İstatistik, Journal of the Turkish Statistical Association
1998, Vol.1, No. 2, pp. 1 - 8

A SIMPLIFIED SOLUTION OF THE CENTRAL LIMIT PROBLEM IN HILBERT SPACE*

V.V. Sazonov
Steklov Mathematical Institute
Moscow, Russia
sazonov@genesis.mi.ras.ru

Abstract

A new simpler solution of the Central Limit Problem for rowwise independent arrays of a Hilbert space valued random variable is given. By its form this solution is a direct generalization of the classical solution of the Central Limit Problem for real random variables.

AMS 1990 Subject Classification: 60B12, 60F05.

Key Words: Central limit problem, uniform asymptotic negligibility, Hilbert Space.

1. Introduction

The aim of the present paper is to give a more simple and close (noun) to the classical in form solution of the Central Limit Problem for arrays of uniformly asymptotically negligible (uan) rowwise independent random variables (r.v.'s) with values in a Hilbert space. The pioneering work in solving this problem in a Hilbert space setting was made by Varadhan (1962) (see also a more accessible book [2], containing a complete exposition of Varadhan's results in [1], which will be used below for references). Later Gihman and Skorohod (1980) gave a solution which is closer to the classical one (for a more detailed proof of their result see also [4]). A drawback of the Gihman-Skorohod solution is a complicated and not constructive choice of normalizing shifts. In what follows we show that these shifts can be chosen by a direct generalization of what is usually done in the classical case of real random variables. This leads to what may be called a proper solution of the Central Limit Problem in a Hilbert space. The proof we present here is based on the important information

*Written while visiting the Center for Stochastic Processes, University of North Carolina with support from Army Research Office Grant No. DAAH04 95 1 0422

already contained in Varadan's compactness type results and this is why it is rather short. A direct (but considerably longer) proof like in [3] (or [4]) can also be given.

2. Notation and the Main Result

Let H be a real separable Hilbert space with inner product $(.,.)$ and norm $|.|$. Measurability in H will be understood as the Borel measurability and all measures in H considered below will be Borel measures, i.e. defined on the σ-algebra $\mathcal{B}$ of Borel subsets of H. $\mathcal{P}$ and $\mathcal{M}$ will denote the set of all probability measures and all finite measures in H respectively, $\mathcal{M}^0 = \{\mu \in \mathcal{M} : \mu(\{0\}) = 0\}$. Furthermore, $\mathcal{S}$ will stand for the set of all linear bounded nonnegative symmetric operators S with finite trace trS in H. Recall that a sequence $(S_n)_{n\geq 1}$ in $\mathcal{S}$ is called compact if

$$\sup_n trS_n < \infty \qquad \text{and} \tag{2.1}$$

$$\lim_{N\to\infty} \sup_n \sum_{j=N}^{\infty} (S_n e_j, e_j) = 0 \tag{2.2}$$

where $(e_j)_{j\geq 1}$ is a complete orthonormal basis in H (actually if (2.2) is true in some basis then it is true in any basis in H). Recall also that a P in $\mathcal{P}$ is infinitely divisible if and only if its characteristic function

$\hat{P}(y) = \int_H \exp\{i(x,y)\}P(dx)$, $y \in H$ has the form

$$\hat{P}(y) = \exp\{i(a,y) - \frac{1}{2}(Sy,y) + \int_{H^0} K(x,y)\mu(dx)\} := \psi(a,S,\mu;y),$$

where $a \in H, S \in \mathcal{S}, H^0 = H\backslash\{0\}$,

$$K(x,y) = \left(e^{i(x,y)} - 1 - \frac{i(x,y)}{1+|x|^2}\right)\frac{1+|x|^2}{|x|^2}$$

and $\mu \in \mu^0$. This representation of $\hat{P}$ is unique, i.e. a, S and μ are defined uniquely (see [3], Ch. VI, §3; an equivalent form of this formula was obtained earlier in [1]).

Let $\{X_{nk} : n = 1,2,\ldots,, k = 1,\ldots,k_n\}$ be an array of H-valued rowwise independent random variables satisfying the uan (uniform asymptotic negligibility) condition: for any $\epsilon > 0$

$$\lim_{n\to\infty} \max_{1\leq k\leq k_n} P(|X_{nk}| \geq \epsilon) = \lim_{n\to\infty} \max_{1\leq k\leq k_n} P_{nk}(B_\epsilon^c) = 0,$$

where P_{nk} is the distribution of X_{nk}, $B_\epsilon = \{x \in H : |x| < \epsilon\}$ and E^c is the complement of a set E. Fix a $\tau > 0$ and denote

$$a_{nk} = \int_{B_\tau} xP_{nk}(dx), a'_n = \sum_{k=1}^{k_n} a_{nk}.$$

Furthermore, denote the distribution of $X_{nk} - a_{nk}$ by Q_{nk} and define

$$a_n = a' + \sum_{k=1}^{k_n} \int_H \frac{x}{1+|x|^2} Q_{nk}(dx);$$

also define S_{nk}, s_n in $\mathcal{S}$ by

$$(S_{nk}y, y) = \int_H \frac{(x,y)^2}{1+|x|^2} Q_{nk}(dx), \quad S_n = \sum_{k=1}^{k_n} S_{nk}(y \in H),$$

and μ_n in $\mathcal{M}$ as

$$\mu_n(A) = \sum_{k=1}^{k_n} \int_A \frac{|x|^2}{1+|x|^2} Q_{nk}(dx) \quad (A \in \mathcal{B}).$$

Finally let $P_n = \prod_{k=1}^{k_n} P_{nk}$ (in the sense of convolution) be the distribution of $\sum_{k=1}^{k_n} X_{nk}$, $n = 1, 2, \ldots$ and $\delta_u, u \in H, \delta_u \in \mathcal{P}$ be such that $\delta_u(\{u\}) = 1$.

Theorem 2.1 The sequence $(P_n)_{n\geq 1}$ converges weakly in $\mathcal{P}$ if and only if

(i) $|a_n - a| \to 0$ as $n \to \infty$ for an a in H;

(ii) $(S_n y, y) \to (S'y, y)$, as $n \to \infty$, for an S' in $\mathcal{S}$ and all $y \in H$, and $(S_n)_{n\geq 1}$ is compact;

(iii) $(\mu_n)_{n\geq 1}$ converges weakly to some μ' in $\mathcal{M}$.

Moreover, if conditions (i) - (iii) are satisfied, then the limiting distribution P of $(P_n)_{n\geq 1}$ is infinitely divisible with the characteristic function

$$\hat{P}(y) = \psi(a, S, \mu; y) \quad (y \in H),$$

where S is defined by

$$(Sy, y) = (S'y, y) - \int_{H^0} \frac{(x,y)^2}{|x|^2} \mu(dx) \quad (y \in H)$$

and $\mu = \mu' - \mu'(\{0\})\delta_0$.

3. Proof

Sufficiency. We will divide the proof into a few steps

1. The uan condition implies that for any $T > 0$

$$\lim_{n\to\infty} \sup_{|y|\leq T} \max_{1\leq k\leq k_n} |\hat{P}_{nk}(y) - 1| = 0 \tag{3.1}$$

(see [2], p. 190 or [4], Lemma 7.8.1), and also

$$\lim_{n\to\infty} \max_{1\leq k\leq k_n} |a_{nk}| = 0 \tag{3.2}$$

([2], Ch. VI, Lemma 6.1). From (3.2) it easily follows that an array of random variables with distributions $Q_{nk}(\cdot) = P_{nk}(\cdot + a_{nk})$ is also uan and hence (cf. (3.1)).

$$\lim_{n\to\infty} \sup_{|y|\le T} \max_{1\le k\le k_n} |\hat{Q}_{nk}(y) - 1| = 0. \tag{3.3}$$

2. For any $T > 0$ and all large $n (n \ge n_0$, where n_0 depends only on τ and on the array (P_{nk}) under consideration)

$$\sup_{|y|\le T} |\hat{Q}_{nk}(y) - 1| \le c(\tau, T) \int_H \frac{|x|^2}{1+|x|^2} Q_{nk}(dx). \tag{3.4}$$

Indeed, for $|y| \le T$ we can write (omitting the subscripts)

$$\left|\hat{Q}(y) - 1\right| \le \left|\int_{B_\tau - a} (e^{i(x,y)} - 1 - i(x,y)) Q(dx)\right|$$

$$+ \left|\int_{B_\tau - a} (x,y) Q(dx)\right| + 2Q(B_\tau^c - a) = J_1 + J_2 + J_3$$

Furthermore, since $|a| \le \tau$,

$$J_1 \le \frac{T^2}{2} \int_{B_\tau - a} |x|^2 \, Q(dx)$$

$$\le \frac{T^2}{2}(1 + 4\tau^2) \int_{B_\tau - a} \frac{|x|^2}{1+|x|^2} Q(dx)$$

and also

$$Q(B_\tau^c \le a) \le \frac{1 + 4\tau^2}{(\tau - |a|)^2} \int_{B_\tau^c - a} \frac{|x|^2}{1+|x|^2} Q(dx),$$

so that

$$J_2 = \left|\int_{B_\tau} (x - a, y) P(dx)\right|$$

$$\le \left|(a,y) - (a,y) \int_{B_\tau} P(dx)\right|$$

$$\le \tau T Q(B_\tau^c - a).$$

$$\text{Thus } J_2 + J_3 \le (2 + \tau T) Q(B_\tau^c - a)$$

$$\le C(\tau, T) \int_{B_\tau^c - a} \frac{|x|^2}{1+|x|^2} Q(dx).$$

The above inequalities together with (3.2) imply (3.4).

3. Observe that (3.4) together with condition (ii) of the theorem gives

$$\sup_{n\ge n_0} \sup_{|y|\le T} \sum_{k=1}^{k_n} \left|\hat{Q}_{nk}(y) - 1\right| \le c(\tau, T) \sup_{n\ge 1} \sum_{k=1}^{k_n} \int_H \frac{|x|^2}{1+|x|^2} Q_{nk}(dx)$$

$$= c(\tau, T) tr S_n < \infty. \tag{3.5}$$

Furthermore, (3.3) implies that for all $y : |y| \le T$ and all large enough n $\log \hat{Q}_{nk}(y)$ is well defined and using (3.5) we have

$$\sup_{|y|\le T} \left| \sum_{k=1}^{k_n} (\log \hat{Q}_{nk}(y) - (\hat{Q}_{nk}(y) - 1)) \right| \le \sup_{|y|\le T} \sum_{k=1}^{k_n} \left| \hat{Q}_{nk}(y) - 1 \right|^2$$

$$\le \sup_{|y|\le T} \max_{1\le k\le k_n} |\hat{Q}_{nk}(y) - 1| \sum_{k=1}^{k_n} \left| \hat{Q}_{nk}(y) - 1 \right| \to 0 \text{ as } \quad n \to \infty. \tag{3.6}$$

4. Since $\hat{P}_n(y) = \exp\{i(a_n', y)\} \prod_{k=1}^{k_n} \hat{Q}_{nk}(y)$, we can write

$$\log \hat{P}_n(y) = i(a_n', y) + \sum_{k=1}^{k_n} (\hat{Q}_{nk}(y) - 1) + R_n(y),$$

where $R_n(y) = \sum_{k=1}^{k_n} (\log \hat{Q}_{nk}(y) - (\hat{Q}_{nk}(y) - 1))$. Moreover, in the notation of the theorem we have

$$i(a_n', y) + \sum_{k=1}^{k_n} (\hat{Q}_{nk}(y) - 1)$$

$$= i(a_n, y) - \frac{1}{2}(S_n y, y) + \int_H \left(K(x, y) + \frac{1}{2} \frac{(x, y)^2}{|x|^2} \right) \mu_n(dx). \tag{3.7}$$

Note that the integrand in the right hand side here is a bounded continuous function if we define it as zero at the origin. Hence (3.6) and conditions (i) - (iii) of the theorem imply

$$\lim_{n\to\infty} \hat{P}_n(y) = \psi(a, S, \mu; y)$$

uniformly over $y : |y| \le T$ for any $T > 0$. Note that S is indeed in $\mathcal{S}$. To show this it is enough to prove that it is nonnegative, the other properties being obvious. But if $\epsilon > 0$ is such that $\mu'\{x : |x = \epsilon\} = 0$, then

$$\int_{H^0} \frac{(x, y)^2}{|x|^2} \mu'(dx) = \int_{B_\epsilon \setminus \{0\}} + \int_{B_\epsilon^c}$$

$$\le |y|^2 \mu'(B_\epsilon \setminus \{0\}) + \lim_{n\to\infty} \int_{B_\epsilon^c} \frac{(x, y)}{1 + |x|^2} Q_{nk}(dx)$$

$$\le |y|^2 \mu'(B_\epsilon \setminus \{0\}) + (S' y, y),$$

and hence, by the definition of S

$$(Sy, y) \ge -|y|^2 \mu'(B_\epsilon \setminus \{0\});$$

it remains to let $\epsilon \to 0$.

5. To finish the proof of the sufficiency it is enough to prove now that the sequence $(P_n)_{\geq 1}$ is shift compact (see [2], Ch. VI, Theorem 4.5). Note that the exp of (3.7) is the characteristic function of the infinitely divisible distribution

$$I_n = \prod_{k=1}^{k_n} e(Q_{nk}) * \delta_{a'_n}. \tag{3.8}$$

where

$$e(Q_{nk}) = \delta_0 + Q_{nk} + \frac{Q_{nk}^{*2}}{2} + \dots.$$

By Theorem 6.2, Ch. VI in [2] it is enough thus to prove the shift compactness of $(I_n)_{n\geq 1}$. This, according to Theorem 5.2 and Remark after Theorem 5.3, Ch. VI in [2], will be proved if we prove that

(a) the restriction of $\mu'_n = \sum_{k=1}^{k_n} Q_{nk}$ to the complement of any neighbourhood of the origin is weakly conditionally compact;

(b) the sequence of operators $(S'_n)_{n\geq 1}$ in $\mathcal{S}$ defined by

$$(S'_n y, y) = \int_{B_t} (x,y)^2 \mu'_n(dx), \quad t > 0, \tag{3.9}$$

is compact.

But by (iii) $(\mu_n)_{n\geq 1}$ converges weakly, hence is tight and uniformly bounded (in the sence that $\mu_n(H)$, $n \geq 1$, are uniformly bounded). Then of course its restriction as well as the restriction of $(\mu'_n)_{n\geq 1} \geq 1 \quad (d\mu'_n = (1+|x|^2)/|x|^2) d\mu_n)$ to the complement of any neighbourhood of the origin is also tight and uniformly bounded and hence weakly conditionally compact. This proves (a). To prove (b) observe that

$$(S'_n y, y) \leq 2 \int_{B_1} \frac{(x,y)^2}{1+|x|^2} \mu'_n(dx) \leq 2(S_n y, y),$$

and it remains only to apply condition (ii) of the theorem.

Necessity. By Theorem 4.5 and 6.2, Ch. VI in [2] the convergence of $(P_n)_{n\geq 1}$ to a limit P implies the convergence of $(I_n)_{n\geq 1}$, defined in (3.8), to the same P. The characteristic function of I_n may be written as

$$\hat{I}_n(y) = \exp\{i(a_n, y) + \int_H K'(x,y) \mu'_n(dx)\},$$

where

$$K'(x,y) = e^{i(x,y)} - 1 - i(x,y)/(1+|x|^2).$$

Applying Theorem 5.2 and Remark after Theorem 5.3, Ch. V in [2] we see that

1. the restrictions of $\mu'_n, n \geq 1$, to the complement of any neighbourhood of the origin is weakly ocnditionally compact and hence the same is true for $\mu_n, n \geq 1$, since $d\mu_n = (|x|^2/(1+|x|^2)) d\mu'_n$,

2. the sequence $(S'_n)_{n\geq 1}$ of operators defined in (3.9) is compact.

It follows that

(a) $(\mu_n)_{n\geq 1}$ is weakly conditionally compact. Indeed

$$\mu_n(H) = \mu_n(B_1) + \mu_n(B_1^c)$$

$$\leq \int_{B_1} |x|^2 \, \mu_n'(dx) + \mu_n'(B_1^c)$$

$$= trS_n' + \mu_n'(B_1^c),$$

so that $\sup_{n\geq 1} \mu_n(H) < \infty$, and together with Lemma 6.3, Ch. VI in [2], this implies, by Prohorov's theorem,the weak conditional compactness of $(\mu_n)_{n\geq 1}$;

(b) $(S_n)_{n\geq 1}$ is compact. Indeed, $trS_n = \mu_n(H)$, so that $\sup_{n\geq 1} trS_n < \infty$. Furthermore, if $(e_j)_{j\geq 1}$ is an orthonormal basis in H and for an x in H $|x|_N^2 = \sum_{j=N}^{\infty}(x, e_j)^2$, then for any $t > 0$

$$\sum_{i=N}^{\infty}(S_n e_i, e_i) = \int_H \frac{|x|_N^2}{1+|x|^2}\mu_n'(dx) = \int_{B_t} + \int_{B_t^c}$$

$$\leq \int_{B_t} |x|_N^2 \, \mu_n'(dx) + \mu_n(B_t^c). \tag{3.10}$$

Taking t large enough we can make the second summand in (3.10) arbitrary small uniformly in $n \geq 1$ (by the tightness of $(\mu_n)_{n\geq 1}$) and for a fixed $t > 0$ the first summand in (3.10) is arbitrary small for all large N and all $n \geq 1$ (by the compactness of $(S_n')_{n\geq 1}$).

Represent now $\hat{I}_n(y)$ as

$$\hat{I}_n(y) = \exp\left\{i(a_n, y) - \frac{1}{2}(S_n y, y) + \int_H \left(K(x,y) + \frac{1}{2}\frac{(x,y)^2}{|x|^2}\right)\mu_n(dx)\right\} \tag{3.11}$$

and choose a subsequence (n') of $(1, 2, \ldots)$ such that along this subsequence $\mu_{n'}$ converges weakly to a $\overline{\mu}'$ in $\mathcal{P}$ and $(S_{n'}y, y) \to (\overline{S}'y, y), y \in H$, where $\overline{S}$ is in $\mathcal{S}$ (the existence of such a $\overline{S}'$ follows e.g. from the compactness of $(S_n)_{n\geq 1}$ and Lemma 5.1, Ch. VI in [2]). Letting in (3.11) is bounded and continuous (when defined as zero at the origin), we see that the limit of $\exp\{i(a_{n'}, y)\}$ exists for all $y \in H$ and is continuous at the origin. This implies that (a_n, y) converges for all y in H, and since H is weakly complete there is an $\overline{a}$ in H such that $(a_{n'}, y) \to (\overline{a}, y)$ for all y in H. Thus in the limit as $n' \to \infty$ (3.11) becomes

$$\hat{P}(y) = \exp\{i(\overline{a}, y) - \frac{1}{2}((\overline{S}'y, y) - \int_{H^0} \frac{(x,y)^2}{|x|^2}\overline{\mu}'(dx)) + \int_{H^0} K(x,y)\overline{\mu}'(dx)\}.$$

This shows that P is infinitely divisible and from the unicity of representation of the characteristic function of an infinitely divisible distribution it follows that $a_n \to a = \overline{a}$ weakly, $(S_n y, y) \to (S'y, y), y \in H$, with $S' = \overline{S}'$, and for any two limiting points

$\overline{\mu}', \widetilde{\mu}'$ of $(\mu_n)_{n\geq 1}$ we have $\overline{\mu}' - \overline{\mu}(\{0\})\delta_0 = \widetilde{\mu}' - \widetilde{\mu}'(\{0\})\delta_0$. The sequence $(\mu_n)_{n\geq 1}$ indeed converges to $\mu' = \overline{\mu}'$. To prove this it is enough now to prove that $\mu_n(H)$ converges. But for any positive integer N

$$|\mu_n(H) - \mu_m(H)| = |trS_n - trS_m| \leq \sum_{j=1}^{N-1} |(S_n e_j, e_j) - (S'_n e_j, e_j)|$$

$$+ \sum_{j=N}^{\infty} ((S_n e_j, e_j) + (S_n e_j, e_j)) \tag{3.12}$$

and the compactness of $(S_n)_{n\geq 1}$ and the convergence $(S_n y, y) \to (S'y, y), y \in H$ imply that $(\mu_n(H))_{n\geq 1}$ indeed converges.

To finish the proof it remains only to show that $a_n \to a$ in norm. Observe that for any $T > 0$ uniformly in $y : |y| \leq T$

$$\widehat{I}_n(y) \to \widehat{I}(y) = \widehat{P}(y) \tag{3.13}$$

$$(S_n y, y) \to (S'y, y) \tag{3.14}$$

$$\int_H \left(K(x,y) + \frac{1}{2}\frac{(x,y)^2}{|x|^2} \right) \mu_n(dx) \to \int_H \left(K(x,y) + \frac{1}{2}\frac{(x,y)^2}{|x|^2} \right) \mu'(dx). \tag{3.15}$$

Indeed (3.13) is a general property (see e.g. Theorem 4.4, Ch. VI in [2]), (3.14) follows easily (cf. (3.12) from the convergence of $(S_n y, y)$ to $(S'y, y)$ for each y in H and the compactness of $(S_n)_{n\geq 1}$, and (3.15) is a corollary of the uniform boundedness and equicontinuity at all x in H of the integrands when $|y| \leq T$ (see Theorem 6.8, Ch. II in [2]). Now (3.13) - (3.15) together with (3.11) imply $(a_n, y) \to (a, y)$ as $n \to \infty$ uniformly in $y : |y| \leq T$, and this finishes the proof of the theorem.

References

[1] Varadhan, S.R.S. (1962) Limit theorems for sums of independent random variables with values in a Hilbert space. *Sankhya* Vol. 24, 213-238.

[2] Parthasarathy, K.R. (1967) Probability measures on metric spaces. Academic Press, New York.

[3] Gihman, I.I. and Skorohod, A.V. (1980) The theory of stochastic processes I. Springer-Verlag, Berlin.

[4] Laha, R.G. and Rohatgi, V.K. (1979) Probability theory. Wiley, New York.

ÖZET

Hilbert uzayındaki bağımsız satır rasgele değişkenleri için Merkezi Limit Probleminin daha basit bir .özümü verilmiştir. Çözümün bu formu reel değerli rasgele değişkenler için verilen çözümün direkt olarak genelleştirilmesidir.

İstatistik, Journal of the Turkish Statistical Association
1998, Vol.1, No. 2, pp. 9 - 21

CHANGE-POINT DETECTION FOR DEPENDENT DATA AND APPLICATION TO HYDROLOGY

Daniela Jaruškova
Department of Mathematics,
Czech Technical University, Faculty of Civil Engineering
Thàrukova 7. CZ-166 29 Praha 6, Czech Republik
E-mail: jarus@mbox.cesnet.cz

Abstract

Scientists are afraid that the impact of human activity on nature may cause a change in nature. This apprehension precipitated the launching of the world project to study important hydrological and meteorological series. The contribution of the Czech Republic consisted in two projects directed by the researchers of the Czech Hydrometeorological Institute. Statistical methods that have been applied to decide whether hydrological and meteorological are stationary are called change-point methods. Applying these methods we realized that the dependence between neighboring observations affects considerably results of statistical tests. The paper uses theoretical results and results obtained by simulations for suggestions how to adapt procedures for i.i.d. random variables to the dependent observations.

Key Words: Change-point detection, statistical tests, time series

1. Introduction

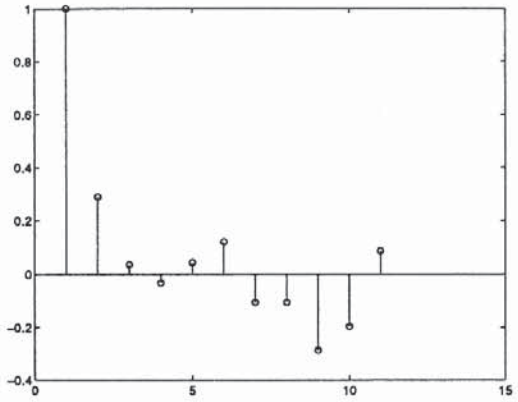
Few years ago I cooperated with researchers of the Czech Hydrometeorological Institute. In the scope of two projects

- *Analysis of long hydrometeorological series*
- *Analysis of hydrological observations in the Czech Republic*

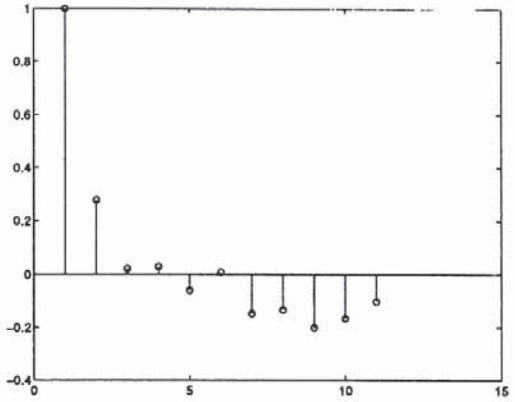
we studied hydrological and meteorological series to decide whether they can be supposed to be stationary or whether some changes can be discovered that might be caused by human activity.

Before applying statistical methods to make such decisions statistical properties of the studied series have to be established. One of the most typical features of hydrological and meteorological data is the dependence between neighboring observations

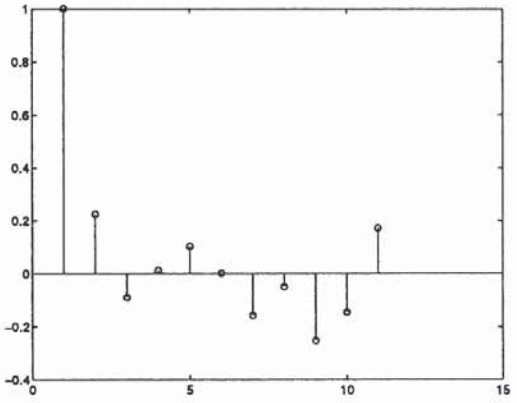
caused by a certain persistance in the behavior of nature. Monthly averages usually exhibit strong dependence but annual averages may be also dependent, see the autocorrelation functions of the water discharges of several Czech rivers in Figure 1. The dependence has a great impact on decision whether a series is stationary or not.



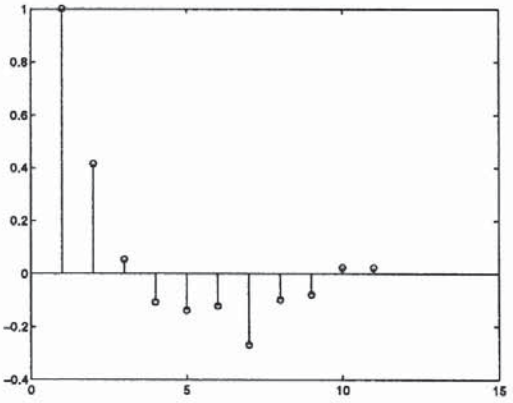
ACF of Svitava (Bílovice), average runoff 5.2 m^3/s



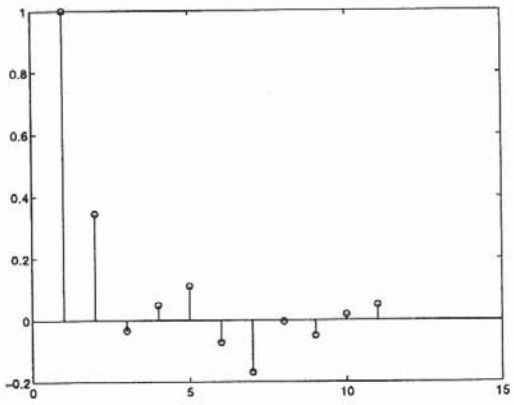
ACF of Lužnice (Bechyně), average runoff 23.7 m^3/s



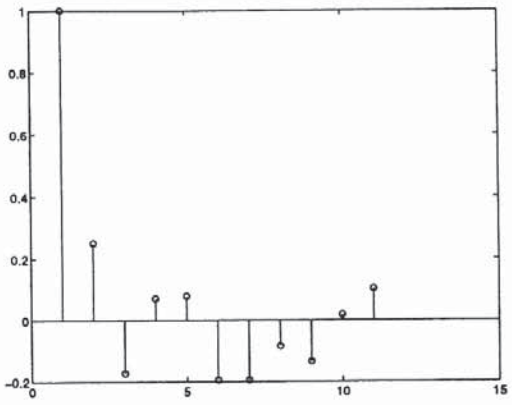
ACF of Morava (Kroměříž), average runoff 52.0 m^3/s



ACF of Otava (Písek), average runoff 2.9 m^3/s



ACF of Metuje (Hronov),
average runoff 23.4 m^3/s



ACF of Jizera (Vilémov),
average runoff 4.9 m^3/s

Completed dams or water reservoirs may cause an abrupt change in the behavior of a series. In that case the time of change as well as the extent of change is usually known. The increase of human economic activity changes series gradually. The detection of gradual change (almost unobservable at the beginning but more and more apparent later) is very important.

2. Mathematical Formulation

In the scope of mathematical statistics the decision whether a series changed is usually based on hypothesis testing. In the case that we expect an abrupt change we test the null hypothesis H_0 against the alternative A:

$$\begin{array}{lll} H_0: & X_i = \mu + e_i, & i = 1, ..., n, \\ A: & \exists\, k \in \{0, ..., n-1\} \quad \text{such that} & \\ & X_i = \mu_1 + e_i, & i = 1, ..., k, \\ & X_i = \mu_2 + e_i, & i = k+1, ..., n, \quad \mu_1 \neq \mu_2. \end{array} \tag{1}$$

The errors $\{e_i\}$ are supposed here to be i.i.d. with $Ee_i = 0, Ee_i^2 = \sigma_e^2$ and $E|e_i|^{2+\delta} < \infty$ for some $\delta > 0$. Supposing σ^2 is unknown the most frequently applied test statistic $T(n)$ is the maximum of the absolute values of two sample $t-$test statistics

$$T(n) = \max_{1\le k<n} |T_k| = \max_{1\le k<n} \sqrt{\left(\frac{(n-k)k}{n}\right)} |\bar{X}_k - \bar{X}_k^*| \cdot \frac{1}{s_k}, \tag{2}$$

where

$$\bar{X}_k = \frac{\sum_{j=1}^{k} X_j}{k}, \qquad \bar{X}_k^* = \frac{\sum_{j=k+1}^{n} X_j}{n-k},$$

$$s_k = \sqrt{(\sum (X_i - \bar{X}_k)^2 + \sum (X_i - \bar{X}_k^*)^2)/(n-2)}.$$

For n large critical values may be calculated using the asymptotic distribution of $T(n)$ derived by Yao and Davis (1986):

$$\lim_{n\to\infty} P\left(\frac{T(n) - b_n}{a_n} > x\right) = 1 - \exp\left(\frac{-2\,e^{-x}}{\sqrt{\pi}}\right), \tag{3}$$

where $a_n = (2\log\log n)^{-1/2}$ and $b_n = a_n^{-1} + (a_n/2) \cdot \log\log\log n$.

In the case we expect that the series might change gradually and after the change point it increases (or decreases) linearly we test the null hypothesis H_0 against the alternative A:

$$\begin{array}{lll} H_0: & X_i = \mu + e_i, & i = 1, ..., n, \\ A: & \exists\, k \in \{0, ..., n-1\} \quad \text{such that} & \\ & X_i = \mu + e_i, & i = 1, ..., k, \\ & X_i = \mu + b \cdot (i - k) + e_i, & i = k+1, ..., n, \quad b \neq 0. \end{array} \tag{4}$$

For unknown σ^2 the test statistic has the form

$$\tilde{T}(n) = \max_{1\le k<n} |\tilde{T}_k| = \max_{1\le k<n} \frac{|\tilde{U}_k|}{\sqrt{RSS - \tilde{U}_k^2}} \sqrt{n-2} \tag{5}$$

where

$$\tilde{U}_k = \frac{\sum_{i=k+1}^{n} (X_i - \bar{X})(i-k)}{\sqrt{\frac{(n-k)(n-k+1)(2n-2k+1)}{6} - \frac{(n-k)^2(n-k+1)^2}{4n}}}, \qquad k = 1, \ldots, n-1$$

and RSS is the residual sum of squares under the null hypothesis. For n large critical values may be calculated using the asymptotic distribution

$$\lim_{n\to\infty} P\left(\frac{\tilde{T}(n) - \tilde{b}_n}{\tilde{a}_n} > x\right) = 1 - \exp\left(\frac{-\sqrt{3}e^{-x}}{2\pi}\right), \tag{6}$$

where $\tilde{a}_n = (2\log\log n)^{-1/2}$ and $\tilde{b}_n = (2\log\log n)^{1/2}$. It is well known that the convergence in (2) and (6) is slow. For moderate values of n we recommend to use the critical values that we obtained by simulations:

n	5% critical values	1% critical values
50	3.15	3.76
100	3.16	3.71
200	3.19	3.72
300	3.21	3.73
500	3.24	3.73

Table 1. Several examples of 5 % and 1 % critical values of the statistic $T(n)$.

n	5% critical values	1% critical values
50	2.62	3.27
100	2.63	3.21
200	2.65	3.22
300	2.65	3.22
500	2.68	3.22

Table 2. Several examples of 5 % and 1 % critical values of the statistic $\tilde{T}(n)$.

3. Effect of dependence

Many authors, e.g. Antoch, Hušková and Prášková (1997), Bai (1993), Kim (1995), Tang and Mac Neil (1993), studied the effect of dependence on decision whether a change in a series occurred or did not occur. The typical behavior of a process, where the neighboring observations are positively correlated, is such that it moves slowly from one level to another level. Therefore, the change must be more apparent to be detected.

To get some information how the dependence affects the distribution of $\tilde{T}(n)$ under H_0 we prepared a simulation study. For $n = 80, 120, 200$ and $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$ we simulated 100 000 realizations of an AR(1) sequence with the first autoregressive coefficient ρ and errors distributed according to $N(0,1)$. Figure 2 shows the sample quantile function for $\rho = 0, 0.3, 0.6$ and for $n = 200$. The 95 % quantile of the statistic $\tilde{T}(n)$ for i.i.d. random variables corresponds approximately to 81.5 % quantile of $\tilde{T}(n)$ for AR(1) sequence with $\rho = 0.3$ and to 54 % quantile for AR(1) sequence with $\rho = 0.6$. It means for example that 46 % of realizations of statistic $\tilde{T}(n)$, if the observations $\{X_i\}$ form an AR(1) sequence with $\rho = 0.6$, exceed the 5 % critical value (95 % quantile) of $\tilde{T}(n)$ supposing $\{X_i\}$ are independent. It indicates that the dependence affects enormously the "rejection-acceptance" decision.

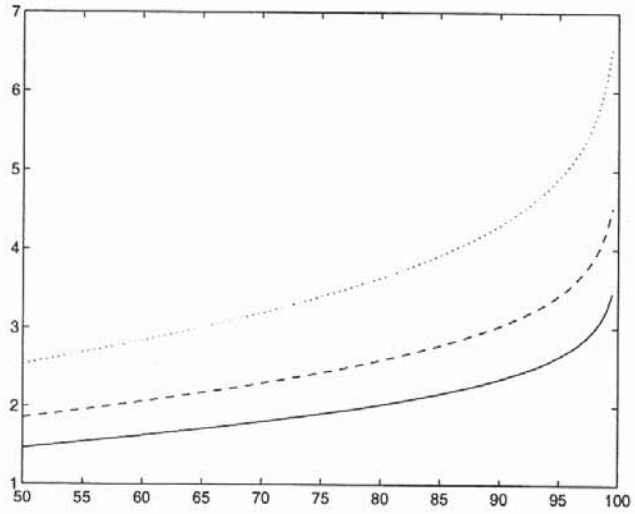


Figure 2. Quantile functions of $\tilde{T}(n)$ for i.i.d. random variables (solid line) and for AR(1) sequences with $\rho = 0.3$ (dashed line) and $\rho = 0.6$ (dotted line).

Table 3 and 4 show critical values of $\tilde{T}(n)$, n=80,120,200, obtained by our simulations.

	ϱ						
n	0.0	0.1	0.2	0.3	0.4	0.5	0.6
80	2.63	2.83	3.08	3.39	3.76	4.23	4.88
120	2.63	2.84	3.09	3.40	3.78	4.25	4.88
200	2.64	2.85	3.11	3.42	3.80	4.27	4.89

Table 3. 5 % critical values of $\tilde{T}(n)$.

	ϱ						
n	0.0	0.1	0.2	0.3	0.4	0.5	0.6
80	3.24	3.50	3.84	4.24	4.73	5.34	6.21
120	3.22	3.48	3.81	4.20	4.68	5.29	6.11
200	3.23	3.50	3.83	4.22	4.70	5.31	6.13

Table 4. 1 % critical values of $\tilde{T}(n)$.

Antoch, Hušková and Prášková (1997) studied the model (1) with errors $\{e_t\}$ forming a linear process

$$e_t = \sum_{j=0}^{\infty} w_j \varepsilon_{t-j}, \tag{7}$$

where $\{\varepsilon_t\}$ are i.i.d. such that $E\varepsilon_t = 0, E\varepsilon_t^2 = \sigma_\varepsilon^2$ and $E|\varepsilon_t|^{2+\delta} < \infty$ for some $\delta > 0$ and $\{w_j\}$ are real constants satisfying $\sum w_j \neq 0, \sum w_j^2 < \infty$ and moreover

$\sum j|w_j| < \infty$. They showed that for such a linear process the critical values (obtained from (3) for i.i.d. random variables) have to be multiplied by $\sqrt{A}$ where

$$A = \frac{2\pi h(0)}{\gamma} = \frac{(\sum w_j)^2}{\sum w_j^2}, \tag{8}$$

$h(\cdot)$ denotes spectral density of the process $\{e_t\}$ and $\gamma = Var\, e_t$. Similarly, for the linear process (7) such that $\sum j^2|w_j| < \infty$ we have proved that critical values of $\tilde{T}(n)$ obtained by (6) have to be adapted in the same way. Notice that stationary ARMA sequences are linear processes satisfying the upper conditions. If for example the process $\{e_t\}$ forms AR(1) sequence, the critical values should be multiplied by $\sqrt{(1+\rho)/(1-\rho)}$, where ρ is the first autoregressive coefficient. Our simulation study showed that the critical values for independent random variables multiplied by $\sqrt{(1+\rho)/(1-\rho)}$ slightly overestimate the corresponding critical values obtained by simulations. In the worst case ($\rho = 0.6$) the 5% critical value for i.i.d. multiplied by $\sqrt{(1+\rho)/(1-\rho)}$ corresponds to the 3% critical value obtained by simulations and the 1% critical value for i.i.d. multiplied by $\sqrt{(1+\rho)/(1-\rho)}$ corresponds to the 0.5% critical value obtained by simulations. For practical purposes it is important to notice that all critical values change as n increases only very slowly. It enables to make interpolations and extrapolations to get critical values which are not listed in our tables.

Sometimes, the value of $A = 2\pi h(0)/\gamma$ may be assessed from an experience with the series similar to the series under study due to the similar environmental conditions. But sometimes it has to be estimated. If we are sure that a certain portion of series is stationary we can use this part of the series to estimate A. However, we have to bear in mind that if in reality the part of the series used for estimation is not stationary then changes affect the sample autocorrelation function and the estimator of A considerably. Antoch, Hušková and Prášková (1997) suggested to use the nonparametric estimator of A:

$$\hat{A} = \hat{\rho}(0) + 2\sum_{k=1}^{L}(1 - \frac{k}{L})\hat{\rho}(k), \quad L << n, \tag{9}$$

where $\hat{\rho}(i)$ denotes the value of sample autocorrelation function for the lag i. Supposing that the process $\{e_t\}$ forms an ARMA sequence, the estimators of the parameters of the ARMA model provide us with the parametric estimator of A.

4. Examples

The first example describes the decrease of precipitations in the North-West Africa. Figure 3 presents the annual rainfall departures in Sahel (1901 – 1990) constructed by Nicolson (1994). The sequence $\{\widehat{T}_k, k = 1, \ldots, 89\}$ is displayed in Figure 4.

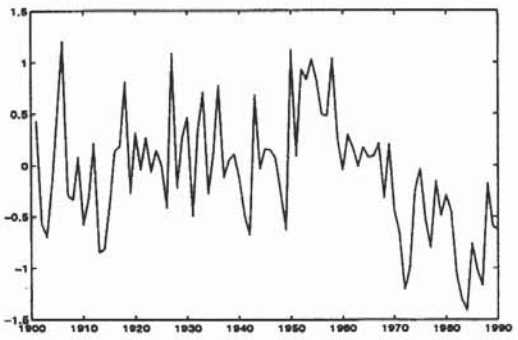


Figure 3. Annual rainfall departures in Sahel 1901 – 1990.

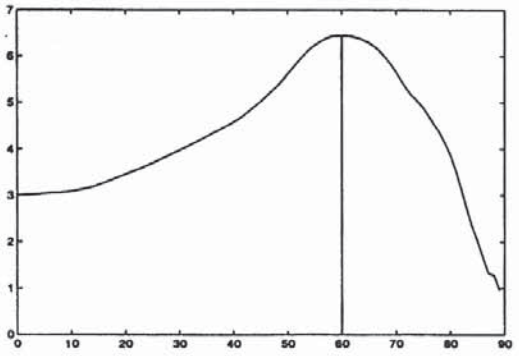


Figure 4. Sequence $\{\widehat{T}_k\}$ for annual rainfall departures in Sahel.

The testing statistic $\widehat{T}(n)$ attains the value 6.44 and exceeds clearly the 1 % critical value for i.i.d. random variables. The null hypothesis that the series is stationary is rejected and the change is detected in 1960. Figure 5 shows the data with the fitted model. The autocorrelation function of the residuals indicates a slight dependence between observations, see Figure 6. The non-parametric estimator of A computed of the residuals (using (9) with $L = 12$) equals approximately 1.7.

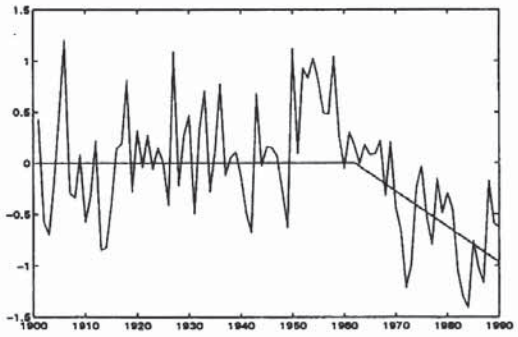


Figure 5. Annual rainfall departures in Sahelwith the fitted model.

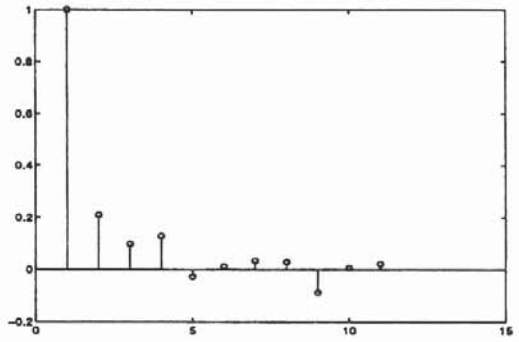


Figure 6. Autocorrelation function of residuals.

As the dependence is slight, our decision to reject the null hypothesis remains without any change. We would like to remark that the value of $\widehat{T}(n)$

is so great that we would not change our decision even for more strongly dependent variables.

The second series that serves as an example how to apply the suggested method presents the decrease of the width of ozone layer that protects the Earth from too much radiation. The example shows that the over-all tendency in behavior of the series is more apparent if we deal with annual averages. On the other side by studying monthly averages we may get a more detailed insight into character of changes of our data. However, the applied statistical inference has to take into account the dependence of the observations that is usually much stronger than in the case of the annual averages.

Figure 7 shows the annual averages of total amount of ozone in D.U. measured in Hradec Králové (Czech Republic) in years 1962-1995.

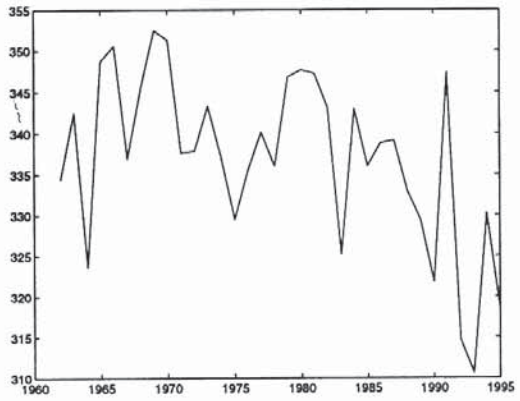


Figure 7. Annual averages of ozone in D.U. measured in Hradec Králové, 1962 – 1995.

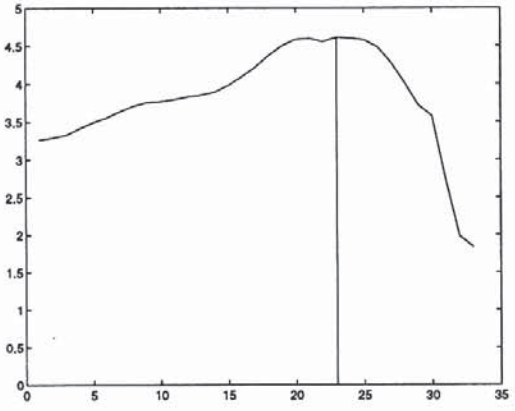


Figure 8. Sequence $\{\tilde{T}_k\}$ for annual averages of ozone.

Figure 8 shows the behavior of the sequence $\{\widehat{T}_k,\ k = 1, \ldots, 33\}$. Its maximum - the statistic $\widehat{T}(n)$ - attains the value 4.61. Supposing the observations are independent the value is significant even on the level 0.001 and detects a change in 1984. Figure 9 shows the data including the model under the alternative (4).

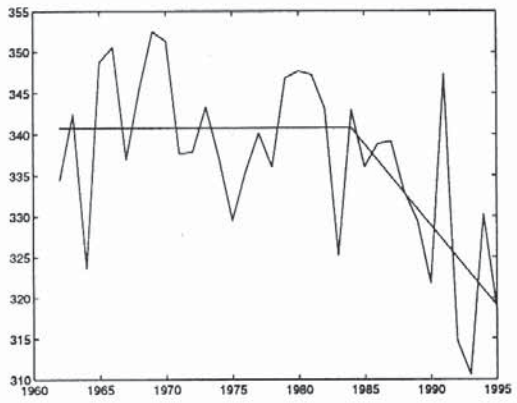


Figure 9. Annual averages of ozone with the fitted model.

The residuals and the autocorrelation function of residuals are shown in Figure 10 and Figure 11. It seems that the model fits the data well.

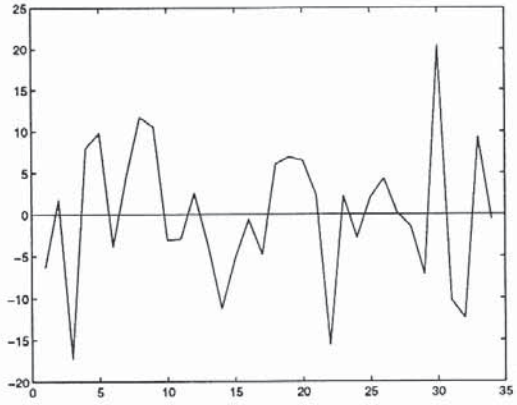


Figure 10. Residuals after fitting the model.

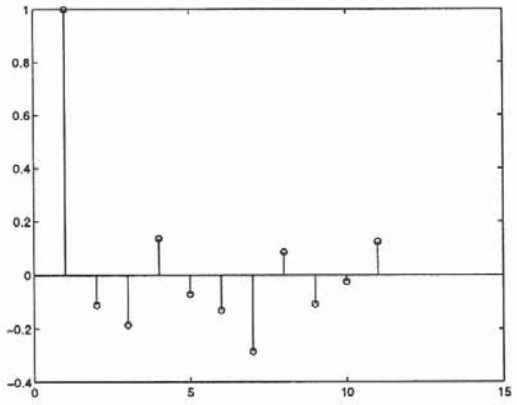


Figure 11. Autoregression function of the residuals.

Figure 12 shows the corresponding monthly averages. To estimate the seasonal component and to estimate the autocorrelation function we supposed that the series was certainly stationary in the first ten years (1962 – 1971) and we used these data for estimation. The autocorrelation function of the first part of series adjusted by removing the seasonal component ($n = 120$) is given in Figure 13.

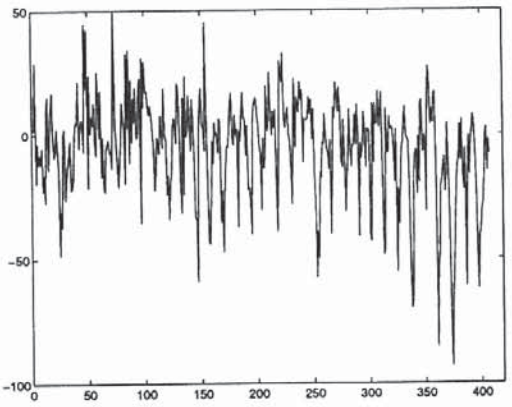


Figure 12. Monthly averages of ozone between 1962 and 1995 after removing the seasonal component.

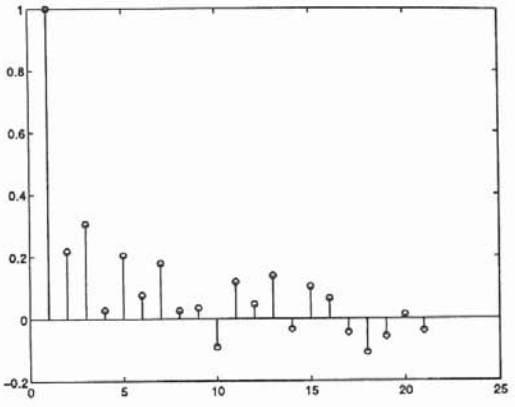


Figure 13. Autoregression function for seasonal adjusted monthly averages between 1962 and 1972.

The nonparametric estimator of A calculated from (9) is 2.9. According to several information criteria the AR(2) sequence

$$e_t - a_1\, e_{t-1} - a_2\, e_{t-2} = \varepsilon_t,$$

where $a_1 = 0.16067$ and $a_2 = 0.27473$, has been chosen as the best ARMA model. The parametric estimator of A is then

$$\widehat{A} = \frac{1}{(1-a_1-a_2)^2}\,\frac{1+a_2}{1-a_2}\left((1-a_2)^2 - a_1^2\right) = 2.76.$$

For the second part of the monthly series between 1972 and 1995 ($n = 288$) adjusted by removing the seasonal component (estimated from the first part of the series) we calculated the corresponding sequence $\{\tilde{T}_k\}$, see Figure 14. Its maximum - the statistic $\tilde{T}(n)$ - attains the value 5.25 for the change point $k^* = 153$ which corresponds again to year 1984. If we take the 5 % critical value 2.65 or 1 % critical value 3.21 from Table 2 and multiply them by $\sqrt{A}$ (supposing $A = 2.9$) we get 4.51 as the 5 % critical value and 5.48 as the 1 % critical value. The null hypothesis claiming that the monthly series is stationary is rejected at the 5 % significance level. The seasonal adjusted monthly series (1972 – 1995) with the fitted model is presented in Figure 15.

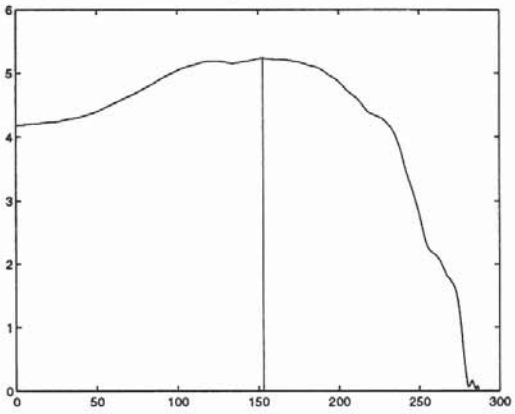


Figure 14. Sequence $\{\tilde{T}_k\}$ for seasonal adjusted month ly averages (1972 – 1995).

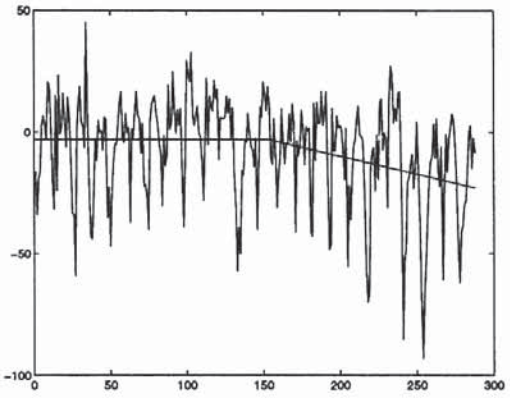


Figure 15. Seasonal adjusted monthly averages (1972 – 1995) with the fitted model.

Figure 16 shows the seasonal components (January average, February average ...) for years 1962 – 1985 and 1986 – 1995 and Figure 17 shows their differences. The most significant decrease occurred in those months in which the total amount of ozone is greatest, i.e., January, February, March and April.

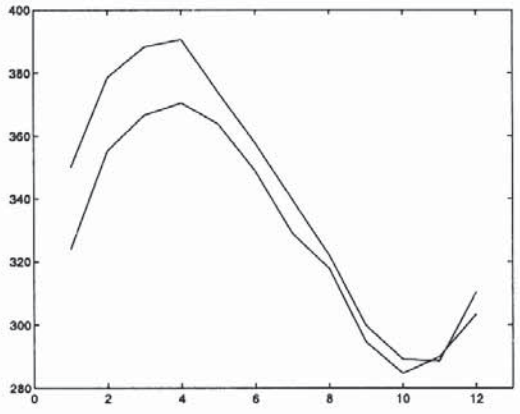


Figure 16. Seasonal component corresponding to years 1962 – 1984 and 1985 – 1995.

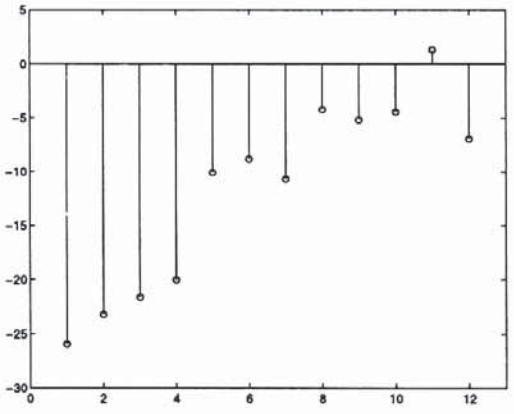


Figure 17. Difference between seasonal component for years 1962 – 1984 and 1985 – 1995.

The study of series corresponded to those months showed that a decrease started already before 1972, e.g. in case of February data (see Figure 18) and in case of March observations even before 1962 as it is shown in Figure 19.

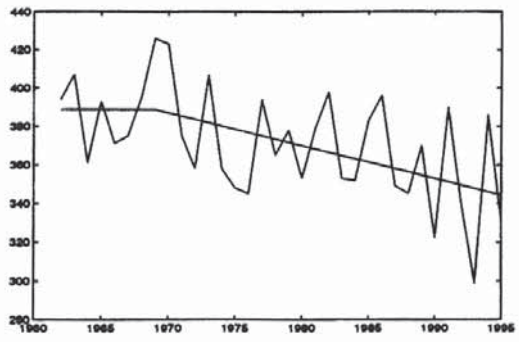


Figure 18. February averages, 1962 – 1995.

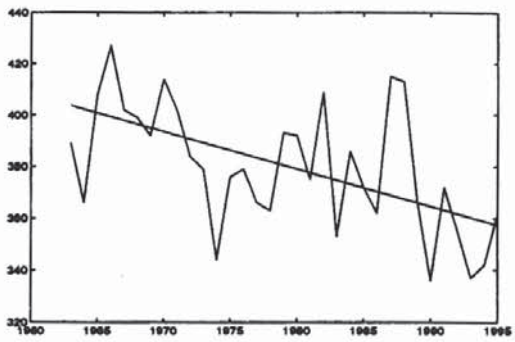


Figure 19. March averages, 1962 – 1995.

Unfortunately, this seems to be typical for natural series related to a damage of environment. One does not usually monitor them before they start to change their behavior.

Acknowledgement: The work was partially supported by grant GAČR 97/201/1163.

References

[1] Antoch J. and M. Hušková (1998), Estimators of changes, *Nonaparametrics, Asymptotics and Time Series, A tribut to Madan Lai Puri*, ed. S. Ghosh, M. Dekker, New York.

[2] Antoch J., M. Hušková and Z. Prášková (1997), Effect of dependence on statistics for determination of change, *Journal of Statistical Planning and Inference* **60**, 291 – 310.

[3] Bai J. (1994), Least squares estimation of a shift in linear processes, *J. Time Series Analysis* **15**, 453 – 472.

[4] Jaruškov´a D. (1997), Some problems with application of change-point detection methods to environmental data, *Environmetrics*, **8**, 469 – 483.

[5] Kim H.J. (1995), Detection of a change in longitudinal data, *Proceedings of International Conference on Statistical Methods and Statistical Computing for Quality and Productivity Improvement*, Seoul (Korea), 748 – 754.

[6] Nicholson S. E. (1994), Century-scale series of standardized annual departure of African rainfall. In *Trends'93: A Compendium of Data on Global Change,* T. A. Boden, D. P. Kaiser, R.J. Sepanski and F.W. Stoss (eds.) 952 – 962. ORNL/CDIAC - 65, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tenn., USA.

[7] Tang S.M. and Mac Neil I.B. (1993), The effect of serial correlation on tests for parameter change at unknown time, *Ann. Statist.* **21**, 552 – 575.

[8] Yao Yi-Ching and R.A. Davis (1986), The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhya* **48**, 339 – 353.

ÖZET

Hidrolojik ve meteorolojik süreçlerin durağan olup olmaması " değişim noktası olarak adlandırılan bir yöntemla araştırılmıştır.

İstatistik, Journal of the Turkish Statistical Association
1998, Vol.1, No. 2, pp. 23 - 37

SMALL SAMPLE ESTIMATES OF SEMIVARIOGRAMS

Zekai Şen
Istanbul Technical University, Faculty of Aeronautics and Astronautics,
Maslak 80626, Istanbul, Turkey

Abstract

It is noticed that small sample estimations of semivariograms are affected by finite sample lengths as well as the persistence of the regionalized variables. In order to quantify this point the small sample expectation of semivariogram is derived analytically by taking into account the autocorrelation structure of the regionalized variable. Analytical derivations are based on the Kendall's bias estimation procedure for the autocorrelation of autoregressive, (AR), processes. His procedure is modified for the semivariogram of AR process and explicitly expanded to cover the autoregressive integrated moving average, ARIMA(1,0,1), processes.These processes subsume AR, moving average, (MA), independent, (IP) and Brownian motion processes. For simple correlations within the regionalized variables which are exemplified by AR process, the sample semivariograms overestimate the population counterparts. The AR process semivariograms do not exhibit any nugget effect even for small samples. However , in the case of ARIMA(1,0,1) process the amount of bias is relatively smaller and for small samples there appears nugget effect the amount of which diminishes with the increase in sample lengths. As the sample size increases and / or the correlation coefficient decreases the discrepancy between the sample and theoretical semivariogram decreases. Analytical expressions of small sample semivariogram estimates explain explicitly the linearity, nugget and hole effects in the semivariogram depending on the type of stochastic process.

Key words: Autocorrelation;bias;small samples;semivariogram;stochastic process.

1. Introduction

The regional variability in any geological phenomena has been modeled thought the use of semivariograms for almost three decades starting with the original work by Matheron(1962).The main purpose of the semivariogram has been to define the distance over which these phenomena are interdependent.Depending on the shape

of semivariogram over small distances the geological phenomenon is said to have either independent structure in the statistical sense, i.e., the occurrences of the phenomenon at different sites do not affect each other or dependent structure from which one can conclude whether that the phenomenon is perfectly, moderately or weakly continuous. Furthermore, the shape of experimental semivariogram over the initial distances determines the model as linear,power,spherical,Gaussian,De Wijsian,etc.,type,(Clark,1979).Last but not the least the comparison of semivariograms for the same phenomenon but at different directions provides information as to the isotropy and heterogeneity of the phenomenon concerned.

The semivariogram concept has developed rapidly especially in ore reserves estimation, (David1977; Journel and Huijbregts, 1978), ground water storage and quality predictions (Aboufirassi and Marino, 1982; Myers et at.,1982; Subyani and Sen, 1989), earthquake evaluations, (Carr and Glass, 1985) as well as in many other earth science domains. Besides, the semivariogram is a model used in the estimation procedure of kriging which provides the best unbiased and linear estimation of any regional variable. The success in the kriging procedure relies significantly on the most suitable choice of the semivariogram type. Such a task can be achieved first of all by identification of an unbiased empirical semivariogram from the available data.

In addition to the aforementioned advantages of the semivariogram, a considerable degree of confusion has arisen, recently, over its basic terminology. There appeared a series of critical discussions whereby the very basis of the semivariogram concept has been questioned, (Philip and Watson, 1986; Shurtz, 1986). Concerning the semivariogram, these discussions concentrated on the arbitrary way that a theoretical semivariogram model is fitted to sample semivariograms and furthermore, on the effects of averaging procedure as well as extreme values of the data in addition to the irregular distribution of sampling sites within the study area. However, cumulative semivariograms avoid almost all these points, (Sen, 1989).

Another significant point in the sample semivariogram estimations from finite-length samples is whether these estimations are biased or not?. It is stated by semivariogram advocates that in the theoretical model fitting only the initial portion of the sample semivariogram should be used. In fact, in the Kriging applications, the small distance semivariogram values play the major role.It is already confirmed by Sharp(1982a,b) through extensive Monte Carlo simulations for some stochastic processes that there are significant deviations between the sample and theoretical semivariograms.He stated that serious attention should be given to the relative importance of short range fitting of semivariograms upon kriging estimations.It is noted that departures between the theoretical and sample semivariograms may reach to significant levels.This is clearly the result of finite length of observations. Hence, this point raises important questions as regards the theoretical model fitting to sample semivariograms and its effective consequences in the kriging estimates of regionalized variable.Even though the theoretical model may predict, for instance, an exponential rise the actual series may produce sample semivariograms which might not be

exponential for very long samples, (Sharp, 1982a).However, the cause of this effect could not be documented analytically except by the Monte Carlo simulations.It will be sequel of this paper analytically that such deviations are due to the bias effects which might appear as a result of different reasons among which finite sample length constitutes the major role.

It is the main aim in this paper to derive analytical expressions for the bias effects resulting from the small sample semivariogram estimations.In order to verify the bias effects, herein, the autoregressive integrated moving average, ARIMA(1,0,1) process proposed by Box and Jenkins(1970)is adopted for the analytical derivation of bias effects on the small sample semivariogram estimates.This stochastic process subsumes other simple processes such as the independent, IP, autoregressive, AR, and moving average, MA, processes.

2. Bias effect

An assumption which is a prerequisite in any simulation, estimating or modeling scheme is that the field data represent only a finite length sample from the underlying phenomenon.It is not possible to have full information on a random variable, (RV), unless all of the possible values are recorded exhaustively.This cannot be the case for observations of geologic phenomena and as a result, the information content in each recorded sequence will be lacking in some way and accordingly the parameter estimates will be operationally biased, that is

$$\alpha_n = \alpha \tag{1}$$

In which α_n is the estimate of any concert parameter from finite length data and α is the corresponding population value that could result from extremely long i.e., complete data set only. Hence, by no means it is possible to extract the population values from finite data sets. As in eq. (1), if the finite length estimate from a single set of data is not equal to its population counterpart then this parameter is said to be in operational bias and the amount of bias, B, is

$$B = \alpha - \alpha_n \tag{2}$$

This bias amount diminishes as the number of data increases but for small samples it always exists and there is no way to get rid of it entirely. However, it can be reduced to a minimal value by different regionalization techniques.

On the other hand, quite distinct from the operational bias which is due to the sampling variability, is the statistical bias associated with parameter estimation. Provided that the underlying stochastic generating mechanism of the geological phenomenon is known a priori, then the estimates of parameters drawn from the finite historic sequence could first be corrected for bias and subsequently these corrected

estimates could be assumed to be the equivalents of corresponding population values. In general two types of statistical resemblance can be achieved between the field observation set and the equally likely generation (by simulation) of these variables by stochastic processes, (Sen, 1974). The first type of resemblance is asymptotically satisfied for large samples;although for small samples bias effect may exist in some of the parameter estimates, and small sample resemblance may not be satisfied. If $\alpha_p$denotes the estimate of any desired parameter desired from a set of finite data, then it plays the role of population parameter. If α_{ns} is the estimate of the same parameter from a generated synthetic sequence, then the field and generated data are said to resemble each other asymptotically provided that $\alpha_{ns} \rightarrow \alpha_p$ as the length of generated data goes to infinity. On the contrary, when dealing with a set of finite data, it would appear more logical to maintain the resemblance for small samples as well. This kind of resemblance is defined in terms of ensemble of m available sequences of finite length n from the same phenomenon. Each one of these finite- length sequence provides an estimate of the parameter concerned and these estimates are then averaged over the ensemble to give an overall estimate denoted by α_{ns^*}; then $\alpha_{ns^*} \rightarrow \alpha_p$, as the number of sequence m goes to infinity. The asymptotic value of the bias amount has to be equal to zero, which ensures that the finite field data and generated counterparts resemble each other asymptotically as well as for the small samples. The statistical bias contrary to the operational bias can be eliminated completely provided that there are suitable analytical expressions which quantify the bias amounts.

The magnitude of bias in parameter estimates of a stochastic process depends on (i) the marginal probability distribution function of the phenomenon concerned, (ii) the autocorrelation structure, (iii) the type of estimate employed, (iv) the length of data, and finally (v) whether or not the mean value of the process is known or has to be estimated. Herein, only the normal probability distribution will be considered together with the ARIMA (1,0,1) process autocorrelation structure as given by Box and Jenkins (1970). As for the types of parameter estimates are concerned there are four different procedures available in the literature which are (i) the moment estimate, (ii) the maximum likelihood estimate, (iii) the maximum entropy estimate and finally (iv) the Bayesian estimate. However, in this study the maximum likelihood estimates will be employed due to certain statistical advantages, (Box and Jenkins, 1970), as well as extensive use in practice.

3. Small sample estimation of semivariogram

One of the most important property of any geologic phenomenon is that its variability of regionalized variable can be measured with the semivariograms which are defined as the half of the summation of successive square differences between observations at two sides h apart. The lag-k semivariogram estimate, γ_{k^*}, from a traverse over which there are n equidistant data values, Z_i, $(i = 1, 2, ..., n)$ within the study area can be written as, (Clark, 1977):

$$\gamma_{k^*} = \frac{1}{n-1} \sum_{i=1}^{n-k} (z_i - z_{i+k})^2 \tag{3}$$

Herein, Z_i's may represent observed sample values such as ore grades, layer thicknesses, ground water level elevations, density, etc. Notice that γ_{0^*}corresponds to zero distance and its value will be zero. Definition in eq. (3) does not include the sample mean value estimate as it is the case for the autocorrelation estimations and therefore semivariograms have the advantage that for small lags it is independent of long term variations. In order to be able to quantify the effect of only regional dependence within a geological phenomenon by means of the semivariogram estimates, first of all the observed data values, (Z_i' s) will be rendered into a standard sequence, Z_i, thorough a standardization procedure as,

$$z_i = \frac{z_i - \overline{z}}{S} \tag{4}$$

in which $\overline{Z}$ and S are the arithmetic average and the standard deviation values estimated from the given set of finite data. The standardized sequence has unit standard deviation and zero mean value. Besides, after the standardization the series become second order stationary. In addition, such a standardization procedure does not affect the autocorrelation structure of the original observations at all. In other words, Z_i 's and z_i 's have the same autocorrelation structure. In an ensemble of standardized series with finite lengths, the estimates of semivariogram and autocorrelation will be different for each ensemble member. Hence, the application of expectation operation on both sides of eq.(4) leads after some simple algebraic calculations to

$$E(\gamma_k) = 1 - E(\rho_k) \tag{5}$$

in which the second term on the right hand side indicates the small sample estimation of the lag-k serial autocorrelation coefficient. Hence, it is apparent that the small sample properties of the semivariogram can be fount only after the analytical expressions are derived for the small sample expectation of autocorrelation coefficient.

The most commonly used estimate of the serial correlation coefficient is a circular series approach, (Sen, 1974), in the sense that the end of a sample series is assumed to be followed by its beginning, and the general from of lag-k autocorrelation coefficient estimate, ρ_{k^*} is given by as

$$\rho_{k^*} = \frac{\sum_{i=1}^{n} \left(z_i - \frac{1}{n} \sum_{i=1}^{n} z_i \right) \left(\sum_{i=1}^{n} z_{i-k} - \frac{1}{n} \sum_{i=1}^{n} z_i \right)}{\sum_{i=1}^{n} \left(z_i - \frac{1}{n} \sum_{i=1}^{n} z_i \right)^2} \tag{6}$$

This expression also represent the maximum likelihood estimation of the autocorrelation coefficient. For the sake of discussions in the following sections a brief

derivation of small sample expectation of γ_{*k} will be exposed briefly as presented by Kendall (1954). Accordingly, Eq.(6) can be written succinctly as a ratio similar to which was first proposed by Kendall as:

$$\rho_{k*} = \frac{A}{B} \tag{7}$$

in which A represent the k-th order covariance function in the numerator and B corresponds to the variance in the denominator of eq.(6), respectively. It is obvious that A and B RV's for within the ensemble finite sequences and so is γ_{k*}. Let the first order moments of these RV's be represent by E(A) and E(B). Then eq.(7) can be written in terms of shifted variables as

$$\rho_{k*} = \frac{E(A)+a}{E(B)+b}$$

where a and b are new RV's with zero means, i.e., expectations and hence,

$$\rho_{k*} = \frac{E(A)}{E(B)}\left[1+\frac{a}{E(A)}\right]\left[1+\frac{b}{E(B)}\right]^{-1} \tag{8}$$

The right hand side of this expression can be expanded into an infinite summation, first, by applying the Binomial expansion formula and subsequently taking the expectation of both sides and then performing the necessary algebraic calculations with the view that the $E(b^2) = Var(b)$,the following approximate but general formula can be obtained,

$$E(\rho_{k*}) = \frac{E(A)}{E(B)} - \frac{Cov(a,b)}{E^2(B)} + \frac{E(A)Var(b)}{E^3(B)} \tag{9}$$

in the derivation of which third and higher order terms have been ignored. Accordingly, from eq.(5) the small sample expectation of the semivariogram can be found in general as

$$E(\gamma_{k*}) = 1 - \frac{E(A)}{E(B)} + \frac{Cov(a,b)}{E^2(B)} - \frac{E(A)Var(b)}{E^3(B)} \tag{10}$$

The explicit forms of various terms on the right hand side of eq.(10) are given by Kendall (1954) as

$$E(A) = \frac{1}{n-k}\left[(n-k-1)\,\rho_k - \frac{1}{n-k}\sum_{i=1}^{n-k-1}(n-k-i)\,(\rho_{k+i}+\rho_{k-i})\right] \tag{11}$$

$$E(B) = 1 - \frac{1}{n-k} - \frac{2}{(n-k)^2}\sum_{i=1}^{n-k-1}(n-k-i)\,\rho_i \tag{12}$$

$$Cov(a,b) = \frac{2}{n-k} \sum_{i=1}^{n-k-1} \rho_i \rho_{i+k} \tag{13}$$

and finally

$$Var(b) = \frac{2}{n-k} \sum_{i=1}^{n-k-1} \rho_i^2 \tag{14}$$

The substitution of eqs. (11)-(14) into eq.(10) yields to a huge expression which has been avoided herein.

4. Biased semivariogram of stochastic processes

In the aforementioned formulations, eq.(10) provides the small sample expectations of the classical semivariogram in its general but implicit form without any reference to any particular stochastic process. However, specifically autoregressive integrated moving average process, ARIMA(1,01,0) will be adopted, herein, in deriving the analytical expressions for $E(\gamma_{k^*})$. These process can be reduced easily into other simpler processes such as the independent, (IP), autoregressive, (AR) or moving average, (MA) processes depending on the values of model parameters as will be explained in the sequel. Besides, the ARIMA(1,0,1) processes are capable for representing long memory effects in natural phenomenon. In general, two subsequent values, Z_i and Z_{i-1}, of the phenomena concerned are related recursively to each other as

$$Z_i = \phi Z_{i-1} + \epsilon_i - \theta \epsilon_{i-1} \tag{15}$$

in which ϕ and θ are model parameters and ϵ_i's are random variable with Gaussian distribution. Notice that for $\phi = 0$ eq.(15) yields MA process and when $\theta = 0$ it reduced to an AR process which is commonly known as a first order Markov process and, finally if $\phi = \theta = 0$ then the resulting process is purely random which is usually referred to as white noise, i.e., IP. In the case of AR process, $\phi = \rho_1$ where ρ_1 is the lag-one autocorrelation coefficient of this process. A common property in all of this processes is that they are stationary. Last but not least, when $\phi = 1$ and $\theta = 0$ eq.(15) leads to random increments, $Z_i - Z_{i-1} = \epsilon_i$, which is known as either random walk or Brownian motion process. As stated by Jenkins and Watt (1968) this process is non stationary in both the mean and the variance.

The autocorrelation structure of ARIMA (1,0,1) process is given in terms of ϕ and θ by Box and Jenkins (1970) explicitly a

$$\rho_0 = 1$$

$$\rho_1 = \frac{(\phi - \theta)(1 - \phi\theta)}{1 + \theta^2 - 2\phi\theta} \tag{16}$$

$$\rho_i = \phi \rho_{i-1}$$

in which ρ_0 and ρ_i are the lag-zero and lag-i autocorrelation coefficients. The substitution of eq.(16) in to eqs.(11)-(14) and the completion of the necessary algebraic calculations yield explicitly that

$$E(A) = \frac{1}{n}\left[(n-k)\,\rho_1^{k-1} + k\phi^{n-k-1}\right] - \frac{1}{n} - \frac{2\rho_1}{n(1-\phi)} \tag{17}$$

$$E(B) = 1 - \frac{1}{n} - \frac{2\rho_1}{n(1-\rho_1)} \tag{18}$$

$$Cov(a,b) = \frac{2}{n}\left[\rho_1\phi^{k-2}(\phi + k\rho_1)\right] + 2\frac{\rho_1^2\phi^k}{(1-\phi)} \tag{19}$$

and

$$Var(b) = \frac{2}{n}\left[1 + \frac{2\rho_1}{1-\phi^2}\right] \tag{20}$$

The substitution of these equations into eq.(10) leads to the small sample expectations of ARIMA(1,0,1) process semivariograms as

$$E(\gamma_{k^*}) = 1 - \rho_1\phi^{k-1} + \frac{1}{n}\left[\frac{(1-\phi+2\rho_1)\left(1-\rho_1\phi^{k-1}\right)}{(1-\phi)} + k\rho_1(\phi+2\rho_1)\phi^{k-1} - \right.$$

$$\left. k\rho_1\phi^{n-k-1} + \frac{4\rho_1^2\phi^{k-1}}{(1-\phi^2)}(\phi-\rho_1)\right] \tag{21}$$

This expression is general in the sense that it gives the small sample expectations of semivariograms for IP, MA and AR processes by substitution of relevant ϕ and θ values as mentioned earlier in this section.

5. Analytical solutions and discussions

It is possible to derive many useful expression concerning the properties such as the nugget and hole effect, linearity, etc., of the semivariograms as follows.

(i) even tough the sample sizes are small the AR processes do not exhibit any nugget effect, i.e., the semivariogram value at the origin is equal to zero. This statement can be proven from eq.(21) by substituting first $\phi = \rho_1$ and then $k = 0$ and after the necessary algebraic manipulations one can find that $E(\gamma_{0^*}) = 0$ irrespective of sample sizes or autocorrelation structure.

(*ii*) on the contrary, for small samples ARIMA(1,0,1) processes possess nugget effect the value of which decreases with increase of sample size. It is not possible to substitute into eq.(21) directly $k = 0$ since as is obvious from eq.(16) that k is meaningful only when it is greater then 1. The recurrence relationship in eq.(16) indicates that in general $\rho_k = \rho_1 \phi^{k-1}$. Therefore, in eq.(21) first of all $\rho_1 \phi^{k-1}$ terms should be replaced with ρ_k and then $k = 0$ must be substituted with consideration that $\rho_0 = 1$. Quantitatively, one can obtain the analytical expression for the nugget effect after these algebraic manipulations as:

$$E(\gamma_{0^*}) = \frac{4\rho_1(\phi - \rho_1)}{n(1 - \phi^2)} \tag{22}$$

which shows clearly the nugget effect for finite sample lengths. However, for large sample sizes ($n \to \infty$) there is no nugget effect, i.e., $E(\gamma_{0^*}) \to 0$.Notice from eq.(22) that for the AR process, $\rho_1 = \phi$, and hence there is no nugget effect.

(*iii*) in the MA process case there appears no nugget effect but it yields a distinct hole effect at lag one as was observed by Sharp (1982) on the basis of extensive Monte Carlo simulations. However, the analytical expression of this effect can be obtained easily from eq.(21) by substituting $\phi = 0$ and $k = 1$ which leads to

$$E(\gamma_{1^*}) = 1 + \frac{1}{n}(1 + 2\rho_1)$$

or, since $\rho_1 = -\theta/(1 + \theta^2)$ the substitution yields

$$E(\gamma_{1^*}) = 1 + \frac{1}{n}\left(1 - \frac{2\theta}{1 + \theta^2}\right) \tag{23}$$

(*iv*) one of the extremes in the semivariograms appear when the AR parameter, ϕ, is set equal to unity simultaneously with $\theta = 0$ which implies that $\phi = \rho_1$. This situation corresponds to the case of Brownian motion. The substitution of these conditions into eq.(21) leads to

$$E(\gamma_{k^*}) = \frac{2k}{n} \tag{24}$$

which implies that for finite sample lengths the semivariogram appears as linear trends. However, for very large sample lengths ($n \to \infty$) the semivariogram becomes equivalent with the horizontal axis for $k << n$.

(*v*) another extreme of the semivariogram is due to random independent process whereby $\phi = \theta = \rho_1 = 0$ and the substitution into eq.(21) gives

$$E(\gamma_{k^*}) = 1 - \frac{1}{n} \tag{25}$$

It is obvious that for small samples there appears some bias in the beginning but increase of sample size leads to a complete nugget effect as $E(\gamma_{k^*}) = 1$

(*vi*) It is interesting to notice that for large sample sizes and small lags, i.e., $n >> k$, eq.(21) reduce reduce to the already known relationship between the semivariogram and autocorrelation function for large samples as

$$E(\gamma_{k^*}) = 1 - \rho_1^k \tag{26}$$

This may be the main reason why in theoretical semivariogram model fitting only the initial portion of the sample semivariogram is consider as reliable in model fitting.

6. Numerical solution and discussion

The small sample expectations of the semivariograms for the aforementioned processes can be calculated for different combinations of model parameter values on digital computers. To this and, first of all the numerical solutions of eq.(21) are presented in Figures 1-3 for various sample lengths and a set of ρ_1 values for AR processes.

The general impression that can be taken from these figures is that the finite sample length as well as the autocorrelation structure distort the population semivariograms. Small sample semivariograms deviate from the population semivariograms more and more with the increase of correlation coefficient. However, the smaller the dependence the smaller will be the deviation. Besides,for instance in Figure 1 at large lags the small sample semivariogram starts to decrease after reaching the maximum at some moderate lags. Furthermore, especially small sample semivariograms for big auto correlations give the impression of a power model where as the population semivariogram is of exponential type. This point indicates clearly that a simple mathematical function must not be fitted heuristically to the small sample semivariograms prior to eliminating the eliminating the bias effect. It is suggested herein that as a preliminary guide the graphs in Figure 1-3 can be used as standard curves and then the one that matches the sample semivariogram is picked up which deduced the underlying population value and accordingly the parameter estimate. Otherwise, all parameters derived from convectional semivariogram fittings are hypothetical and subjective. Comparisons of Figures 1-3 indicate that as and sample sizes increases the small sample semivariogram estimates are expected to approach the population semivariogram. However, as the sample size increases there appears cases where by after initial curvature point the small sample semivariogram remains constant for the majority of moderate lags. This is very similar to the hypothetically suggested spherical model where by the constancy after some distance corresponds to sill, i.e., the variance in the stochastic process terminology. For instance, in Figure 2-3 all

of the small sample semivariograms with different autocorrelation coefficients expect $\rho_1 = 0.9$ appear in the forms of spherical model whereby the underlying population (theoretical) semivariograms have exponential increase. This point also shows pitfalls in fitting a theoretical semivariogram to the sample semivariogram points without any further interpretation. In fact, there should be a physical and analytical basis in fitting theoretical semivariogram otherwise the small sample semivariogram might lead to some model which may not be consistent with the geological phenomenon occurrence at all.

Figures 4-5 show the effect of small samples very clearly for different sample lengths. It can be concluded that as the structural dependence in regionalized variable increases the chances are that the sample semivariogram becomes more biased and therefore does not represent the population semivariogram. On the other hand, for independently distributed regionalized variables there will be relatively very minor bias effect from the finite sample lengths, (see eq.25). Therefore, the best case for the semivariogram application away from the undesired bias effect is the independent regionalized variables which have limitations in practical application. Otherwise, the finite sample length as well as autocorrelation effects must be considered and accounted for in the semivariogram fittings.

It is possible to obtain numerous similar graphs for any set of parameters (ϕ, θ and n) from eq.(21) which shows the variation of $E(\gamma_{k^*})$ with the lag. Herein, Figures 6-7 represent two samples only from the small sample expectations of ARIMA(1,0,1) process for different model parameter values and sample lengths. By comparing these graphs with each other as well as with the previous figures for the AR process the following additional significant points can be observed :

(i) for the same AR parameters ϕ and ρ, invariably the AR processes semivariograms will have more bias than the ARIMA(1,0,1) process provided that $\theta < \phi$ which is the case with natural phenomena.

(ii) for fixed ϕ values increase in θ value leads to less biased semivariograms initial portions which is obvious from Figures 6 or 7.

(iii) comparison of Figures 6 and 7 indicates that as expected, any increase in the sample size decreases the bias effect.

(iv) furthermore, increase in ϕ parameter brings additional biased to the small sample semivariogram estimates.

7. Conclusions

The small sample properties of the semivariograms resulting from the different stochastic process are presented analytically. It is observed that the smaller the sample length the more will be the bias effect and, therefore, the semivariogram estimations at small distances will be in gross error compared to the parent population

values. Furthermore, paucity of data affects the sample semivariogram estimates, especially, if the regionalized variable has a strong dependence structure. This effect occurs in the form of bias which causes overestimations in the semivariogram calculations. The analytical expression is obtained that relates the small sample semivariogram expectations and the autocorrelation structure of the ARIMA(1,0,1) process in term of model parameters. This expression provides a common basis for the autoregressive, AR, moving average, MA, independent, IP, and Brownian motion processes small sample semivariogram expectations. The analytical expression of these expectations showed quantitatively the linearity, nugget and hole effect in the semivariogram estimates.

References

[1] Aboufirassi, M., and Marino, M. A., 1984. A geostatistically based approach to the identification of aquifer transmissivities in Yolo Basin, California: Math. Geol., v.16, n.26, p.125-137.

[2] Box, G.E.P. and Jenkins, G.M., 1970. Time series analysis, forecasting and control: Holden-Day, San Francisco, pp. 553

[3] Carr, J.R. and Glass, C.E., 1985. Treatment of earthquake ground motion using regionalized variables: Math. Geol., v.17, n.3, pp221-241.

[4] Clark, I., 1979. The semivariogram - Part I. Engin. Min. Jour., v.180, n.7, pp90-94.

[5] David, M., 1977. Geostatistical ore reserve estimation: Elsevier, Amsterdam, pp. 364.

[6] Jenkins, G.M. and Watt, D.G., 1968. Spectral analysis and its applications. HoldenDay, San Francisco, pp.525.

[7] Journel, A.G. and Huijbregts, Ch, J., 1978. Mining Geostatistics: Academic Press, London, pp. 600.

[8] Kendall, M.G., 1954. Note on the bias in the estimation of autocorrelation, Biometrica, v.42, pp.403-404.

[9] Matheron, G., 1962. Traite de geostatistique applique. Tome 1: Editions Technic, Paris, pp.334.

[10] Myers, D.E., Begovic, C.L, Butz, T.R., and Kane, V.E., 1982. Variogram models for regional ground water chemical data: Math. Geol., v.14,

n.6, pp. 629-644.

[11] Philip, G.M., and Watson, D.F., 1986. Matheronian Geostatistics-Quo Vadis?. Math. Jeol., v.18, n.1, pp. 93-119.

[12] Sharp, W.E., 1982a. Stochastic simulation of semivariograms. Math. Geol., v.14, n.5, pp.445-456

[13] Sharp, W.E., 1982b. Estimation of semivariograms by the maximum entropy method. Math. Geol., v.14, n.5, pp. 457-474.

[14] Shurtz, R.E., 1986. A critique of A. Journel's "Deterministic Side of Geostatistics", Math. Jeol., v.17, n.8, pp. 861-869.

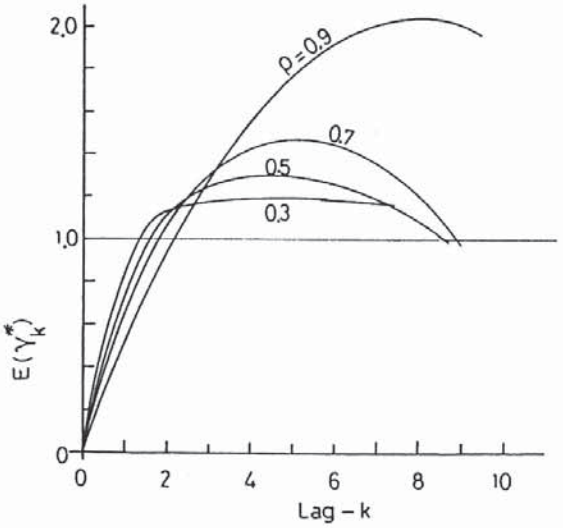
[15] Subyani, A., and Sen, Z., 1989. Geostatistical modeling of Wasia aquifer in central Saudi Arabia, Jour. Hydrol., v.110, pp. 295-314.

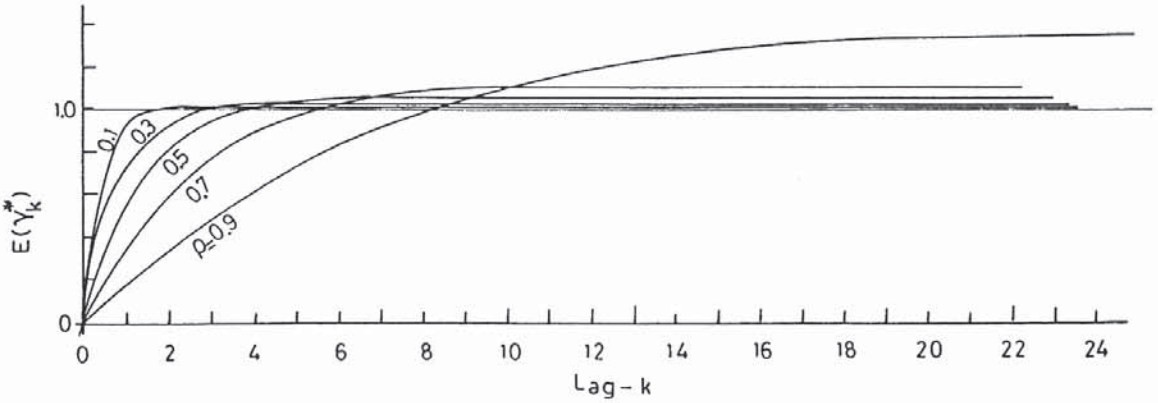
[16] Sen, Z., 1974. Small sample properties of stationary stochastic models and the Hurst phenomenon in hydrology. Unpublished Ph.D. Thesis, Imperial College of Science and Technology, University of London. pp. 284.

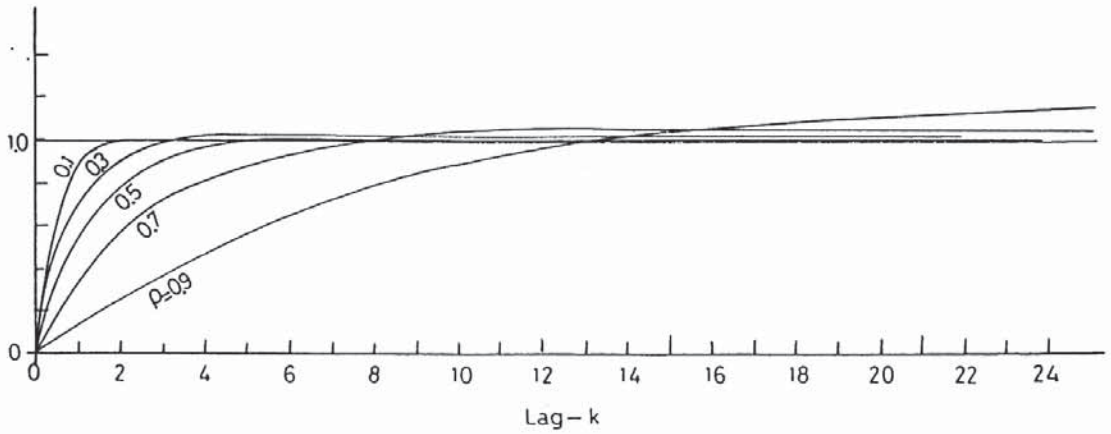
[17] Sen, Z., 1989. Cumulative semivariogram models of regionalized variables. Math. Geol., v.21, n.8, pp. 891-903.

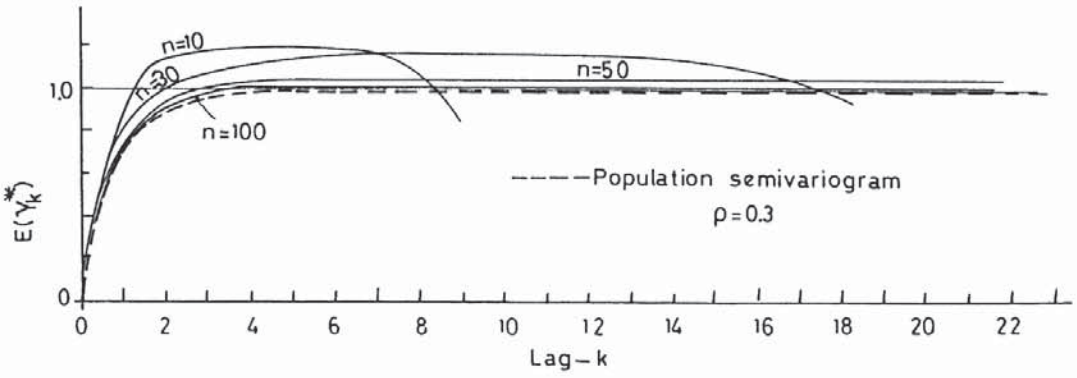
List of figure captions

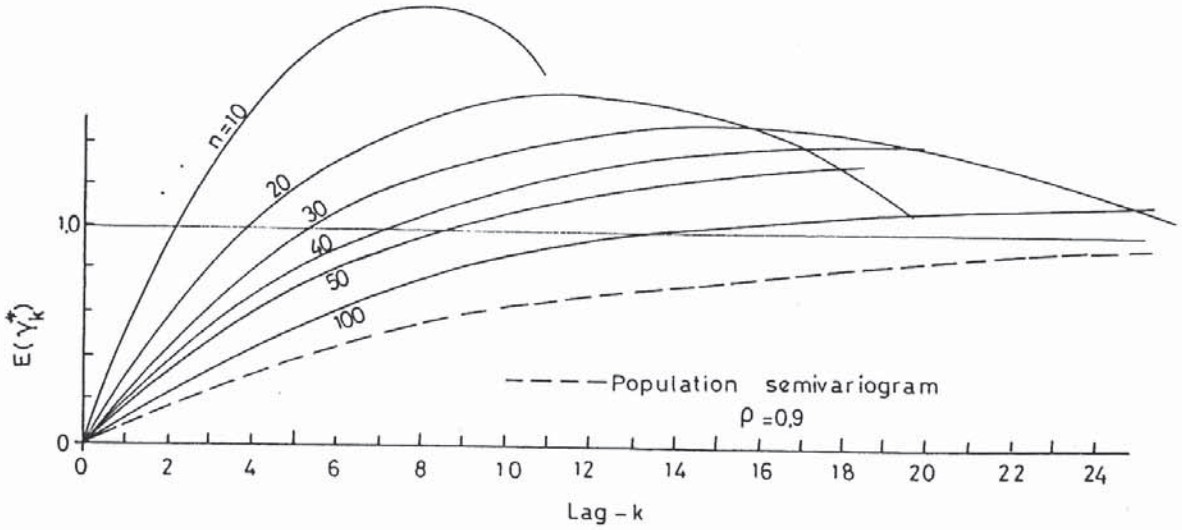
FIGURE 1. Semivariogram for AR process, ($n = 10$).
FIGURE 2. Semivariogram for AR process, ($n = 50$).
FIGURE 3. Semivariogram for AR process, ($n = 100$).
FIGURE 4. Semivariogram for AR process, ($\rho_1 = 0.3$).
FIGURE 5. Semivariogram for AR process, ($\rho_1 = 0.9$).
FIGURE 6. Semivariogram for ARIMA(1,0,1) process, ($\phi = 0.8$, $n = 20$).
FIGURE 7. Semivariogram for ARIMA(1,0,1) process, ($\phi = 0.9$, $n = 30$).

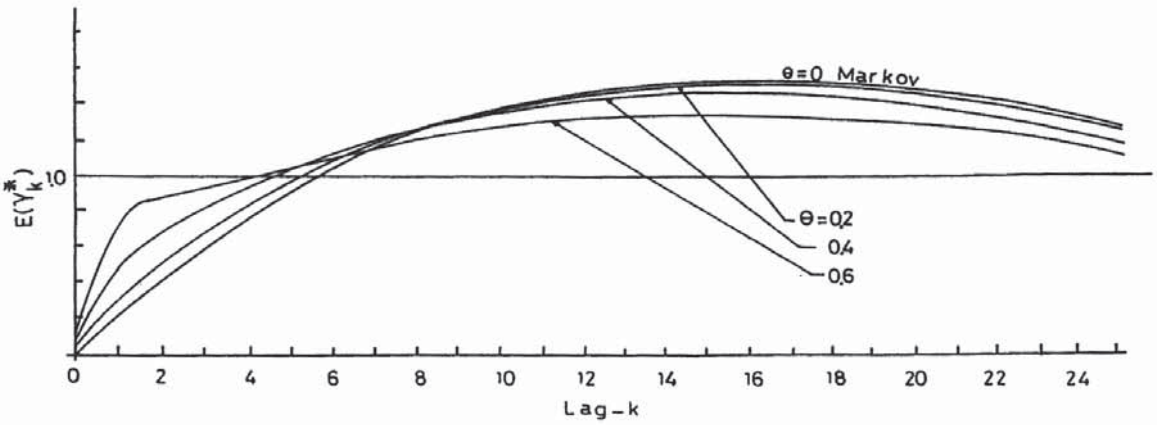


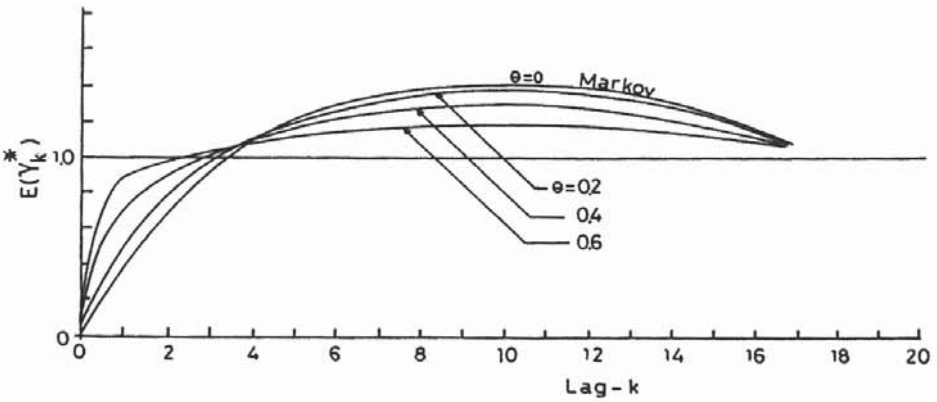



1.0
0.1
0.3
0.5
0.7
ρ=0.9
0
2
4
6
8
10
12
14
16
18
20
22
24
Lag – k


n=10
n=30
n=50
n=100
1.0
0
E(Y*k)
Population semivariogram
ρ=0.3
0
2
4
6
8
10
12
14
16
18
20
22
Lag – k


n=10
20
30
40
50
100
1.0
0
E(Y*k)
Population semivariogram
ρ=0.9
0
2
4
6
8
10
12
14
16
18
20
22
24
Lag – k





ÖZET

Yarı-Variogramların tahminleri sonlu örneklem uzunluklarına ve yöre verilerin durumlarından etkilenir. Yöresel verilerin otokorelasyonlarının yapısına bal ık yarı-variogramların beklenen değerleri sonlu örneklemler için analitik olarak hesaplanmıştır

İstatistik, Journal of the Turkish Statistical Association
1998, Vol.1, No. 2, pp. 39 - 46

PREDICTION OF HIGH WATER LEVELS IN THE BALTIC

A. Svensson and J. Holst
Department of Mathematical Statistics
Lund Institute of Technology, Box 118, SE-22100 Lund, Sweden
Phone: +46 46 2224679 (ASv); +46 46 2229538 (JH)
email: anderss@maths.lth.se (ASv)
janh@maths.lth.se (JH)

Abstract

This paper presents a level crossing predictor for Gaussian ARMAX processes, which is optimal in the sense that it minimizes the number of false alarms for a given probability of detecting the level-crossings. It is applied to real data for predicting and warning for high water levels at the Danish coast in the Baltic Sea. The optimal alarm system is shown to work better than a simpler and more conventional alarm system. A method to optimally predict the crossings also when the external signals are not known is presented. In this particular case most of the variability of the predictions are due to system noise, so the performance of the system with predicted external signals are almost identical to the performance when the external signals are known. A smaller simulation study shows that the water level process is hard to predict and that the choice of model can be rather important.

KEY WORDS: level crossings, flooding alarm, catastrophe prediction, optimal alarm, ARMAX process.

1. Introduction

A flooding incident can be disastrous, especially if people are not warned. Hence, in many situations it is important to be able to give an alarm some time before the incident occurs. It is also important to give as few false alarms as possible, but still find a sufficient number of the flooding incidents.

In a more general setting, the problem is to predict level crossings, catastrophes, of a stochastic process a sufficient time in advance. This catastrophe prediction problem was treated by de Maré and Lindgren, and a definition of the optimal catastrophe predictor was given as the predictor that gives a minimum number of false alarms for a given

detection probability. This idea was further treated in Svensson, Holst, Lindquist & Lindgren , and leads to an explicit catastrophe predictor for Gaussian ARMA processes with constant catastrophe level. Since the construction of the optimal catastrophe predictor requires quite a large amount of calculations, two suboptimal predictors were also introduced. In Svensson & Holst

the technique was extended to cover both ARMAX and SETARMAX processes with a deterministic but changing catastrophe level. This made it possible to use the optimal catastrophe predictor on real data, describing water levels in the Baltic Sea, presented in this paper. Modelling of the water levels in the Baltic Sea is treated in Berntsen, Nielsen and Spliid & Nielsen . A complication with ARMAX processes is that the external signals might not be known in advance, which means that they have to be predicted too. An idea how this can be treated in the same framework as above is also included in this paper and applied to the data sets used.

2. The data set

The data sets used in this paper are from 1978, 1979 and 1980. They consist of the following measurements.

Location	Water level	Head wind	Side wind	Air pressure	Temp.
Korsør	X				
Rødbyhavn	X				
Gedser	X				
Visby	X				
Kadetrenden/ Maribo (78)		X	X	X	X
Møn-Sydøst lightship (79,80)		X	X	X	X
Møn lighthouse:		X	X	X	X
Christiansø lighthouse	X	X	X	X	X
Hammer Odde lighthouse:		X	X	X	X

Only three of these signals are used in the final model describing the water level at Rødbyhavn. They are the water level at Rødbyhavn, the head wind at Christiansø lighthouse and the air pressure at Kadetrenden/Maribo (78) or Møn-Sydøst lightship (79,80). The original data sets contained measurements every hour, but since the process is oversampled, only one sample per 3 hours was used for modelling the water level. Before they have been used for modelling, the mean value using data from all three years has been subtracted. However, the catastrophe levels used later are related to the original data. In Figure /refwater78 the water level at Rødbyhavn is shown for the data set from 1978. The complete data sets with a short description can be found at the address: http://www.maths.lth.se/matstat/staff/anderss/data/data.html.

3. The models

The water level at Rødbyhavn has been modelled as ARMAX and SETARX processes, denoted X_t, with two external signals, denoted $u_{1,t}$ and $u_{2,t}$. The structure of an ARMAX(p,q,r_1,r_2) process is

$$X_t + a_1 X_{t-1} + \ldots + a_p X_{t-p} =$$
$$b_{1,0}u_{1,t} + \ldots + b_{1,r_1}u_{1,t-r_1} + b_{2,0}u_{2,t} + \ldots + b_{2,r_2}u_{2,t-r_2} + c_0 e_t + \ldots + c_q e_{t-q},$$

or shorter

$$A(z^{-1})X_t = B_1(z^{-1})u_{1,t} + B_2(z^{-1})u_{2,t} + C(z^{-1})e_t,$$

where $\{e_t\}_{t=-\infty}^{\infty}$ is white noise and e_t is uncorrelated with X_s, $u_{1,s}$ and $u_{2,s}$ for $s < t$. It is furthermore assumed that $e_t \in N(0,1)$.

After trying a number of different models three were chosen and estimated on the data from 1978, and optimal alarm systems were calculated. The models are ARMAX(2,1,1,1), ARMAX(4,2,1,1) and SETARX(2;2,2;1,1). The noise is assumed to be independent and Gaussian with variance 1.

The ARMAX(2,1,1,1)-model is

$$\begin{aligned} A(z^{-1}) &= 1.0000 - 1.2794z^{-1} + 0.3786z^{-2} \\ C(z^{-1}) &= 5.5678 + 3.1836z^{-1} \\ B_1(z^{-1}) &= -0.0072z^{-1} \\ B_2(z^{-1}) &= 0.0232z^{-1} \end{aligned}$$

The empirical density functions for the one and two-step prediction error for the data set from 1978 are shown in Figure , together with the normal density function, and normal probability plots. It can be seen that the residuals have slightly heavier tails than in the normal distribution. However, in spite of these deviations the normal distribution has been used for modelling and calculation of the alarm systems. It seems to work rather well.

The ARMAX(4,2,1,1)-model is

$$\begin{aligned} A(z^{-1}) &= 1.0000 - 1.7227z^{-1} + 1.7602z^{-2} - 1.5950z^{-3} + 0.6652z^{-4} \\ C(z^{-1}) &= 5.1942 + 0.4259z^{-1} + 3.7263z^{-2} \\ B_1(z^{-1}) &= -0.0074z^{-1} \\ B_2(z^{-1}) &= 0.0225z^{-1}. \end{aligned}$$

The SETARX-model is composed of two ARX-models where

$$\begin{aligned} A(z^{-1}) &= 1.0000 - 1.5280z^{-1} + 0.6137z^{-2} \\ C(z^{-1}) &= 6.0661 \\ B_1(z^{-1}) &= -0.0049z^{-1} \\ B_2(z^{-1}) &= 0.0155z^{-1} \end{aligned}$$

is used when the process value $X_{t-2} < 30$ and

$$\begin{aligned} A(z^{-1}) &= 1.0000 - 1.4226z^{-1} + 0.7434z^{-2} \\ C(z^{-1}) &= 6.0661 \\ B_1(z^{-1}) &= 0.0202z^{-1} \\ B_2(z^{-1}) &= 0.1312z^{-1} \end{aligned}$$

when the process value $X_{t-2} \geq 30$.

In cases when models for the external signals are needed these signals have been modelled as AR processes. The model for the head wind at Christiansø lighthouse is an AR(3) process with the parameters

$$\begin{aligned} A_1(z^{-1}) &= 1.0000 - 1.7042z^{-1} + 0.5411z^{-2} + 0.1713z^{-3} \\ C_1(z^{-1}) &= 7.7460 \end{aligned}$$

and the model for the air pressure at Kadetrenden/Maribo (78) or Møn-Sydøst lightship (79,80) is an AR(1) process with the parameters

$$\begin{aligned} A_2(z^{-1}) &= 1.0000 - 0.9657z^{-1} \\ C_2(z^{-1}) &= 27.8675. \end{aligned}$$

It could be considered using one model for the air pressure at Kadetrenden/Maribo (78) and another model for Møn-Sydøst lightship (79,80), but since the locations are rather close to each other, the same model has been used. This also requires fewer calculations.

4. The optimal alarm system

The optimal alarm systems used in this paper are optimal in the sense that they minimize the probability of false alarms for a given probability of detecting the catastrophes. Optimality is reached by the alarm system defined through the likelihood ratio,

$$\frac{dP_{Y(t-k)}(y|C_t^*)}{dP_{Y(t-k)}(y|C_t)} \leq \text{constant},$$

where $Y(t)$ denotes the available information at time t, C_t is the event that a catastrophe occurs at time t and C_t^* is the complementary event that no catastrophe occurs at time t. This condition can be simplified, so that the alarm system can be based on only the predictor $(\hat{x}_{t-1}, \hat{x}_t)$ of the process X_t at times $t-1$ and t, instead of all the available information $Y(t)$. The result is

$$P(C_t|\hat{x}_{t-1}, \hat{x}_t) > P_b,$$

which was shown in Svensson et al. , to be the optimal alarm system for ARMA processes with a constant catastrophe level. It is then possible to calculate the alarm region in advance, which makes the alarm system rather fast. A typical alarm region

in the $(\hat{x}_{t-1}, \hat{x}_t)$-plane is shown in figure . The model is the ARMAX(2,1,1,1) described above, with the influence of the external signals subtracted. The predictor is using 6 old process values and the prediction horizon is 2. This idea was further developed in Svensson & Holst, to cover ARMAX and SETARX processes when the external signals are known and the process is stationary.

5. Alarm system using predicted external signals

Since predictions of the process values are needed in the level crossing predictor, also predictions of the external signals are needed when the external signals are not known in advance. In case of known external signals, the effect can be included in the catastrophe level, giving a catastrophe level that changes through time, see Svensson & Holst . However it is not that simple in case of stochastic external signals. A few assumptions on the signals have to be added in order to get an explicit level crossing predictor.

If we assume that the external signals and the process are stationary Gaussian processes, the covariance of the process value predictor $Cov(\hat{x}_{t-1}, \hat{x}_t)$ will include both the effects of the process noise and the external signals.

Suppose the process can be written,

$$\begin{aligned} A(z^{-1})X_t &= B(z^{-1})u_t + C(z^{-1})e_t \\ A_1(z^{-1})u_t &= C_1(z^{-1})w_t. \end{aligned}$$

Due to linearity, X_t can be decomposed into one part, $X_{u,t}$, describing the influence of the external signals and one part, $X_{e,t}$, describing the influence of the system noise.

$$\begin{aligned} X_t &= X_{u,t} + X_{e,t} \\ A(z^{-1})X_{e,t} &= C(z^{-1})e_t \\ A(z^{-1})A_1(z^{-1})X_{u,t} &= B(z^{-1})C_1(z^{-1})w_t. \end{aligned}$$

The same deductions can be done for the predictions, leading to

$$\hat{X}_t = \hat{X}_{u,t} + \hat{X}_{e,t}.$$

If the noise processes e_t and w_t are assumed to be independent, the covariance of the predictions $\hat{x}_{t-1}, \hat{x}_t$ is

$$Cov(\hat{x}_{t-1}, \hat{x}_t) = Cov(\hat{x}_{u,t-1}, \hat{x}_{u,t}) + Cov(\hat{x}_{e,t-1}, \hat{x}_{e,t}).$$

This means that if $X_{u,t}$ can be optimally predicted, the technique presented in Svensson & Holst , can still be used and thus the resulting critical levels for the stochastic part of the process will be

$$[L_{cat}(t-1|t-k), L_{cat}(t|t-k)] = [L(t-1), L(t)] - [\hat{x}_u(t-1|t-k), \hat{x}_u(t|t-k)]\,.$$

The catastrophe level $L(t)$ for the original process, is assumed to be deterministic and known, and need not be predicted. The part of the process that is due to the external signals, influences the mean value of the process and will thus enter as an addition to the catastrophe level. Predictions for times $t-1$ and t are needed and the information is available up until $t-k$.

One model, the ARMAX(2,1,1,1)-model with the external signals modelled as AR(3) and AR(1) as above, have been tested and the results are shown and compared to the other alarm systems in Table , Table and Table . The alarm system works well for the data set that was used for estimating the model, but poorer for the other two data sets. The reason for this could be that the fixed models for the process and the external signals are not totally correct. This is similar to the alarm system where the external signals are not predicted, which is expected since almost all the variability is due to process noise.

In order to check how much the departures from normality and model type influence the performance, a smaller simulation study based on the ARMAX(2,1,1,1) model above with the external signals simulated as AR(3) and AR(1), was also performed. It shows that the process is very hard to predict, and will give a large amount of false alarms if a high detection probability is desired. An alarm is denoted false if it does not predict the catastrope exactly in time. The influence of the inputs are rather easy to predict when the prediction horizons are short, leading to almost the same alarm system as for known inputs. The variability of the predictions of process values is almost entirely due to the influence of the system noise, e_t. The results from the simulation are shown in Table . When the wrong model is used the detection probability can become a lot lower than calculated. This is obvious, especially for the SETARX model. The performance would have been better if the models had been estimated on the simulated data and not on the water data. Worth noting is that the maximal detection probability for the naive-naive alarm system is 0.29, so it is not comparable to the other alarm systems.

6. Results

The optimal alarm systems for the different models were compared to some simpler alarm systems. The simplest alarm system, called the naive-naive alarm system, gives an alarm when the process value k steps before a possible catastrophe crosses a certain level. This alarm system did work, but not as well as the optimal alarm systems. The most important disadvantage is that the naive-naive alarm system will have a maximum detection probability, that cannot be exceeded and is rather low.

Another simple alarm system models the process and gives alarm when the predicted process values crosses a level, that was determined from the data sets. This alarm system did not work, hence it has not been included in the tables below.

The optimal alarm system has a nonlinear alarm region, that changes depending on the catastrophe level and the process values. This makes the optimal alarm system

rather complex. In many cases when the performance is important this is the alarm system that ought to be used. In other cases it might be good to compare a proposed simpler alarm system to the optimal in order to check how close to the optimal the simpler alarm system is.

The parameters describing the process have all been estimated on the data set from 1978, and then tested on all three data sets. The alarm level for the naive-naive alarm system has been optimized over the three data sets together. As can be seen in Table the naive-naive alarm system has a rather low maximal detection probability, and thus is not possible to use if a high detection probability is required. The performance of the optimal alarm systems for these three data sets does not differ very much from each other and they have almost the same number of false alarms. The detection probabilities used are shown in parenthesis. They were in most cases set to 90%. ARMAX2111pred is the alarm system where also the external signals are predicted.

The optimal alarm systems with the highest detection probabilities have quite a few false alarms according to the strictest definition, where an alarm is considered false if it does not predict the catastrophe exactly right in time, but it could be questioned if all of these should be considered false. In Figure , it can be seen that a few of the so called false alarms are early alarms, or alarms given when the levels are still critical. In case of early alarms, at least for one or two steps early which means 3-6 hours early, the additional cost should not be too large. Also, the confidence in the alarm system will not be damaged too much. In the case of alarms when still over the critical level, it means that it will take a little longer to get back to normal state from the emergency state, caused by the process being alarmed. The cost should be small compared to the cost of the catastrophe. If these ideas, i.e. one and two steps early alarms are counted as correct alarms and alarms given when in catastrophe state are not counted at all are taken into account, Table will turn into Table .

The alarm level for the naive-naive alarm system is optimized over all three sets. It only reaches a total detection probability of approximately 40 %, which is far below the detection probabilities reached by the different optimal alarms. However, a higher detection probability will inevitably lead to more false alarms, and that is a trade-off that has to be made in each individual case.

In Figure close-ups at some different times are shown to give an explanation for the rather high rate of false alarms. 95% one-dimensional confidence intervals based on the one and two step predictions are also shown. The process is rather hard to predict which leads to wide confidence intervals and a high number of false alarms if a high detection probability is wanted.

7. Conclusions

This paper has presented an optimal alarm for processes described by linear or piecewise linear processes applied to prediction of high water levels in the Baltic. The

optimal alarm technique gives as few false alarms as possible for a given probability of detecting the catastrophes.

Data are collected in the southern part of the Baltic and high water levels in Rødbyhavn in Denmark are to be predicted.

The models that are used to describe the water levels all contain external variables, with future values that are unknown at prediction time. This means that also these external signals have to be predicted, which influences alarm levels and probabilities for detection and for alarm. Three different models for the water level have been considered.

The optimal alarm systems presented in the paper work well, and have the ability to reach any specified detection probability. The more conventional alarm algorithm that the optimal alarm is compared to, i.e. the alarm is sounded when the process reaches a certain level, has a maximal detection probability which in these cases is rather low. This means that if a high detection probability is required, the optimal alarm system has to be used. A drawback with a high detection probability is that the number of false alarms also becomes rather large, even though the optimal alarm systems give a minimum of false alarms. In particular the SETARX model for the water level shows this balance, it has a fast response and detect almost all catastrophes on all datasets, but at the expence of giving a high amount of false alarms, in particular on a dataset (from 1980) to which the model was not adapted.

A possibility to lower the number of false alarms is to find a better model, e.g. by using more external information for the predictions or by taking the timevariations of the water level process into account. Furthermore, in the flooding data case the prediction errors are not exactly normally distributed, which introduces further approximations in the calculations.

ÖZET

Bu çalışmada Gaussian ARMAX süreçleri için yanlış alarmların sayılarının minimize edilmesi anlamında optimal kestiriciler incelenmiştir. Sonuçlar Baltık Denizi'ndeki su seviyelerinin kestirimleri için uygulanmış ve burada verilen optimal alarm sisteminin daha basit sistemlere göre daha iyi sonuç verdiği gözlenmiştir.

İstatistik, Journal of the Turkish Statistical Association
1998, Vol.1, No. 2, pp. 47 - 57

ON SOME STATISTICS FOR MEASUREMENT OF DEPENDENCE AND/OR INDEPENDENCE OF TWO RANDOM VARIABLES

Yalçın Tuncer
Ankara University, Faculty of Science, Department of Statistics
06100, Tandogan, Ankara, Turkey

Abstract

Two basic statistics seem to have been introduced in literature to measure and detect a possible quadrant dependence between two random variables. This work compares these two statistics and dwells on one of these discussing distributional aspects for the empirical case. The results presented are preliminary findings of an ongoing research on the subject.

Key Words: Probability ratio, bivariate empirical processes, simple random walk, negative binomial (Pascal) probability.

1. Introduction

Two arbitrary random variables X and Y are in consideration. These variables have the respective given marginal distributions $F_X(x)$ and $F_Y(y)$ and a symmetric joint distribution $F(x,y)$ at the point $(x,y) \in R^2$, such that little information exist about the latter for reasons to be given later, the marginal and joint distributions are not assumed to converge concurrently to the same functional values as $x \to \pm\infty$ and (or) $y \to \pm\infty$.

When independence of X and Y to be emphasized, the joint distribution $F(x,y)$ will restrictively be denoted by $F^{(0)}(x,y)$, i.e.,

$$F(x,y) = F^{(0)}(x,y) \equiv F_X(x)F_Y(y); \tag{1.1}$$

$F^{(1)}(x,y)$ will in general stand for quadrant dependence case (c.f., for the concept, Lehmann (1966)), i.e.,

$$F^{(1)}(x,y) \neq F_X(x)F_Y(y). \tag{1.2}$$

When neither $F^{(1)}(x,y)$ nor $F^{(0)}(x,y)$ are to be emphasized, the joint distribution will be denoted by $F(x,y)$. To measure and detect dependence of X and Y , two statistics

seem to be favored in literature which form the bases of many other alternatives used for the same purpose:

$$\begin{array}{l} i)\ S_F(x,y) = \frac{F(x,y)}{F^{(0)}(x,y)}, \quad F^{(0)}(x,y) \neq 0\ , \quad (\ x,y) \in \Re_2 \\ ii)\ H_F(x,y) = F(x,y) - F^{(0)}(x,y), \quad (\ x,y) \in \Re_2 \end{array} \tag{1.3}$$

such that $S_F(x,y) = 1$ and $H_F(x,y) = 0$ discloses in dependence of X and Y at the point $(x,y) \in R^2$. To give some examples of other statistics based on $S_F(x,y)$, the *generalized likelihood (probability) ratio* is a typical one used in sequential analysis (c.f., Wald (1947), pp.37-61), i.e., $S_F(x,y) = \frac{F^{(1)}(x,y)}{F^{(0)}(x,y)}$; also, the *Kullback- Leibler separator* (c.f., Kullback-Leibler (1951)) is another example used in measurement of dependence, i.e.,

$$s_F(x,y) = \ln S_F(x,y) = \ln \frac{F^{(1)}(x,y)}{F^{(0)}(x,y)};$$

and we have also the *concentration ratio (function)* of Cifarelli and Regazzini (1987) and of Scarsini (1991), i.e.,

$$S_F^{(1)}(x,y) = \frac{F^{(1)}(x,y)}{F^{(0)}(x,y)}.$$

Similarly, it is well-known that

$$\sup_{(x,y)\in\Re^2} |H_F(x,y)| = \sup_{(x,y)\in\Re^2} \left|F(x,y) - F^{(0)}(x,y)\right|$$

corresponds to the *bivariate Kolmogorov-Smirnov statistic*;

$$[H_F\,(x,y)]^2 = \left[F\,(x,y) - F^{(0)}\,(x,y)\right]^2$$

yields the bivariate version of *Cramer-von Mises statistic* (cf., Kendall and Stuart (1973), p.467); and

$$\frac{[H_F(x,y)]^2}{F^{(0)}(x,y)} = \frac{\left[F\,(x,y) - F^{(0)}\,(x,y)\right]^2}{F^{(0)}(x,y)}$$

forms the well-known *chi-square statistic.*

In all of the above variations of $S_F(x,y)$ and $H_F(x,y)$, the statistics cannot distinguish between dependence and independence at the points $x \to \pm\infty$ and/or $y \to \pm\infty$, where $F(x,y)$ on the one hand, $F_X(x)$ and $F_Y(y)$ on the other converge concurrently to the same limiting value. Hence, these two basic statistics have similar performances at the extremes. This means that the restriction $F^{(0)}(x,y) \neq 0$ for $S_F(x,y)$ is not disadvantage for the statistic in question, because $H_F(x,y)$ has also the same disadvantage. However, because of the major works of Kolmogorov (1933) and Smirnov

(1935), and in view of time-proven applicability ease of chi-square, $H_F(x,y)$ seem to favored and known more in applications.

In the following sections, empirical counterparts $S_{F,n}(x,y)$ and $H_{F,n}(x,y)$ of the dependence statistics $S_F(x,y)$ and $H_F(x,y)$ will be discussed; then, an attempt will made to tackle the distributional features of the former empirical statistic $S_{F,n}(x,y)$. The results presented are preliminary findings of an ongoing research.

2. Empirical dependence statistics

When, on the other hand, all the distributions $F_X(x)$, $F_Y(y)$ and $F(x,y)$ are unknown and when we have a finite-sized (i.e., $n < \infty$) random sample of observations (X_i, Y_i), $i = 1, 2, \ldots, n$, on (X, Y), the following empirical counterparts of (1.3) are used:

$$S_{F,n}(x,y) = \frac{F_n(x,y)}{F_n^{(0)}(x,y)}, \quad F_n^{(0)}(x,y) \neq 0 \ , \tag{2.1}$$

$$H_{F,n}(x,y) = F_n(x,y) - F_n^{(0)}(x,y)$$

where, by definition (cf., Gaenssler and Stute(1985) and Tuncer (1995), we have

$$F_n(x,y) \ = \frac{1}{n}\sum_{i=1}^{n} I_{A\times B}(X_i \ , Y_i) = \frac{1}{n}\sum_{i=1}^{n} I_A(X_i)\cdot I_B(Y_i) \tag{2.2}$$

$$F_n^{(0)}(x,y) \ = F_{X,n}(x)\cdot F_{Y,n}(y) = \left(\frac{1}{n}\sum_{i=1}^{n} I_A(X_i)\right)\cdot\left(\frac{1}{n}\sum_{j=1}^{n} I_B(Y_j)\right)$$

with the usual indicator function for an arbitrary set E being defined as

$$I_E(\omega\ ) = \begin{cases} 1, \text{ if } \omega \in E, \\ 0, \text{ if } \omega \notin E, \end{cases} .$$

and $A = (-\infty, x]$ and $B = (-\infty, y]$.

For convenience of notation, we shall re-set

$$\begin{gathered} \zeta_i = I_{A\times B}(X_i, Y_i) = I_A(X_i) I_B(Y_i) = \xi_i \eta_i, \\ \xi_i = I_A(X_i) \text{ and } \eta_i = I_B(Y_i) \end{gathered} \tag{2.3}$$

which, as they will be easily noted, are independent Bernoulli trials yielding two-state Markovian processes

$$Z_m = \sum_{i=1}^{m} \zeta_i \ , \ T_m = \sum_{j=1}^{m} \xi_j \text{ and } U_m = \sum_{k=1}^{m} \eta_k \ , \tag{2.4}$$

$m = 0, 1, \ldots$, to be discussed in the next section.

Note however that for $n > 1$ and for any $(x, y) \in \Re_2$,the events that $S_{F,n}(x, y) = 1$ and $H_{F,n}(x, y) = 0$ are analytically negligible, inasmuch as,

$$\frac{1}{n}\sum_{k=1}^{n}\xi_k \cdot \eta_{\star k} \neq \left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) \cdot \left(\frac{1}{n}\sum_{j=1}^{n}\eta_j\right), \tag{2.5}$$

unless of course all ξ's (or all η_j's or both) either vanish or are equal to unity —clearly, for $n = 1$, $S_n(x, y) \equiv 1$ — This point of negligibility is also supported by the fact that the set

$$M = \{(x, y) \mid S_n(x, y) = 1, (x, y) \in \Re_2\}$$

is a line in $\Re_2$ which has Lebesque measure zero, and therefore, for any distribution $G(x, y)$ that is absolutely continuous with respect to Lebesque measure, the probability measure is zero:

$$\mathcal{P}(M) = \int\int_{M} dG(x, y) = 0.$$

This undoubtedly is true for continuous distributions. Corresponding to the case where the quantities $\frac{1}{n}\sum_{k=1}^{n}\xi_k \cdot \eta_k$ and $\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) \cdot \left(\frac{1}{n}\sum_{j=1}^{n}\eta_j\right)$are each obtained from two distinct samples from an identical population, Karlin and Taylor(1981, pp. 113-116) find on the other hand that the probability $\mathcal{P}(M)$ in general is equal to $\frac{1}{2n-1}$. This point of negligibility warrants that there is almost no information lost when such points are discarded from the analysis.

In fact, as it is the case with $S_F(x, y)$ and $H_F(x, y)$ at $x \to \pm\infty$ and/or $y \to \pm\infty$, the empirical measures $S_{F,n}(x, y)$ and $H_{F,n}(x, y)$ will not discern dependence from independence at points $x_{n:n} = \max\{x_{1:n}, x_{2:n}, ..., x_{n:n}\}$ and/or $y_{n:n} = \max\{y_{1:n}, y_{2:n}, ..., y_{n:n}\}$. These observations yield thus negligible results.

Note that the distributions that govern $H_{F,n}(x, y)$ are well-worked-out in literature, but this is hardly the case with $S_{F,n}(x, y)$. The latter is a ratio of two positive integers, which may be statistically dependent on each other. As a matter of fact, when the numerator and denominator of $S_{F,n}(x, y)$ are both estimated from the same sample, then, as will be noted in(2.3), both will be based on identical observations, so that dependence will be inevitable. The form of such a dependence will be taken up in the next section. Nonetheless, when the numerator and denominator are each computed from two distinct (independent) samples, such a dependence is eliminated, so that it may be possible to enquire into some discrete distributions, say, (cf. Johnson and Kotz (1969, p.31))

$$\mathcal{P}[X = r/s] = (e-1)^2\left(e^{r+s}-1\right)^{-2}, \tag{2.6}$$

which has positive finite moments and where r and s positive integers. However, when based on the same distribution, the numerator and the denominator are dependent, so that the distributions that govern the numerator and the denominator of $S_{F,n}(x, y)$ must be studied separately, and ways must be sought to eliminate the dependence. This line of reasoning will be followed in the following pages:

Letting $N = \{1, 2, ..., n\}$, the set of first positive integers, then

$$S_{F,n}(x,y) = \frac{n \cdot Z_n}{T_n \cdot U_n} = \frac{Z_n^*}{V_n} ,$$
$$Z_n^* = n \cdot Z_n \text{ and } V_n = T_n . U_n$$

where, for a given n, Z_n, T_n and U_n are as in (2.4). Accordingly, $V_n \in N^2 = \{k \cdot l / \ k \in N \text{ and } l \in N\}$ and $Z_n \in N$, the latter holding under specific conditions to be discusses in the next section. Thus, $S_{F,n}(x,y)$ becomes a quotient of two inter-dependent positive integers, and in the next section, we shall discuss the distributions of Z_n as eliminated from its dependence on V_n and the subsquent section will shortly dwell on the distribution of V_n. Under the circumstance, it will be possible to enquire into potential uses of such distributions as in (2.6). The last point will be resumed in another article.

3. Bivariate empirical distribution $n \cdot F_n(x,y)$

Since the variable ζ_i in (2.3) is a Bernoulli trial which takes the value $\zeta_i = 1$ with $\mathcal{P}((X_i, Y_i) \in A \times B) = p$ and the value $\zeta_i = 0$ with $\mathcal{P}((X_i, Y_i) \notin A \times B) = q$, then their sums $Z_m = \sum_{i=1}^{m} \zeta_i$, $(m = 0, 1, 2, \ldots)$,will obviously be a simple random walk governed by the probability p. The sample size n is fixed, so that the process cannot continue indefinitely. Our concern will thus be the probability governing the sum $Z_n = n \cdot F_n^{(1)}(x,y) = (\sum_{i=1}^{n} \zeta_i)$.

Adopting the vectorial notation $\underset{\sim}{\xi}^t = (\xi_1, \xi_2, \ldots, \xi_n)$ and $\underset{\sim}{\eta}^t = (\eta_1, \eta_2, \ldots, \eta_n)$, where the components ξ_i's and η_i's, $(i = 1, 2, \ldots, n)$, are the indicator functions defined in (2.3), it is possible to define three distintc sums which are related to the three distinct empidistributions mentioned earlier (cf., 2.4):

$$Z_n = nF_n(x,y) = \underset{\sim}{\xi}^t \underset{\sim}{\eta} , \quad T_n = nF_{X,n}(x) = \underset{\sim}{\xi}^t \underset{\sim}{\xi} \quad \textit{and} \quad U_n = nF_{Y,n} = \underset{\sim}{\eta}^t \underset{\sim}{\eta} ,$$

where the sum Z_n corresponds to the bivariate Markovian process ς_i's ; the sums T_n and U_n result from the respective univariate Markovian processes ξ_i's and η_i's. On the other hand, the Cauchy-Schwarz inequality as applied to the case, i.e.,

$$(\underset{\sim}{\xi}^t \underset{\sim}{\eta})^2 \leq (\underset{\sim}{\xi}^t \underset{\sim}{\xi}) \cdot (\underset{\sim}{\eta}^t \underset{\sim}{\eta})$$

yields the following relations between these three sums:

$$(\underset{\sim}{\xi}^t \underset{\sim}{\eta}) \leq \min \left\{ (\underset{\sim}{\xi}^t \underset{\sim}{\xi}) , (\underset{\sim}{\eta}^t \underset{\sim}{\eta}) \right\}, \tag{3.1}$$

and all the more

$$(\underset{\sim}{\xi}^t \underset{\sim}{\eta}) \leq (\underset{\sim}{\xi}^t \underset{\sim}{\xi}) \cdot (\underset{\sim}{\eta}^t \underset{\sim}{\eta}). \tag{3.2}$$

The sum Z_n is thus bounded from above by the random quantity $r = \min\left\{(\underset{\sim}{\xi}^t\underset{\sim}{\xi}), (\underset{\sim}{\eta}^t\underset{\sim}{\eta})\right\}$ as in (3.1) above. As such, the sum Z_n can only take values $0, 1, 2, \ldots, r$; whereas the sums T_n and U_n can freely take values $0, 1, 2, \ldots, n$. Since these latter sums obey Binomial probability laws with respective parameters $(n, F_X(x))$ and $(n, F_Y(y))$ (cf., Karlin and Taylor(1981), pp. 112-123; Gaensler and Stute(1985), pp.1-9), one is tempted to think that Z_n is also governed by the Binomial law with parameters $(n, F(x,y))$. This, however, is in strict conflict with the random bound given by (3.1).

In fact, corresponding to the bivariate case, the random sums

$$Z_m = \sum_{i=1}^{m} \zeta_i, \;\; m = 0, 1, 2, ..., n, \quad \textit{with } Z_0 = \zeta_0 = 0, \tag{3.3}$$

display a two-state simple Markov process property, such that the process Z_m starts at Z_0 and is either absorbed at $Z_m = r$, $m < n$, or stopped whenever $0 < Z_n < r$. The fixed sample size n and the bound r play thus some important role on the stochastic behavior of the process.

From an intuitive standpoint, since n is fixed and since the bound r must be obeyed, Negative Binomial probability law seems to be appropriate for the issue in hand. This intuitive result is also supported by analytical methods often utilized in practice, i.e., by probability generating function technique (cf., Cox and Miller (1965) pp.22-75). In fact, the probability generating function of such a process (for analytical derivations, cf., Appendix) is

$$\Phi(\,t) = \left(\frac{tp}{1 - \,tq}\right)^r = \sum_{j=b}^{\infty} t^{\,j} \binom{j-1}{r-1} p^r (q)^{j-r}, \tag{3.4}$$

where $t \in \Re_1$, such that $t < q^{-1}$, and, as before,

$$\begin{gathered} r = \min\{\textstyle\sum_{i=1}^{n} \xi_i), (\sum_{i=1}^{n} \eta_i)\} > 0, \\ p = \mathcal{P}(\,X \leq x\;, Y \leq y\,) \\ q = 1 - p. \end{gathered} \quad . \tag{3.5}$$

The coefficient of t^n in (3.4) is instrumental for and is central to the subsequent discussions on obtainement of the probability governing Z_n which we aim to obtain. The coefficient corresponds to

$$\mathcal{P}(N = n; r, p) = \binom{n-1}{r-1} p^r q^{n-r}, \quad n \geq r \geq 1. \tag{3.6}$$

and $\mathcal{P}(N = n; r, p)$ in (3.6) is the probability mass function of a random variable N, which represents varying sample size and yields the relative frequency of the number of observations required to obtain $Z_n = r$ is n. The quantity r is the parameter of this mass function, which, as it will be noted in (3.1) above, itself is random. Another well-known version of (3.6) is

$$\mathcal{P}(W = w; r, p) = \binom{w+r-1}{w} p^r q^w, \; w \geq 0 \text{ and } r \geq 1. \tag{3.7}$$

which stands for the relative frequency of the waiting time w (in terms of failures) to $Z_n = r$.

As it is noted earlier, the positive integer r is random, and therefore, the probabilities in (3.6) and (3.7) must be compounded with the distribution of this random positive integer. The distribution of R is known to be Binomial with parameters n and π, where π is either $\mathcal{P}(X \leq x) = F_X(x)$ or $\mathcal{P}(Y \leq y) = F_Y(y)$,depending on $\min\{F_X(x), F_Y(y)\}$. However, since $r > 0$, the relevant Binomial distribution is the so-called *modified Binomial distribution,* in which only positive Binomial observations are considered (cf., for the concept, Johson and Kotz (1969), pp.204-209):

$$\mathcal{P}(R^* = r; n, \pi) = \frac{\mathcal{P}(R=r;n,\pi)}{1-\mathcal{P}(R=0;n,\pi)} = \frac{1}{1-(1-\pi)^n}\binom{n}{r}\pi^r(1-\pi)^{n-r}, \quad r = 1, 2, \ldots, n \tag{3.8}$$

where R^* is the modified Binomial variate and R is the standard Binomial variate. The corresponding probability function is:

$$\Phi_{R^*}(t) = \alpha + (1-\alpha) \cdot \Phi_R(t),$$

with $\alpha = -\frac{(1-\pi)^n}{1-(1-\pi)^n}$ and $\Phi_R(t) = ((1-\pi) + \pi t)^n$, so that

$$\Phi_{R^*}(t) = \frac{(1-\pi)^n}{1-(1-\pi)^n} \sum_{k=1}^{n} \binom{n}{k} (\frac{\pi}{1-\pi})^k \, t^k. \tag{3.9}$$

Accordingly, the compound distribution has reprective probability and moment generating functions

$$\begin{aligned} \Phi_{W_n^*}(t) &= \alpha + (1-\alpha) \cdot \Phi_{R^*}(\Phi_{W_n}(t)) \\ &= \frac{(1-\pi)^n}{1-(1-\pi)^n} \sum_{r=1}^{n} \binom{n}{r} (\frac{\pi}{1-\pi})^r (\Phi_{W_n}(t))^r \\ &= \frac{(1-\pi)^n}{1-(1-\pi)^n} \sum_{r=1}^{n} \binom{n}{r} (\frac{\pi}{1-\pi})^r (\frac{p}{1-qt})^r \end{aligned} \tag{3.10}$$

$$M_{W^*}(t) = \frac{(1-\pi)^n}{1-(1-\pi)^n} \sum_{k=1}^{n} \binom{n}{k} (\frac{\pi}{1-\pi})^k (\frac{p}{1-qe^t})^k. \tag{3.11}$$

Hence, the mean can be calculated to be equal to

$$\mathcal{E}(W^*) = \frac{nq}{p} \cdot \frac{(1-\pi)^n}{1-(1-\pi)^n} \cdot \sum_{r=1}^{n} \binom{n-1}{r-1} (\frac{\pi}{1-\pi})^r, \tag{3.12}$$

and the variance is given by

$$\mathcal{V}(W_n^*) = \frac{nq^2}{p^2} \cdot \frac{(1-\pi)^n}{1-(1-\pi)^n} \cdot \sum_{r=1}^{n} (r+1)\binom{n-1}{r-1} (\frac{\pi}{1-\pi})^r + \mathcal{E}(W_n^*)(1-\mathcal{E}(W_n^*)). \tag{3.13}$$

It is only obviously a routine matter to calculate $\mathcal{E}(Z_n)$ and $\mathcal{V}(Z_n)$ through the relation $Z_n + W_n^* = n$. Also it should be noted that, in (3.10), instead of the standard characteristic function of w, we can used the characteristic function of the modified Negative Binomial variate w. where $w = 0$ is eliminated.

From the computational standpoint, both versions of Negative Binomial noted in (3.6) and (3.7) above are known to have some affinity with a quotient of Incomplete Beta functions, and hence Binomial distribution. In fact, the variables Z, W and N are related through $N = Z + W$, so that the event $\{Z_n < r\}$ in the Binomial case is related to to the event $\{n < N\}$ in the Negative Binomial case, etc., (cf., for details, Kendall and Stuart (1969), p.130 ; Johson and Kotz (1969), p.127). As such, computation of Negative Binomial probabilities boils down to calculation of the corresponding Beta and hence Binomial percentage points. Nonetheless, separate tables of percentage points are available (cf., Williamson and Bretherton (1963). Furthermore, avenues of Poisson approximation (cf., Rahman (1968), p.218), of Gamma approximation (cf., Woodroofe (1975), p.109) and of Normal approximation (cf., Wilks (1962), p.274) as well as other computational possibilities (cf., Johnson and Kotz(1969), pp.127-131) also exist.

4. Distributions of V_n

Since V_n is the product of T_n and U_n, both being positive integers in N, and since the distribution of these latter variables are modified Binomial distributions in which $T_n = 0$ and $U_n = 0$ are eliminated, then V_n will be a positive integer in N^2 which contains some of the positive integers ranging from one to n^2 and leaves out some like $n^2 - n + 1, n^2 - n + 2, ..., n^2 - 1$. Accordingly, the distribution of V_n will be a discrete distribution in whivh the unit mass is concentrated at points of N^2. The distributions of T_n and U_n in the through the distributionsof T_n and U_n in the usual way:

$$\begin{aligned}\mathcal{P}(V = v) &= \sum_{t=\frac{v}{n}}^{n} \mathcal{P}(t, \frac{v}{t}) \\ &= \sum_{t=\frac{v}{n}}^{n} \mathcal{P}(T = t) \cdot \mathcal{P}(u = \frac{v}{t}) \\ &= AB \sum_{t=\frac{v}{n}}^{n} \binom{n}{t} \binom{n}{\frac{v}{t}} \psi^t \phi^{\frac{v}{t}} \qquad (4.1)\end{aligned}$$

where $AB = \frac{(1-\delta)^n(1-\theta)^n}{(1-(1-\delta)^n)(1-(1-\theta)^n)}$, $\psi = \frac{\delta}{1-\delta}$ and $\phi = \frac{\theta}{1-\theta}$
with $\delta = F_X(x)$ and $\theta = F_Y(y)$.

Acknowledgement

I would like to thanks Drs. Ismihan Bairamov, Yilmaz Akdi and Bilgehan Güven for helpful suggestions on the subject.

References

[1] Cifarelli, D.M.and Regazzini, E.(1987). "On a general definition of concentration function". *Sankhyā,* Vol. B 49, pp. 307-319.

[2] Cox, D.R. and Miller, H.D.(1965). *The Theory of Stochastic Processes.* Chapman and Hall, London.

[3] Gaenssler, P. and Stute, W.(1987). *Seminar on Empirical Processes (*DMV seminar; Bd.9). Birkhauser-Verlag, Basel

[4] Johnson, N.L.(1969). *Distributions in Statistics: Discrete Distributions.* Houghton Mifflin Company, Boston.

[5] Karlin, S. and Taylor, H.M.(1981). *A Second Course in Stochastic Processes.* Academic Press, Inc., San Diego, California.

[6] Kendall, M.G. and Stuart,A.(1969). *The Advanced Theory of Statistics:Distribution Theory.* Third edition. Hafner Publishing Company, New York.

[7] Kendall, M.G. and Stuart,A.(1973). *The Advanced Theory of Statistics:Inference and Relationship.* Third edition. Hafner Publishing Company, New York.

[8] Kolmogorov, A.(1933). Sulla determinazione empirica di una legge de distribuzione.*Giorn. Inst. Ital.Altuari,* Vol.4, pp.83-91.

[9] Kullback, S. and Liebler, R.A.(1951). On Information and suffiency. *Annals of Mathematical Statistics,* Vol. 22, pp.79-86.

[10] Rahman, N.A.(1968). *A Course in Theoretical Statistics.* Hafner Publishing Company, New York.

[11] Scarsini, M.(1990). "An ordering of dependence" in *Topics in Statistical Dependence,* edit. by H. Block et. al., Institute of Mathematical Statistics Lecture Notes and Monograph Series, Vol.10, pp.403-414.

[12] Smirnov, N.V.(1935). Über die Verteilung des allgemeinen Giedes in der Variationsreihe. *Metron,* Vol.12, pp.59-81.

[13] Wald, A.(1947). *Sequential Analysis.* John Wiley and Sons, Inc., New York.

[14] Wilks, S.S.(1962). *Mathematical Statistics.* Wiley International Edition. John Wiley and Sons, Inc., New York.

[15] Williamson, E. and Bretherton, M.H.(1963). *Tables of Negative Binomial Probability Distribution. John Wiley and Sons, London.*

[16] Woodroofe, M.(1975). *Probability with Applications.* McGraw-Hill, New York.

[17] Tuncer, Y.(1995). On some measures of dependence and complete dependence. *Journal of Applied Statistical Science,* Vol. **3**, pp.107-128.

Appendix

Let Z_0 stand for the initial stage of the process, so that we may have either $Z_0 = 0$ or $Z_0 = r > 0$. In the latter case, with probability one, the process starts and ends automatically, and this will be denoted with

$$p^{(0)}_{(r,r)} = P(\textit{pocess starts and ends at initial stage}) = 1,$$

where the superscript (0) indicates the stage 0 and the subscript (r, r) shows that the process starts at r and ends at r. In this appendix, we shall be concerned with probabilities like $p^{(n)}_{(0,r)}$ and shall seek the generating function which produces them.When the process starts at $Z_0 = 0$, Z_m for any $m = 1, 2, \ldots, n$ can take the values (can reach the states) $0, 1, \ldots, r$ with probabilities$p^{(m)}_{(0,0)}$, $p^{(m)}_{(0,1)}$,. . ., $p^{(m)}_{(\ 0,r)}$. Likewise, for an intermediary state c, $0 < c < r$, we have

$$p^{(m)}_{(c,r)} = P(c \leq X_1, X_2, \ldots, X_{m-1} < r, X_m = r \mid X_0 = c).$$

Clearly, since a process cannot start at c and end at r at the initial stage whenever $c \neq r$, we shall have

$$p^{(0)}_{(c,r)} = 0.$$

Together with $p^{(0)}_{(r,r)} = 1$, $p^{(0)}_{(c,r)} = 0$ forms initial conditions. Note that a markovian process has the property that the realization of Z_s, $s < t$, does not depend on the realization of X_u, $t < u$. Accordingly, if we start at the intermediary point c,i.e., $Z_0 = c$, and then the next will be

$$\zeta_1 = Z_1 = \begin{cases} c+1, \textit{with } p = \mathcal{P}((X_1, Y_1) \in A \times B) \\ c, \quad \textit{with } q = \mathcal{P}((X_1, Y_1) \in (A \times B)^c) \end{cases}, \tag{A.1}$$

where A and B will respectively be, say, $A = (-\infty, x\ ]$ and $B = (-\infty, y\ ]$, and this will be independent of the realizations of all $(m-1)$ subsequent steps. Therefore, when $Z_1 = c+1$, the $(m-1)^{th}$ step will be independent, so that we can multiply the relevant probabilities to have

$$p^{(1)}_{(c,c+1)} \cdot p^{(m-1)}_{(c+1,r)}. \tag{A.2}$$

Similarly, when $Z_1 = c$, the next steps will be independent, so that we have

$$q^{(1)}_{(c,c)} \cdot p^{(m-1)}_{(c,r)}. \tag{A.3}$$

Obviously, these two events are mutually exclusive, and these imply that. $Z_m = b$. Thus, adding (A.1) and (A.2), we obtain $p^{(m)}_{(c,r)}$

$$p^{(m)}_{(c,r)} = p^{(1)}_{(c,c+1)} \cdot p^{(m-1)}_{(c+1,r)} + q^{(1)}_{(c,c)} \cdot p^{(m-1)}_{(c,r)}$$

which denotes the probability that the process starts at c and is absorbed at r at the m^{th} stage. Clearly, $p_{(c,r)}^{(m)}$ is an unknown function of two discrete variables m and c. By definition, the probability generating function which will produce $p_{(c,r)}^{(m)}$ is

$$\varphi_{(c,r)}(t) = \sum_{m=0}^{\infty} t^m p_{(c,r)}^{(m)} = t\left[p \cdot \sum_{m=1}^{\infty} t^{m-1} p_{(c+1,r)}^{(m-1)} + q \cdot \sum_{m=1}^{\infty} t^{m-1} p_{(c,r)}^{(m)}\right]$$

because, as stated earlier in connection with initial conditions, $p_{(c,r)}^{(0)} = p_{(c+1,r)}^{(0)} = 0$. As such,

$$\varphi_{(c,r)}(t) = t \cdot p\left[\varphi_{(c+1,r)}(t)\right] + t \cdot q\left[\varphi_{(c,r)}(t)\right],$$

which is a linear homogeneous difference equation in $\varphi(t)$ with the boundary condition

$$\varphi_{(r,r)}(t) = 1, \qquad (A.4)$$

due to the facts that $p_{(r,r)}^{(0)} = 1$ set initially and that $p_{(r,r)}^{(m)} = 0$, i.e., the process is not recurrent. A solution of this difference equation is given by

$$\varphi_{(c+1,r)}(t) = \left(\frac{1 - t \cdot q}{t \cdot p}\right) \cdot \varphi_{(c,r)}(t).$$

The equation now is free from m and is a function of t only. Thus,

$$\varphi_{(c,r)}(t) = \left(\frac{1 - t \cdot q}{t \cdot p}\right)^{c} \cdot \varphi_{(0,r)}(t),$$

i.e.,

$$\varphi_{(r,r)}(t) = \left(\frac{1 - t \cdot q}{t \cdot p}\right)^{r} \cdot \varphi_{(0,r)}(t).$$

However, by the boundary condition (A.4) above $\varphi_{(r,r)}(t) = 1$, so that

$$\varphi_{(0,r)}(\quad = \left(\frac{t \cdot p}{1 - t \cdot q}\right)^{r},$$

as stated in the text.

ÖZET

İki rasgele değişken arasında mümkün olan dairesel (quadrant) bağımlılık için literatürdeki iki istatistik tanıtılmıştır. Bu çalışmada empirik durumlar için bu iki istatistik karşılaştırılmış ve dağılımları incelenmiştir. Burada bulunan sonuçlar bir başlangıç olup bu alanda çalışmalar devam etmektedir.

İstatistik, Journal of the Turkish Statistical Association
1998, Vol.1, No. 2, pp. 59 - 66

CHARACTERIZATION OF GEOMETRIC DISTRIBUTION THROUGH WEAK RECORDS

Fazil A. Aliev
Ankara University, Faculty of Science, Department of Statistics
06100, Tandogan, Ankara, Turkey,
Baku State University, Faculty of Applied Mathematics, Azerbaijan

Abstract

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables (r.v.'s) taking on values $0, 1, \ldots$ with a distribution function F such that $F(n) < 1$ for any $n = 0, 1, \ldots$ and $EX_1 < \infty$. Let $X_{L(n)}$ be the n-th weak record value. In this paper we show that X_1 has a geometric distribution iif $E(X_{L(n+2)} - X_{L(n)} \mid X_{L(n)} = i) = \alpha$, for some $n > 0, \alpha > 0$ and for all $i \geq 0$.

Keywords: records, weak records, characterization of geometric distribution.

1. Introduction

A lot of papers in the field of records are devoted to characterizations of distributions via records (see $[2-9]$, and also the references in $[1], [7]$ among many others). Great interest in records exists because we often come across with them in our everyday life in such a way that singling out and fixing of record values.

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables (r.v.'s) taking on values $0, 1, \ldots$ with a distribution function F such that $F(n) < 1$ for any $n = 0, 1, \ldots$ and $EX_1 < \infty$. Define the sequence of weak record times $L(n)$ and weak record values $X_{L(n)}$ as follows:

$$L(1) = 1, \quad L(n+1) = \min\left\{j > L(n) : X_j \geq X_{L(n)}\right\}, \qquad n = 1, 2, \ldots \tag{1}$$

If we replace the sign $\geq$ by $>$ in (1), then we obtain record times and record values instead of weak record times and weak record values. Denote $p_k = P\{X_1 = k\}$ and $\overline{F}(k) = 1 - F(k) \quad (k \geq 0)$.

It is known that

1) X_1 has a distribution of the form

$$P\{X_1 \geq m\} = \left(\prod_{i=1}^{m}(\alpha+(i-1)\beta)\right)\left(\prod_{i=1}^{m}(1+\alpha+i\beta)\right)^{-1}$$

for some $\alpha > 0, \beta \geq 0$, and $m = 1, 2, ...$, iif $E(X_{L(n+1)} - X_{L(n)} \mid X_{L(n)} = s) = \alpha + \beta s$, for all $s = 0, 1, ...(n > 0)$ (Stepanov(1994)). If $\beta = 0$ this result corresponds to the geometric distribution,

2) if $\{A_i\}_{i=0}^{\infty}$ is any sequence of positive numbers such that $A_{i-1}/(1+A_i) < 1$ for all i and $\prod_{i=1}^{\infty} A_i/(1+A_i) = 0$ then X_1 has distribution of the form $P\{X_1 \geq m\} = \prod_{i=1}^{m} A_{i-1}/(1+A_i)$

for all $m = 1, 2, ...$ iif $E\left\{X_{L(n+1)} - X_{L(n)} \mid X_{L(n)} = s\right\} = A_s$ for all $s = 0, 1, ...(n \geq 1)$ (Aliev(1998)). In the case of $A_s = \alpha + \beta s$ from this result implies the above result of Stepanov(1994).

In this paper we first give a characterization of geometric distribution in terms of $E\left\{X_{L(n+2)} - X_{L(n)} \mid X_{L(n)} = s\right\}$ instead of $E\left\{X_{L(n+1)} - X_{L(n)} \mid X_{L(n)} = s\right\}$.

2. Characterization Theorem

Theorem. *A necessary and sufficient condition for a r.v. X_1 to have a geometric distribution is that*

$$E\left\{X_{L(n+2)} - X_{L(n)} \mid X_{L(n)} = s\right\} = \alpha \quad \textit{for some } n \geq 1,\ \alpha > 0 \ \textit{and all } s = 0, 1, ... \tag{2}$$

Proof. Consider the probability $P\left\{X_{L(n+2)} - X_{L(n)} = k, X_{L(n)} = s\right\}$ $(k, s \geq 0)$. We have

$$P\left\{X_{L(n+2)} - X_{L(n)} = k, X_{L(n)} = s\right\} = P\left\{X_{L(n+2)} = k+s, X_{L(n)} = s\right\} =$$

$$= \sum_{d=n}^{\infty} P(X_{L(n+2)} = k+s, X_{L(n)} = s, L(n) = d) =$$

$$= \sum_{d=n}^{\infty}\sum_{m=d+2}^{\infty} P(X_{L(n+2)} = k+s, X_{L(n)} = s, L(n) = d, L(n+2) = m) =$$

$$= \sum_{d=n}^{\infty}\sum_{m=d+2}^{\infty}\sum_{l=d+1}^{m-1} P(X_m = k+s, X_d = s, L(n) = d, L(n+1) = l, L(n+2) = m) = \tag{3}$$

$$= \sum_{d=n}^{\infty}\sum_{m=d+2}^{\infty}\sum_{l=d+1}^{m-1}\sum_{t=s}^{k+s} P(X_m = k+s, X_d = s, X_{L(n+1)} = t, L(n) = d, L(n+1) =$$

$$l, L(n+2) = m) =$$

$$= \sum_{d=n}^{\infty} \sum_{m=d+2}^{\infty} \sum_{l=d+1}^{m-1} \sum_{t=s}^{k+s} P(X_m = k+s, X_d = s, X_l = t, L(n) = d, L(n+1) = l, L(n+2) = m).$$

Probability under summation may be rewritten as

$$P(X_m = k+s, X_d = s, X_l = t, L(n) = d, L(n+1) = l, L(n+2) = m) =$$

$$= P(X_d = s, L(n) = d, X_{d+1} < s, ..., X_{l-1} < s, X_l = t, \tag{4}$$
$$X_{l+1} < t, ..., X_{m-1} < t, X_m = k+s).$$

Note that the event $\{X_d = s, L(n) = d\}$ is defined only by the r.v.'s $X_1, X_2, ..., X_d$ and, therefore, is independent of

$$\{X_{d+1} < s, ..., X_{l-1} < s, X_l = t, X_{l+1} < t, ..., X_{m-1} < t, X_m = k+s\},$$

consequently, from (4) we have

$$P(X_m = k+s, X_d = s, X_l = t, L(n) = d, L(n+1) = l, L(n+2) = m) =$$

$$= P(X_d = s, L(n) = d) \times P(X_{d+1} < s, ..., X_{l-1} < s,$$
$$X_l = t, X_{l+1} < t, ..., X_{m-1} < t, X_m = k+s) = \tag{5}$$
$$= P(X_d = s, L(n) = d) \cdot F^{l-d-1}(s) \cdot F^{m-l-1}(t) \cdot P(X_l = t) \cdot P(X_m = k+s) =$$
$$= p_t p_{k+s} P(X_d = s, L(n) = d) \cdot F^{l-d-1}(s) \cdot F^{m-l-1}(t).$$

From (3) and (5) changing the order of summation for t, l and m one can write

$$P\left\{X_{L(n+2)} - X_{L(n)} = k, X_{L(n)} = s\right\} =$$

$$= \sum_{d=n}^{\infty} \sum_{m=d+2}^{\infty} \sum_{l=d+1}^{m-1} \sum_{t=s}^{k+s} \left\{p_t p_{k+s} P(X_d = s, L(n) = d) \cdot F^{l-d-1}(s) \cdot F^{m-l-1}(t)\right\} =$$

$$= p_{k+s} \cdot \sum_{d=n}^{\infty} P(X_d = s, L(n) = d) \sum_{t=s}^{k+s} p_t \sum_{l=d+1}^{\infty} F^{l-d-1}(s) \sum_{m=l+1}^{\infty} F^{m-l-1}(t). \tag{6}$$

Using the obvious facts

$$\sum_{m=l+1}^{\infty} F^{m-l-1}(t) = \frac{1}{\overline{F}(t)}, \quad \sum_{l=d+1}^{\infty} F^{l-d-1}(s) = \frac{1}{\overline{F}(s)} \text{ and}$$

$$P(X_{L(n)} = s) = \sum_{d=n}^{\infty} P(X_d = s, L(n) = d) \text{ with (6) we obtain}$$

$$P\left\{X_{L(n+2)} - X_{L(n)} = k, X_{L(n)} = s\right\} = P(X_{L(n)} = s) \cdot p_{k+s} \cdot \frac{1}{\overline{F}(s)} \sum_{t=s}^{k+s} \frac{p_t}{\overline{F}(t)},$$

or, equivalently, for the conditional probability, we may write

$$P\left\{X_{L(n+2)} - X_{L(n)} = k \mid X_{L(n)} = s\right\} = p_{k+s} \cdot \frac{1}{\overline{F}(s)} \sum_{t=s}^{k+s} \frac{p_t}{\overline{F}(t)}. \tag{7}$$

Note that since this probability does not depend on n, we may, without loss of generality, assume that $n = 1$.

From (7) the conditional expectation is

$$E\left\{X_{L(3)} - X_1 \mid X_1 = s\right\} = \frac{1}{\overline{F}(s)} \cdot \sum_{k=0}^{\infty} \left(k\ p_{k+s} \sum_{l=s}^{k+s} \frac{p_l}{\overline{F}(l)} \right). \tag{8}$$

By changing the order of summation in (8) and taking $k + s = z$ one can write

$$E\left\{X_{L(3)} - X_1 \mid X_1 = s\right\} = \frac{1}{\overline{F}(s)} \sum_{l=s}^{\infty} \left(\frac{p_l}{\overline{F}(l)} \cdot \sum_{z=l}^{\infty} ((z-s)\ p_z) \right) =$$

$$= \frac{1}{\overline{F}(s)} \sum_{l=s}^{\infty} \left(\frac{p_l}{\overline{F}(l)} \cdot \left(\sum_{z=l}^{\infty} (z\ p_z) - s\overline{F}(l) \right) \right) = \sum_{l=s}^{\infty} \sum_{z=l}^{\infty} \frac{z p_z p_l}{\overline{F}(s)\overline{F}(l)} - s.$$

Therefore, the basic formula for the conditional expectation for future references is

$$E\left\{X_{L(3)} - X_1 \mid X_1 = s\right\} = \sum_{l=s}^{\infty} \sum_{z=l}^{\infty} \frac{z p_z p_l}{\overline{F}(s)\overline{F}(l)} - s. \tag{9}$$

Necessity. Let X_1 has a geometric distribution with $p_k = P(X_1 = k) = pq^k$ $(k = 0, 1, ..., \quad q = 1 - p)$. Then it is obvious that $\overline{F}(s) = q^s$ for all $s \geq 0$. Using the known formula $\sum_{z=l}^{\infty} (zq^z) = \frac{lpq^l + q^{l+1}}{p^2}$ and (9), it may be trivially seen that

$$E\left\{X_{L(3)} - X_1 \mid X_1 = s\right\} = \sum_{l=s}^{\infty} \sum_{z=l}^{\infty} \frac{zpq^z pq^l}{q^s q^l} - s = \frac{p^2}{q^s} \sum_{l=s}^{\infty} \sum_{z=l}^{\infty} (zq^z) - s =$$

$$= \frac{p^2}{q^s} \sum_{l=s}^{\infty} \frac{lpq^l + q^{l+1}}{p^2} - s = \frac{1}{q^s} \sum_{l=s}^{\infty} (lpq^l + q^{l+1}) - s =$$

$$= \frac{p}{q^s} \sum_{l=s}^{\infty} (lq^l) + \frac{1}{q^s} \sum_{l=s}^{\infty} q^{l+1} - s = \frac{p}{q^s} \frac{spq^s + q^{s+1}}{p^2} + \frac{q}{p} - s = \frac{2q}{p}$$

which proves the necessity part of the theorem.

Sufficiency. Let condition (2) hold. Also using (9), we take the equality

$$\sum_{l=s}^{\infty} \sum_{z=l}^{\infty} \frac{z p_z p_l}{\overline{F}(s)\overline{F}(l)} - s = \alpha,$$

or, equivalently,

$$B_s \equiv \sum_{l=s}^{\infty}\sum_{z=l}^{\infty}\frac{zp_zp_l}{\overline{F}(s)\overline{F}(l)} = \alpha + s \qquad \text{for all } s \geq 0. \tag{10}$$

Rewriting (10) for $s = k$ we have

$$\alpha + k = B_k \equiv \sum_{l=k}^{\infty}\sum_{z=l}^{\infty}\frac{zp_zp_l}{\overline{F}(k)\overline{F}(l)} = \sum_{z=k}^{\infty}\frac{zp_zp_k}{\overline{F}(k)\overline{F}(k)} + \sum_{l=k+1}^{\infty}\sum_{z=l}^{\infty}\frac{zp_zp_l}{\overline{F}(k)\overline{F}(l)} =$$

$$= \frac{p_k}{\overline{F}^2(k)}\sum_{z=k}^{\infty}(zp_z) + \sum_{l=k+1}^{\infty}\sum_{z=l}^{\infty}\frac{zp_zp_l}{\overline{F}(k)\overline{F}(l)} = \tag{11}$$

$$= \frac{p_k}{\overline{F}^2(k)}\sum_{z=k}^{\infty}(zp_z) + \frac{\overline{F}(k+1)}{\overline{F}(k)}\sum_{l=k+1}^{\infty}\sum_{z=l}^{\infty}\frac{zp_zp_l}{\overline{F}(k+1)\overline{F}(l)} = \frac{p_k}{\overline{F}^2(k)}\sum_{z=k}^{\infty}(zp_z) + \frac{\overline{F}(k+1)}{\overline{F}(k)}B_{k+1}.$$

By the condition (10) $B_{k+1} = \alpha + k + 1$, therefore, from (11) for all $k \geq 0$ one can write

$$\alpha + k = \frac{p_k}{\overline{F}^2(k)}\sum_{z=k}^{\infty}(zp_z) + \frac{\overline{F}(k+1)}{\overline{F}(k)} \cdot (\alpha + k + 1) \tag{12}$$

Observing that $\overline{F}(k+1) = \overline{F}(k) - p_k$ (12) gives the identity

$$\alpha + k = \frac{p_k}{\overline{F}^2(k)}\sum_{z=k}^{\infty}(zp_z) + \frac{\overline{F}(k) - p_k}{\overline{F}(k)} \cdot (\alpha + k + 1),$$

or, equivalently,

$$\alpha + k = \frac{p_k}{\overline{F}^2(k)}\sum_{z=k}^{\infty}(zp_z) + \alpha + k + 1 - \frac{p_k}{\overline{F}(k)} \cdot (\alpha + k + 1).$$

From the last identity we may write

$$p_k\left(\alpha + k + 1 - \frac{1}{\overline{F}(k)}\sum_{z=k}^{\infty}(zp_z)\right) = \overline{F}(k) \quad \text{for all } k \geq 0. \tag{13}$$

Here in the case of $k = 0$ we have

$$EX_1 + \frac{1 - p_0}{p_0} = \alpha, \quad \text{or} \quad \sum_{z=1}^{\infty}(zp_z) + \frac{1 - p_0}{p_0} = \alpha$$

and rewriting (13) in the form

$$p_k = \frac{1 - (p_0 + \ldots p_{k-1})}{\alpha + k + 1 - \frac{1}{1-(p_0+\ldots p_{k-1})}\left(\alpha - \frac{1-p_0}{p_0} - \sum_{z=1}^{k-1}(zp_z)\right)}$$

we have recurrent relation for determining p_k for any given p_0. It is clear that the set of probabilities $p_0, p_1, p_2, ...$ must satisfy the conditions

$$\begin{gathered} p_0 + p_1 + p_2 + p_3 + ... = 1 \text{ and} \\ \sum_{z=1}^{\infty}(zp_z) + \frac{1-p_0}{p_0} = \alpha. \end{gathered} \tag{14}$$

For proving that such set of $p_0, p_1, p_2, ...$ exists and is unique rewrite (13) in terms of $\overline{F}(k)$. Having the obvious equality

$$\frac{1}{\overline{F}(k)}\sum_{z=k}^{\infty}(zp_z) = \frac{kp_k + (k+1)p_{k+1} + (k+2)p_{k+2} + ...}{\overline{F}(k)} =$$

$$= (k-1) + \frac{\overline{F}(k) + \overline{F}(k+1) + \overline{F}(k+2) + \overline{F}(k+3) + ...}{\overline{F}(k)}$$

with (13) and the identity $p_k = \overline{F}(k) - \overline{F}(k+1)$,we reach

$$(\overline{F}(k) - \overline{F}(k+1))\left(\alpha + 2 - \frac{\overline{F}(k) + \overline{F}(k+1) + \overline{F}(k+2) + \overline{F}(k+3) + ...}{\overline{F}(k)}\right) = \overline{F}(k).$$

Last equality may be equivalently changed as

$$\alpha\overline{F}(k) = \overline{F}(k+1) + \overline{F}(k+2) + \overline{F}(k+3) + ... + \frac{\overline{F}(k)\overline{F}(k+1)}{\overline{F}(k) - \overline{F}(k+1)}. \tag{15}$$

Now using (15) for k and $k+1$ and subtracting we take

$$\alpha\left(\overline{F}(k) - \overline{F}(k+1)\right) - \overline{F}(k+1) - \frac{\overline{F}(k)\overline{F}(k+1)}{\overline{F}(k) - \overline{F}(k+1)} = -\frac{\overline{F}(k+1)\overline{F}(k+2)}{\overline{F}(k+1) - \overline{F}(k+2)}.$$

Denote $\beta_k = \overline{F}(k+1)/\overline{F}(k)$ $(k \geq 0)$, noting that $\overline{F}(k+1) = \beta_k\overline{F}(k)$ and $\overline{F}(k+2) = \beta_k\beta_{k+1}\overline{F}(k)$ from the previous formula we have the recurrent relation for β_k

$$\beta_{k+1} = 1 + \frac{\beta_k(1-\beta_k)}{\alpha(1-\beta_k)^2 - 3\beta_k + 2\beta_k^2}. \tag{16}$$

Consider the second part of conditions (14) for β_k. Note that $\beta_0 = 1 - p_0 = \overline{F}(1)$, $\beta_0\beta_1...\beta_k = \overline{F}(k+1)$ and $0 \leq \beta_k \leq 1$ for all k. Together with the second part of (14) and the known equalities for expectation

$$\sum_{z=1}^{\infty}(zp_z) = EX_1 = \sum_{k=1}^{\infty}\overline{F}(k),$$

we have

$$\beta_0 + \beta_0\beta_1 + ... + \beta_0\beta_1...\beta_k + ... + \frac{\beta_0}{1-\beta_0} = \alpha. \tag{17}$$

In this step first note that taking $\beta_0 = \frac{\alpha}{\alpha+2}$ in (16) we have $\beta_0 = \beta_1 = \beta_2 = ... = \frac{\alpha}{\alpha+2}$ which implies that $\overline{F}(k) = (\frac{\alpha}{\alpha+2})^{k-1}$. Thus as well as one solution for (16)-(17) (also satisfying (13)-(14) and (15)) exists. This solution corresponds to the geometric distribution with $p_0 = p = \frac{2}{\alpha+2}$. Let us show that this solution is unique. Consider the real valued function $f(x) = 1 + \frac{x(1-x)}{\alpha(1-x)^2-3x+2x^2}$ $(0 \leq x \leq 1)$ with two points of discontinuity. For all continuity points x of $f(x)$ we may write

$$f'(x) = \frac{\alpha(1-x)^2 + x^2}{(\alpha(1-x)^2 - 3x + 2x^2)^2} > 0.$$

Therefore $f(x)$ is monotonly increasing function in continuity intervals. Let x_1 and x_2 $(x_1 \leq x_2)$ be the discontinuity points of $f(x)$. It may be verified that this points are different, $x_1 \in (0,1)$ and $x_2 > 1$ for any $\alpha > 0$. Furthermore, $f(x) > 1$ for $0 < x < x_1$ and from (17) we may have $\beta_0 > 0$ and $\beta_0 < 1$. (16) may be written as $\beta_{k+1} = f(\beta_k)$ and we have $\beta_0 > x_1$,because in vice versa we have $\beta_1 > 1$, which contradicts with condition $0 \leq \beta_1 \leq 1$. By the same way it may be seen that from the condition $0 < \beta_1 = f(\beta_0)$ we have $\beta_0 > 1 - (\alpha+1)^{-1/2}$. Note that last point $1-(\alpha+1)^{-1/2}$ is the small one of two roots of equation $f(x) = 0$. For all x such that $1-(\alpha+1)^{-1/2} < x < 1$ $f(x)$ is strictly increasing function and (17) becomes to the form

$$\begin{gathered} \beta_0 + \beta_0 \cdot f(\beta_0) + \beta_0 \cdot f(\beta_0) \cdot f(f(\beta_0)) + \\ + ... + \beta_0 \cdot f(\beta_0) \cdot ... \cdot f(...f(f(\beta_0))...) + ... + \frac{\beta_0}{1-\beta_0} = \alpha. \end{gathered} \tag{18}$$

Because $f(\beta_0)$ is monotonly increasing function of β_0 and $\frac{\beta_0}{1-\beta_0}$ is also monotonly increasing expression of β_0 $(0 \leq \beta_0 \leq 1)$ we say that the left hand side of (18) is monotonly increasing expression of β_0. Therefore, for the constant right hand side of (18) we may have only one β_0 satisfying (18), which completes the proof of the Theorem.

References

[1] Ahsanullah, M. (1995) *Record Statistics.*Nova Science Publishers, Inc., Commack, NY.

[2] Ahsanullah, M. and Holland, B. (1984) Record values and the geometric distribution. *Statistische Helfe*, 25, 319-327.

[3] Aliev, F.A. (1998) Characterization of distributions through weak records. *Journal of Applied Statistical Science*, in appear.

[4] Kirmani, S.N.U.A. and Beg, M.I. (1984) On characterization of distributions by expected records. *Sankhya.*, 49, A, 463-465.

[5] Korwar, R.M. (1984) On charecterizing distributions for which the second record value has a linear regtression on the first. *Sankhya.*, 46, B, 108-109.

[6] Nagaraja, H.N. (1988) Record values and related statistics, A review. *Comm. Statist. Theory Methods*, Ser. A, 17, 2223-2238.

[7] Nevzorov, V.B. (1987) Records. *Theory Probab. Appl.*,V.32, 201-228.

[8] Stepanov, A.V. (1994) A characterization theorem for weak records. *Theory Probab. Appl.*,V.39, 762-764.

[9] Vervaat, W. (1973) Limit theorems for records from discrete distributions. *Stoch. Proc. Appl.*, 317-334.

ÖZET

$X_1, X_2, X_3, \ldots$ bağımsız 0,1,2,3,... değerlerini alan aynı F dağılımına sahip rasgele değişkenler olsunlar öyleki herhangi bir n için (n=0,1,2,3,...) F(n)<1 ve $E(X_1) < \infty$. $X_{L(n)}$ n. inci zayıf rekor değerini göstersin. Bu çalışmada X_1 rasgele degişkeninin geometrik dağılıma sahip olması için gerek ve yeter koşul bütün $i \geq 0$ ve bazı k>0, $\alpha > 0$ için $E\left(X_{L(n+2)} - X_{L(n)} \middle| X_{L(n)} = i\right) = \alpha$

İstatistik, Journal of the Turkish Statistical Association
1998, Vol.1, No. 2, pp. 67 - 74

CONSTRUCTION OF CONFIDENCE LIMITS FOR DEPENDENT SAMPLE VALUES

Ismihan G. Bairamov
Ankara University, Faculty of Science, Department of Statistics
06100, Tandogan, Ankara, Turkey

Dmitry A. Kljushin, Yuriy I. Petunin
Kiev Shevchenko University, Department of Cybernetics
252017, Vladimirskaya, 64, Kiev, Ukraine

Abstract

The problem of constructing confidence limit for dependent sample values is investigated by the method of computer simulation. This problem is considered for the random sequence $z_1, z_2, \ldots, z_n, \ldots$ constructed from the sequence of independent identical distrubuted random variables $x_1, x_2, \ldots, x_n, \ldots$ with the continious distribution function F by the method of moving summation:

$$z_1 = x_1 + x_2 + \ldots + x_k, \quad z_2 = x_2 + x_3 + \ldots + x_{k+1}, \quad \ldots, \quad z_n = x_n + x_{n+1} + \ldots + x_{k+n-1}, \quad \ldots .$$

Key Words: Invariant confidence interval, dependent random variables, moving average, computer simulation.

1. Introduction

In the contemporary mathematical statistics the problem of construction of confidence limits for the bulk of distribution of general population and for the parameters has outstanding place. The solution of this problem is necessary for creation of the statistical tests of pattern recognition, test of hypothesis and also for the test the precision of parameter estimations. Unfortunately, the analytical solution for the problem of the constructing the confidence limits and investigation its properties we can realize only in the case of independent sample values, as far as for dependent ones

there exist significant important and some times insuperable difficulties. In this work we shall investigate this problem by the methods of computer simulation.

2. Investigation of confidence limits for the bulk of distribution of general population constructed with the help of the order statistics in the case of dependent sample values

Let $x_1, x_2, \ldots, x_n$ be a sample obtained from the general population G with an unknown distribution function $F(u)$ with the help of the simple sampling and $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be a variational series corresponding to this sample. As well-known Madreimov and Petunin (1982), with the help of the order statistics $x_{(i)}, x_{(j)}$, $i < j$ we can construct the confidence interval $\left(x_{(i)}, x_{(j)}\right)$ with confidence level $\alpha = \dfrac{j-i}{n+1}$:

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

where x is the sample value from the general population G which does not depend on the sample $x_1, x_2, \ldots, x_n$. Bairamov and Petunin (1990) were shown that under some natural conditions any confidence interval

$$\left(a\left(x_1, x_2, \ldots, x_n\right), b\left(x_1, x_2, \ldots, x_n\right)\right),\ a\left(x_1, x_2, \ldots, x_n\right) < b\left(x_1, x_2, \ldots, x_n\right)$$
$$\forall\, x_1, x_2, \ldots, x_n$$

for the bulk of the distribution of the general population G with significance level which does not depend on the distribution function $F(u)$ (so-called confidence invariant interval) will be generated by the order statistics so that $a\left(x_1, x_2, \ldots, x_n\right) = x_{(t)}$, $b\left(x_1, x_2, \ldots, x_n\right) = x_{(j)}$.

Now assume that our sample values represent a sequence of dependent random values $x_1, x_2, \ldots, x_n, \ldots$ (for example, stationary time series or stationary Markov sequence) with the joint distribution function $F\left(u_1, u_2, \ldots, u_n\right)$, whose marginal distributions coincide with $F(u)$. In a natural way the following problem arises: will the formula (1) be correct or is it necessary to revise its right- hand side? We consider this problem for the random sequence $z_1, z_2, \ldots, z_n, \ldots$ constructed from the sequence of independent identical distributed random variables $x_1, x_2, \ldots, x_n, \ldots$ with the continuous distribution function $F(u)$ by the method of moving summation:

$$z_1 = x_1 + x_2 + \ldots + x_k, \quad z_2 = x_2 + x_3 + \ldots + x_{k+1}, \quad \ldots, \quad z_n = x_n + x_{n+1} + \ldots + x_{k+n-1}, \quad \ldots$$

CONFIDENCE LIMITS FOR DEPENDENT SAMPLE

The random sequence $\{z_n\}$ is stationary in a broad sense, as far as the mathematical expectation is

$$E(z_n) = (k-1)\,E(x_n) = (k-1)a,$$

where $a = E(x_n)$, and its correlation function is of the following form for $t > 0$:

$$K(z_n, z_{n+t}) = E((z_n - E(z_n))(z_{n+t} - E(z_{n+t}))) = K(t) = \begin{cases} k-t, & t \le k, \\ 0, & t > k. \end{cases}$$

Note, that correlation coefficient is $r(z_n, z_{n+t}) = \dfrac{K(z_n, z_{n+t})}{\sigma(z_n)\sigma(z_{n+t})} = \dfrac{k-t}{k}$, if $t \le k$; and $r(z_n, z_{n+t}) = 0$, if $t > k$, that is why for large k the statistical dependence between the random values z_n and z_{n+t} may be strongly desired. For the computer simulation and computation of frequency of hitting of the values z_l, $l > n$ ($n = 30$, $l = 31, ..., 60$) into the confidence interval $\left(z_{(i)}, z_{(j)}\right)$ as a random sequence x_n we select the sequence of independent random variables from uniform distribution over the interval $[0, 1]$, obtained with the help of generator of pseudo random real values RANDOM of MS Fortran 5.0. The lower and upper approximate confidence limits for the probability p_{ij} are constructed with the help of the frequency h_{ij} at 5% significance level according to the formulae Van der Waerden (1957)

$$p_{ij}^{(1)} = \frac{h_{ij}m + g^2/2 - g\sqrt{h_{ij}(1-h_{ij})\,m + g^2/4}}{m + g^2},$$
$$p_{ij}^{(2)} = \frac{h_{ij}m + g^2/2 + g\sqrt{h_{ij}(1-h_{ij})\,m + g^2/4}}{m + g^2},$$

where $g = 3$ in accordance with the 3σ rule, m=29; then the corresponding significance interval is $I = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$.

The results show that for small k, the number of terms contained the sum z_n ($n = 1, 2, 3, 4$) is slightly different from the frequency h_{ij} and probability p_{ij}.

In the Table 1 it is shown that as number k is increasing this difference become more and more significant and under k =10 the number of falling of p_{ij} outside the correspondent confidence interval exceeds 50%. This indicates on the bad correspondence of the formula (1) with experimental data.

3. The 3s-rule for depended observations

Classical 3σ-rule states that for random variables, which may occur in the practical statistical calculation, the confidence interval

$$(E(x) - 3\sigma(x),\ E(x) + 3\sigma(x))$$

contains more than 95 % of values of the general population generated by the random variable x. Now this rule is strongly justificated for the general population G with unimodal distribution $F(u)$ (Vysochanskij and Petunin (1980), Pukelsheim (1994), Sellke (1996), Dharmadhikari (1988)). As far as the parameters $E(x)$ and $\sigma(x)$, as a rule, are unknown, then this parameters are replaced by its estimations constructed on the basis of sample values $x_1, x_2, \ldots, x_n$

$$E(x) \approx \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \ \sigma^2(x) \approx s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 .$$

According to practical recommendations stated on the monograph of Cramer (1975), the estimation $\overline{x}$ practically coincides with $E(x)$, if $n \geq 30$, and $s(x)$ coincides with $\sigma(x)$, if $n \geq 150 \div 200$, in the case when samples values $x_1, x_2, \ldots, x_n$ are given as a result of simple sampling. If we replace $E(x)$ by $\overline{x}$, and $\sigma(x)$ - by $s(x)$, then the interval $I_{3s} = (\overline{x} - 3s(x), \overline{x} + 3s(x))$ is called a confidence interval for the bulk of the general population G constructed on the $3s$.

Consider the problem of estimation of significance level for this confidence interval I_{3s}. We shall use the method of computer simulation for this aim. Let $x_1, x_2, \ldots, x_n, \ldots$ be a sequence of independent identically distributed random variables with uniform distribution on the segment $[0, 1]$. As early, we generate the sequence of random variables $z_1, z_2, \ldots, z_n, \ldots$ by the method of moving summation. Let G_z denote general population corresponding to this sequence. To estimate the parameters $E(z)$ and $\sigma(z)$ we shall use the samples whose size is more than 30. After calculation of the confidence interval I_{3s} we consider a new sample whose size is equal 30 from general population G_z and calculate the percent of hitting of these new sample values in the confidence interval I_{3s}. The results of these calculations are represented in the Table 2. Analysis of the data mentioned in the Table 3 shows that the confidence level of this confidence interval I_{3s} constructed on the $3s$-rule practically, is about 100 %, although the sample values $z_1, z_2, \ldots, z_n, \ldots$ for the big values k ($k = 80 \div 100$) will be strongly dependent. Thus the $3s$-rule remains valid for the dependent observations.

4. Confidence limits for probability in MP-model

Consider the following model of identification of homogeneous of sample composed from two different samples. Let G_x and G_y be two general populations with unknown continuous distribution functions $F_x(u)$ and $F_y(u)$, respectively. We have two samples $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_m)$ sampled from G_x and G_y, respectively. It is necessary to test whether these unknown distribution function $F_x(u)$ and $F_y(u)$ coincide (hypothesis H_0) or differ from each other (compound concurrent hypothesis H_1) on the basis of these samples. In the first case $F_x(u) \equiv F_y(u)$ we deal with combined heterogeneous sample $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$; and if $F_x(u) \neq F_y(u)$ then combined sample is heterogeneous.

Construct on the sample the variational series $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, assuming $x_{(0)} = -\infty$, $x_{(n+1)} = \infty$, and consider the random interval $I_{i,q} = \left(x^{(i)}, x^{(i+q}\right)$, where i, q are the fixed numbers $(0 \leq i \leq n;\ 1 \leq q \leq n - i + 1)$. Consider in connection with the interval $I_{i,q}$ the following model of trials: at kth step $(k = 1, 2, \ldots, m)$ we test whether the sample value y_k belongs to the interval $I_{i,q}$ or not. Thus we have a the group of events $A_k = \{y_k \in I_{i,q}\}, k = 1, 2, \ldots, m$, each of which can occur with some probability $p_k = p(A_k), k = 1, 2, \ldots, m$. It should been noted, that even though in the classical Bernoulli trials are independent, the events A_k in considered model will be depended since in each of these events we deal with random interval $I_{i,q}$. Define the random variables α_k in the following way: $\alpha_k = 1$, if the event A_k has occurred, and $\alpha_k = 0$ otherwise; as far as the events A_k are depended then the random variables α_k also will be dependent.

Introduce a random variable Θ which is equal to the number of events A_k occurred in the series of m trials; then $\Theta = \sum_{k=1}^{m} \alpha_k$. If the hypothesis H_0 is valid then the probabilities in the formula (1) are equal to

$$p(A|H_0) = \frac{q}{n+1} = p_q. \tag{2}$$

This model is known as MP- model (N.Johnson and S.Kotz , 1992). In papers Matveychuk and Petunin (1990, 1991) the distribution of probabilities of random variable Θ was obtained as follows:

$$p(\Theta = l|H_0) = \frac{C_{l+q-1}^{l} C_{m+n-l-q}^{m-l}}{C_{m+n}^{m}},\ \ l = 0, 1, \ldots, m, \tag{3}$$

where m, n are sizes of the samples X and Y ; q is a fixed number which is the number of order statistics in the interval $I_{i,q}$; C_r^s is the number of combinations from r by s.

Let h be a frequency of appearance of events A_k under m trials, then $h = \frac{\Theta}{m}$.

Consider the problem of validity of the 3σ-rule for random variable h. It was found that this rule is valid if the hypothesis H_0 is valid, and samples X and Y have equal size, moreover, we shall show that even for significantly more narrow confidence interval $I_{2\sigma}$, constructed on the 2σ-rule, has the significance level not exceeding 6 %. Indeed, in the case when hypothesis H_0 is valid then on the basis of the results of the work [10, 11]

$$E(h|H_0) = p_q,$$

$$\sigma^2(h|H_0) = \frac{p_q(1-p_q)}{m}\frac{m+n+1}{n+2} = \frac{p_q(1-p_q)}{m}\frac{2n+1}{n+2} \approx 2\frac{p_q(1-p_q)}{m}$$

By using these formulae we can calculate the confidence level of the interval $I_{2\sigma}$:

$$p(h \in I_{2\sigma}) = \sum_{\left|\frac{l}{m} - E(h)\right| \le 2\sigma(h)} p\left(h = \frac{l}{m}|H_0\right).$$

Note that with the help of the confidence interval $I_{2\sigma}$ for the bulk of the distribution of the random variable h we can construct the confidence interval for the probability p_q. Really,

$$\begin{aligned} p(h \in I_{2\sigma}) &= p(|h - E(h)| \le 2\sigma(h)) = p(|h - p_q| \le 2\sigma(h)) \\ &= p(p_q \in (h - 2\sigma(h), h + 2\sigma(h))). \end{aligned}$$

The results of calculations of the confidence level of the interval $I_{2\sigma}$ which is equal to the confidence level of the interval $(h - 2\sigma(h), h + 2\sigma(h))$ are represented in Table 3 for $n = m$.

If we have a random confidence interval $I(\Theta) = (h - 2\sigma(h), h + 2\sigma(h))$ containing the probability p_q with the significance level 2β:

$$p(p_q \in I(\Theta)|H_0) = 1 - 2\beta,$$

then we can construct the following criteria to test the hypothesis H_0 against alternative H_1:

1) we construct the variational series on the sample X and take the random interval $I_{i,q}$ for fixed i and q ;

2) we define the statistics Θ which is equal to the number of the elements of the sample Y belonging to the interval $I_{i,q}$;

3) on the statistics Θ with the predefined significance level 2β we construct confidence interval $I(\Theta)$;

4) if $I(\Theta)$ cover p_q then we accept the hypothesis H_0, otherwise we accept the hypothesis H_1.

Summarizing the analysis of obtained results we may state that the presence of the dependence between the sample values not always requires to introduce significant corrections in calculations of the significance level of the confidence; in many cases these significance levels remain practically constant.

References

[1] 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.

[2] Bairamov, I. and Petunin, Yu.I. (1990) Structure of the invariant confidence intervals containing the underlying main distributed mass. Theor. Probability and its Applications, 35, 1, P.25-36.

[3] Van der Waerden, B.L. (1957) Mathematische Statistic, Springer-Verlag, Berlin.

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

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

[6] Sellke, Th. (1996) Generalized Gauss-Chebyshev inequalities for unimodal random variables. Metrika, 43, P.107-121.

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

[8] Cramer, H. (1975) Mathematical Methods of Statistics. Mir, Moskow (Russian tranl. from English), 648 pp.

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

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

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

k	Number of hitting	Number of falling out
1	28	1
2	28	1
3	26	3
4	25	4
5	17	12
6	21	8
7	19	10
8	16	13
9	17	12
10	12	17

Table 1: Frequency of hitting and fallings out confidence interval

k	Hitting frequency	$\overline{z}$	$s^2(\overline{z})$
1	1.00000	1.04048	0.16601
2	1.00000	1.49308	0.22020
3	0.99700	2.03360	0.29452
4	1.00000	2.49091	0.43437
5	0.99600	3.01639	0.48761
6	0.99800	3.55370	0.51377
7	0.99800	4.12859	0.53500
8	1.00000	4.45868	0.66220
9	0.99500	5.01084	0.76228
10	1.00000	5.49083	0.72886
20	0.99800	10.80193	1.87548
30	0.99900	15.59051	2.86506
40	1.00000	20.84352	4.96147
50	0.99700	25.38411	2.96953
60	1.00000	30.37748	4.52146
70	0.94900	36.82066	5.67559
80	0.99800	39.80943	5.33090
90	0.96900	46.38148	6.54674
100	1.00000	50.86939	7.78361

Table 2: Frequencies of hitting in confidence interval for depended observations

k	n=10	n=20	n=30
1	0.043	0.053	0.056
2	0.029	0.046	0.051
3	0.035	0.032	0.040
4	0.035	0.041	0.052
5	0.029	0.048	0.036
6	0.039	0.027	0.042
7	0.033	0.028	0.027
8	0.035	0.036	0.040
9	0.029	0.042	0.037
10		0.044	0.044
11		0.044	0.050
12		0.042	0.054
13		0.038	0.042
14		0.032	0.045
15		0.037	0.031
16		0.048	0.031
17		0.041	0.045
18		0.032	0.042
19		0.046	0.054
20			0.050
21			0.044
22			0.038
23			0.043
24			0.032
25			0.042
26			0.036
27			0.052
28			0.040
29			0.051

Table 3: Confidence level in MP-scheme

ÖZET

Bağımlı örneklemler için güven aralıklarının oluştutulması problemi bir simülasyon metodu ile incelenmiştir. Bağımsız aynı sürekli F dağılımından gelen $X_1, X_2, X_3 \ldots$ rasgele değişkenlerinden elde edilen hareketli toplamlar $Z_1, Z_2, Z_3 \ldots$ rasgele değişkenleri gözönüne alınmıştır. Burada

$$Z_1 = X_1 + X_2 + \ldots + X_k, \quad Z_2 = X_2 + X_3 + \ldots + X_{k+1}, \ldots, \quad Z_n = X_n + X_{n+1} + \ldots + X_{n+k-1} \ldots$$