_3+1)}{(n+1)\,(n+2)}\right)(1-2y_2)\,(1-2y_3)+$$

$$+\alpha^3\left(1-2\frac{r_1+r_2+r_3}{n+1}+4\frac{r_1\,(r_2+1)+r_1\,(r_3+1)+r_2\,(r_3+1)}{(n+1)\,(n+2)}-\right.$$

$$\left.-8\frac{r_1\,(r_2+1)\,(r_3+2)}{(n+1)\,(n+2)\,(n+3)}\right)(1-2y_1)\,(1-2y_2)\,(1-2y_3)\,. \tag{4.3}$$

5. Concomitant of the norm-ordered statistics

Bairamov and Gebizlioğlu (1998) introduced norm-ordered statistics for multivariate data. Let $R^m, m\geq 1$, be the real Euclidean space. Suppose $X_1,X_2,..,.X_n\in R^m$ are independent identically distributed (i.i.d.) random variables ($m>1$ random vectors) (r.v.'s) with distribution function (d.f.) F. Denote by $\|.\|$ the norm defined in R^m. It is clear that $\|X_1\|,\|X_2\|,...,\|X_n\|$ are i.i.d. r.v. with d.f. $P\{\|X_i\|\leq x\}\equiv F^*(x)$, $x\in R$. If F is assumed to be continuous, the probability of any two or more of these r.v. assuming equal magnitudes is zero. Therefore, there exists a unique ordered arrangement within the r.v. $\|X_i\|$, $i=1,2,...,n$. We say that X_1 precedes X_2 (or that X_1 is less than X_2 in a norm sense) if $\|X_1\|\leq\|X_2\|$ and denote $X_1\prec X_2$. Suppose $X^{(1)}$ denotes the smallest of the set $X_1,X_2,..,.X_n$; $X^{(2)}$ denotes the second smallest, etc. ; and $X^{(n)}$ denotes the largest in a norm sense. The distribution of norm-ordered statistics is expressed in terms of the so called structural function $h(x,y)=P\{\|X_1\|\leq\|\bar{x}\|\}$, where $\bar{x}=(x_1,x_2,...,x_m)\in R^m$ which can be estimated empirically. Specifically the p.d.f. of r th norm-ordered statistic $X^{(r)}$ is

$$f_r(x_1,x_2,...,x_k)=$$

$$n\binom{n-1}{r-1}[h(x_1,x_2,...,x_k)]^{r-1}[1-h(x_1,x_2,...,x_k)]^{n-r} f(x_1,x_2,...,x_k) \ ,$$

$$if \ \ \|\bar{x}_1\| < \|\bar{x}_2\| < ... < \|\bar{x}_k\|$$

and

$$f_r(x_1,x_2,...,x_k) = 0 \ , \ otherwise$$

The joint p.d.f. of $(X^{(1)}, X^{(2)}, ..., X^{(n)})$ is

$$f_{1,2,...,n}(x_1,y_1,x_2,y_2,....,x_n,y_n) =$$

$$= \begin{cases} n!f(x_1,y_1)f(x_2,y_2)...f(x_n,y_n) \ , & if \ \ \|\bar{x}_1\| < \|\bar{x}_2\| < ... < \|\bar{x}_k\| \\ 0 \ , & otherwise \end{cases}$$

Here we define concomitants for norm-ordered statistics as follows.

Let (X,Y,Z) be the absolutely continuous r.v. with the d.f. $F(x,y,z)$ and p.d.f. $f(x,y,z)$. Let (X_i,Y_i,Z_i), $i=1,2,...,n$ be the independent copies of (X,Y,Z). Suppose that the first two coordinates of (X,Y) are ordered in a norm sense, i.e., let $X^{(1)} \prec X^{(2)} \prec ... \prec X^{(n)}$ be the norm-ordered statistics of (X_i,Y_i) $i=1,2,...,n$. Denote by $Z_{[r:n]}$ the Z values associated with $X^{(r)}$. We call $Z_{[r:n]}$ the concomitant of r th norm-ordered statistics $X^{(r)}$.

The d.f. of $Z_{[r:n]}$ can be found as follows:

$$P\left\{Z_{[r:n]} \le z\right\} = P\left\{\bigcup_{k=1}^{n} Z_k \le z, X^{(k)} = X_{(r)}\right\}$$

By using total probability formula for continuous random variables

$$P(A) = \int_{-\infty}^{+\infty} P(A \mid X = x)\, dF_X(x)$$

one can write

$$P\left\{Z_{[r:n]} \le z\right\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} F(z \mid x,y)\, f_{r:n}(x,y) \quad {}^{\prime}y \equiv G_{[r:n]}(z) \tag{5.1}$$

and

$$g_{[r:n]}(z) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(z \mid x,y)\, f_{r:n}(x,y)\, dxdy,$$

where

$$\begin{aligned} f_{r:n}(x,y) &= \frac{n!}{(r-1)!\,(n-r)!}[h(x,y)]^{r-1}[1-h(x,y)]^{n-r} f(x,y)\, dxdy \ , \quad (5.2) \\ h(x,y) &= P\{\|(X,Y)\| \le \|(x,y)\|\} \end{aligned}$$

Now consider the following three-variate FGM distributions with unit exponential marginals:

$$F(x,y,z) = \left(1-e^{-x}\right)\left(1-e^{-y}\right)\left(1-e^{-z}\right)\left\{1+\alpha e^{-x-y-z}\right\},\ x,y,z>0\ ;\ -1\le\alpha\le 1. \tag{5.3}$$

This is a trivariate extension of Gumbel's bivariate exponential distribution. The p.d.f. is

$$f(x,y,z) = e^{-x-y-z}\left\{1+\alpha\left[2e^{-x}-1\right]\left[2e^{-y}-1\right]\left[2e^{-z}-1\right]\right\}\ ,\ x,y,z>0 \tag{5.4}$$

Evidently

$$f(z\mid x,y) = e^{-z}\left\{1+\alpha\left[2e^{-x}-1\right]\left[2e^{-y}-1\right]\left[2e^{-z}-1\right]\right\},\ x,y,z>0\ ; \tag{5.5}$$

$$F(z\mid x,y) = \int_0^z f(z\mid x,y)\,dz$$

$$= \left(1-e^{-z}\right)\left\{1+\alpha e^{-z}\left[2e^{-x}-1\right]\left[2e^{-y}-1\right]\right\} \tag{5.6}$$

By using (5.1) we can obtain the d.f. of norm-ordered concomitants for the distribution of the form (5.3). For example, consider the distribution function of $Z_{[1:n]}$. Let $\|(x,y)\| = |x|+|y|$. Then

$$h(x,y) = P\{|X|+|Y|\le|x|+|y|\} = P\{X+Y\le x+y\} = G_{2,1}(x+y).$$

Hence

$$h(x,y) = 1-(1+x+y)\,e^{-x-y} \text{ and} \tag{5.7}$$

$$1-h(x,y) = (1+x+y)\,e^{-x-y}. \tag{5.8}$$

Using (5.2)-(5.8) in (5.1) one can write

$$P\left\{Z_{[1:n]}\le z\right\} = \int_0^{+\infty}\int_0^{+\infty} F(z\mid x,y)\,n\left[1-h(x,y)\right]^{n-1} f(x,y)\,dxdy$$

$$= \int_0^{+\infty}\int_0^{+\infty}\left(1-e^{-z}\right)\left\{1+\alpha e^{-z}\left[2e^{-x}-1\right]\left[2e^{-y}-1\right]\right\}$$

$$n\left[(1+x+y)\,e^{-x-y}\right]^{n-1} e^{-x-y}dxdy$$

$$= n\left(1-e^{-z}\right)\int_0^{+\infty}\int_0^{+\infty}\left\{1+\alpha e^{-z}\left[2e^{-x}-1\right]\left[2e^{-y}-1\right]\right\}(1+x+y)^{n-1}\,e^{-nx-ny}dxdy$$

$$= n\left(1-e^{-z}\right)\left\{I_0+\alpha e^{-z}\left[4I_1-2I_2-2I_3+I_4\right]\right\},$$

where

$$I_0 = I_4 = \int_0^{+\infty}\int_0^{+\infty} (1+x+y)^{n-1} e^{-nx-ny} dxdy = \frac{1}{n},$$

$$I_1 = \int_0^{+\infty}\int_0^{+\infty} (1+x+y)^{n-1} e^{-(n+1)x-(n+1)y} dxdy = \frac{1}{n+1} - \frac{(n-1)!}{(n+1)^{n+1}} \sum_{l=0}^{n-1} \frac{(n+1)^l}{l!},$$

$$I_2 = \int_0^{+\infty}\int_0^{+\infty} (1+x+y)^{n-1} e^{-(n+1)x-ny} dxdy = \frac{(n-1)!}{n^n} \sum_{l=0}^{n-1} \frac{n^l}{l!} - \frac{(n-1)!}{(n+1)^n} \sum_{l=0}^{n-1} \frac{(n+1)^l}{l!},$$

$$I_3 = \int_0^{+\infty}\int_0^{+\infty} (1+x+y)^{n-1} e^{-nx-(n+1)y} dxdy = \frac{(n-1)!}{n^n} \sum_{l=0}^{n-1} \frac{n^l}{l!} - \frac{(n-1)!}{(n+1)^n} \sum_{l=0}^{n-1} \frac{(n+1)^l}{l!},$$

$$I_2 = I_3.$$

We thus have

$$P\left\{Z_{[1:n]} \le z\right\} = \left(1-e^{-z}\right)\left\{1+\alpha e^{-z}\left[4n\left(\frac{1}{n+1} - \frac{(n-1)!}{(n+1)^{n+1}} \sum_{l=0}^{n-1} \frac{(n+1)^l}{l!}\right) - \right.\right.$$

$$\left.\left. -4n\left(\frac{(n-1)!}{n^n} \sum_{l=0}^{n-1} \frac{n^l}{l!} - \frac{(n-1)!}{(n+1)^n} \sum_{l=0}^{n-1} \frac{(n+1)^l}{l!}\right) + 1\right]\right\}.$$

Denote

$$h(n) = 1 + 4\frac{n}{n+1} + 4nn! \sum_{l=0}^{n-1} \frac{1}{l!}\left[(n+1)^{l-n-1} - n^{l-n-1}\right].$$

Then one has

$$P\left\{Z_{[1:n]} \le z\right\} = \left(1-e^{-z}\right)\left\{1+\alpha e^{-z} h(n)\right\} \tag{5.9}$$

and p.d.f. of $Z_{[1:n]}$

$$g_{[1:n]}(z) = \left(1-\alpha h(n) e^{-z}\right) + 2\alpha h(n) e^{-2z}\ ,\ z>0. \tag{5.10}$$

References

[1] Bairamov, I. G. and Gebizlioğlu, Ö. L. (1998). On the ordering of random vectors in a norm sense. *Journal of Appl. Statist. Sci.*, Vol. **6**, Num. 1, 77-86.

[2] Bairamov, I. G. and Kotz, S. (1999). Dependence structure and symmetry of Huang-Kotz FGM distributions and their extensions. (to appear)

[3] Balasubramanian, K. and Beg, M. I. (1997). Concomitant of order statistics in Morgenstern type bivariate exponential distributions. *Journal of Appl. Statist.* Sci., Vol. **5**, Num. 4, 233-245.

[4] Balasubramanian, K. and Beg, M. I. (1998). Concomitant of order statistics in Gumbel's bivariate exponential distribution. *Sankhya, The Indian Journal of Statistics*, Vol. **60**, Series B, Pt. 3, 399-406.

[5] Bhattacharya, P. K. (1984). Induced order statistics. Theory and application. *In Handbook of Statistics* (P. R. Krishnaiah and P. K. Sen eds.), **4**, 383-403. North-Holland, Amsterdam.

[6] David, H. (1981). *Order Statistics.* 2 nd edn., John Wiley, New York.

[7] David, H. (1993). *Concomitant of order statistics*: Review and recent developments. In Multiple Comparisons, Selection and Application in Biometry (F. M. Hoppe, ed.), 507-518, Dekker, New-York.

[8] Farlie, D. J. G. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, **47**, 307-323.

[9] Gumbel, E. J. (1960). Bivariate exponential distributions. *Journ. Amer. Statist. Assoc.*, **55**, 698-707.

[10] Huang, J. S. and Kotz, S. (1984). Correlation structure in iterated Farlie-Gumbel-Morgenstern distributions. *Biometrika*, **71**, 633-636.

[11] Huang, J. S. and Kotz, S. (1999). Modifications of the Farlie-Gumbel-Morgenstern distributions. A tough hill to climb. *Metrika*, **49**.

[12] Johnson, N. L. and Kotz, S. (1975). On some generalized Farlie-Gumbel-Morgenstern distributions. *Comm. Statist.*, **4**, 415-427.

[13] Johnson, N. L. and Kotz, S. (1977). On some generalized Farlie-Gumbel-Morgenstern distributions - II: Regression, correlation and further generalizations. *Comm. Statist.*, **6**, 485-496.

[14] Kotz, S. and Johnson, N. L. (1977). Propriétés de dépendence de distributions itérées, généralisées á deux variables Farlie-Gumbel-Morgenstern, *Comptes rendues, Acad. Sc. Paris*, **285**, Séries A, 277-280.

[15] Lehmann, E. L. (1966). On concepts of dependence. *Ann. Math. Stat.*, **37**, 1137-1153.

[16] Morgenstern, D. (1956). Einfache Beispiele zweidimensionaler Verteilung, *Mitteislingsblatt für Mathematische Statistik*, **8**, 234-235.

ÖZET

Marjinalleri $[0,1]$ aralığında düzgün dağılım fonksiyonu olan iki değişkenli FGM dağılımları incelenmiştir. İlk bileşenin r-inci sıra istatistiğinin eşinin dağılımı elde edilmiştir. Eşlerin momentleri arasındaki indirgeme bağıntıları verilmiştir.