ANNEX 1

The methods of discovering discrepancies in random processes and their application to the analysis of historical texts

By *B. E. Brodsky* and *B. S. Darkhovskiy**

The present work discusses the possibility of using the methods of discovering the alteration points in the probability characteristics of random processes for the analysis of historical texts, presenting the primary concepts of the non-parametric approach to respective statistical problems as developed by the authors.

The methods of analyzing narrative texts developed by A. T. Fomenko made it feasible to provide quantitative answers to a number of questions, which are of interest for historians. In particular, it proved possible to make a mathematically correct formulation of the following historical problem, which is rather remarkable. It is known that a wide range of the ancient historical sources (all sorts of chronicles and manuscripts) consists of individual fragments, or segments, of varying origins. For instance, individual fragments may have been written by different authors and/or in different geographical regions, and can therefore differ from each other substantially in their character, language, style of narration, amount of details, emotional overtones etc. It is possible that these fragments were united into a single book by a later chronologist. After that, the origins of the fragment texts may have become forgotten, and their collection regarded as a single chronicle. Over the course of time, the initial differences between individual fragments would gradually become unnoticeable due to repeated copying, "editing" and so on.

A natural and important (to historians first and foremost) question arises in this respect, namely, whether it is possible to discover these initial ingredients of a single voluminous text via the statistical analysis of their various frequency characteristics, or cut the large text up into the fragments that it consisted of originally.

A. T. Fomenko and A. N. Shiryaev put forth the hypothesis that each one of these fragments is uniform stochastically – in other words, it represents a stationary temporal sequence of some sort (being transformed into a sequence of numbers, which we consider accomplished a priori; as to the methodology of such transformation, see Annex 2), with different fragments corresponding to different stationary sequences with different stochastic characteristics.

This hypothesis proved useful in the analysis of actual historical texts. Corresponding results are contained in Annex 2. Herein we shall provide a more detailed account of the ideology of solving the arising class of statistical problems.

This field of mathematical statistics can be referred to as the methods of detecting the alterations

* Professor B. S. Darkhovskiy is a mathematician and a Doctor of Physics and Mathematics, Moscow; Professor B. E. Brodsky is a Mathematician and a Candidate of Physics and Mathematics. Both of them are experts in the field of probability theory and mathematical statistics.

in the stochastic properties of random processes and fields. We are referring to the following two classes of problems:

Primo. Let us assume that we have a sampling (realization) of a random process (field) under study. Any sort of statistical processing of this sample aimed at the creation of a model, the evaluation of its parameters etc must be based upon the hypothesis (which is the key element of mathematical statistics) that the evaluated phenomenon did not change over the course of data collection. Therefore, a verification of such uniformity happens to be a necessary preliminary stage of any statistical processing. Thus, the question is posed as follows: is the sampling in question uniform statistically, inasmuch as the immutability of its stochastic characteristics is concerned? Should this be answered in the positive, one may proceed with the regular kind of statistical analysis in accordance with the goals of the researcher. If the answer is negative, one is confronted with a problem of discovering the alteration points in stochastic characteristics and separating the original sampling into several fragments that would be uniform statistically.

The class of problems described above became known as that of the retrospective (a posteriori) discrepancy problems. The term "discrepancy" here is a brief reference to any change in stochastic characteristics.

Secundo. The second class of problems can be related as follows. Let us assume that the information pertinent to a random process (its measurement) reaches us in a temporal sequence of some sort, and that at some point that remains unknown to us, one of the stochastic characteristics changes in some way (one of the distribution functions in the most general case). What we need to know is how to discover this change as soon as possible after it had taken place (one understands that it is impossible to do this beforehand, or "predict the future"), without making the false alarms too frequent. The frequency of such signals can be limited by a given value. This problem received the name of rapid discovery of the "discrepancy".

The first works on the subject were published as early as the 1930's. See the reference to Shewhart's work on the rapid discovery problem in [1111]. However, no rigid theory was constructed then. In the 1950's, Page's works came out ([1325] and [1326]), containing the formulated methodology of rapid and retrospective discovery of the "discrepancy". This method became known as the method of cumulative sums later on, and is based on the consecutive calculation of plausibility; it proved convenient from the point of view of organizing the calculations, and effective practically. Around the same time, A. N. Kolmogorov gave a rigid formulation of the problem of rapid discovery of the "discrepancy" for Winer's process, presenting it as an extreme stochastic problem. This problem was solved by A. N. Shiryaev, who had estimated the optimal discovery method for the situation in question. The results of A. N. Shiryaev's research in this field are contained in [976].

General interest in the "discrepancy" problematics began to grow in the mid-1960's due to the growing demand for their application. The main efforts of the researchers were aimed at the development of methods that would require a minimum of a priori data. The matter is that optimal methods, as well as the ones approximating those, are based on the exact knowledge of distribution functions as they manifest before and after the "discrepancy" point, should the latter be of a random nature. Such information is difficult to obtain, and this concerns a large number of interesting practical applications. This resulted in the development of certain minimax methods, which render the information pertaining to the distribution function of the "discrepancy" point unnecessary, as well as non-parametric methods, which require no information concerning the random sequence distribution. Voluminous overviews of the works on this problem that were published over the last 15-20 years can be found in [392], [1406] and [1230].

The works of the authors of the present text were among the first research documents on the topic of non-parametric solution methods applicable to solving "discrepancy" problems. From the very beginning we have strived to synthesize such methods, simple enough to be used practically for problem solution. We deem the non-parametric methods that use no a priori distribution data to be most fitting for the purpose at hand.

Our research in the field of mathematical statistics referred to herein is summarized in [1051]. We shall presently relate the main concepts of our approach to the retrospective methods of finding the "discrepancy", since these methods were used for the analysis of historical texts.

The two main ideas behind our methodology are as follows. The first can be formulated in the following manner: the discovery of a change in every distribution function or some other stochastic characteristic (with any degree of precision) can be rendered to the discovery of an alteration in the mathematical expectation of some new random sequence derived from the original. Let us illustrate with the following example. Assuming that the random sequence under analysis is

$$X = \{x_t\}_{t=1}^{N},$$

which is a collation of the two rigidly stationary random sequences

$$X_1 = \{x_t\}_{t=1}^{n^*}, \quad X_2 = \{x_t\}_{t=n^*+1}^{N},$$

$n^* = [\theta N]$, $0 < \theta < 1$, and one has to evaluate n^* as the collation point.

Let us assume that X_1 and X_2 are known to differ in one of their two-dimensional distribution functions – namely, that the function $P\{x_t \le u_0, x_{t+2} \le u_1\} = = F(u_0, u_1)$ until the moment $t_1^* = n^* - 2$ is equal to $F_1(\cdot)$, and to $F_2(\cdot)$ in case of $t \ge t_2^* = n^* + 1 - F_2(\cdot)$, where $\|F_1(\cdot) - F_2(\cdot)\| \ge \varepsilon > 0$, and $\|\cdot\|$ stands for the regular sup-norm. It is known well enough that the distribution function of the finite random vector can be evenly approximated with any degree of precision by the random vector distribution function with a finite amount of values. This leads us to the premise that the separation of the plane **R** into a large enough number of non-intersecting areas A_j, $j = 1, \dots, r$, allows the vector (x_t, x_{t+2}) to be approximated in terms of distribution by a vector with a finite amount of values. Therefore, if we are to introduce new random sequences

$$V_t^{ij} = I(x_t \in A_i, x_{t+2} \in A_j), \quad 1 \le i \le r, \quad 1 \le j \le r$$

($I(A)$ being the indicator of set A), at least one of these sequences will demonstrate an alteration of mathematical expectation. Therefore, should there be an algorithm to allow us the discovery of changes in mathematical expectation, the very same algorithm could also discover them in the function of distribution. Alterations in arbitrary stochastic characteristics can be discovered in the same manner. For instance, should there be a correlation function change in the sequence, our study of the new sequences $V_t(\tau) = = x_t x_{t+\tau}$, $\tau = 0, 1, 2, \dots$ shall render the problem to that of discovering a change of mathematical expectation in one of the sequences $V_t(\tau)$.

This circumstance makes it sufficient to develop a single basic algorithm that would allow us to discover changes in mathematical expectation instead of creating infinite algorithms for the discovery of changes in various stochastic characteristics.

The second idea that our approach is based upon is discovering the "discrepancy" moments with the use of statistical families such as

$$Y_N(n) = \left[\left(1 - \frac{n}{N}\right)\right] \delta \left[\frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{N-n}\sum_{k=n+1}^{N} x_k\right] \quad (1)$$

where $0 \le \delta \le 1$, $1 \le n \le N - 1$, $X = \{x_k\}_{k=1}^{N}$ is the realization under study, as well as certain derivatives of these statistic.

Family (1) is a generalized variant of the Kolmogorov-Smirnov statistics used to verify coinciding or varying distribution functions in two samples (n being a fixed value). One can demonstrate that the statistics of the (1) kind are asymptotically minimax in their order (with $N \to \infty$, and the correlations between the "collated" realizations remaining the same); that is to say, they minimize the maximal probability of an error in evaluating the "discrepancy" moment.

Said concepts (see [1051] for more details) were embodied in a software suite called VERDIA for PC-compatible machines, which allows for the interactive discovery of "discrepancies" in arbitrary random sequences. We have used the VERDIA suite for analyzing several actual historical texts; the results of this analysis are published in Annex 2 to the present book.