Statistical series for the ordered array approximation of random variable

V G Polosin
Medical Cybernetics and Computer Science Department, Penza State University, 40, Krasnay Str., Penza, 440026, Russia

E-mail: polosin-vitalij@yandex.ru

Abstract. Many papers contain the study of observation results by using statistical moments and signs of the shape of distributions of random variables. Most of the directions of the application of moments are based on the relationship of moments with the parameters of the shape and scale of distributions. Expansion of the boundaries for the distribution parameters area is possible by identifying new qualitative properties of these characteristics. For this reason, it is relevant to study the possibility of using the central moments of the n-th order when constructing a statistical series. The paper demonstrates the possibility of using distribution moments to construct a series of approximations of a sorted data sequence. In particular, a method is developed for the expansion of a monotone varying function in a power series, that is based on obtaining the coefficients of the series by using the central moments of the n-th order which calculated from a random sample of data. There are also examples of sorted series approximation for typical known symmetric and asymmetric distributions. There are also examples of sorted series approximation for typical known symmetric and asymmetric distributions. The article illustrates the use of a statistical series for the study of arrays of random variables by the example of approximating a sorted sequence of one cycle of an electrocardiosignal.

1. Introduction
In various fields of research, statistical methods are often used for the analysis of physical processes in complex systems. The exponential growth in computing power over the past two decades has revolutionized statistical analysis and led to rapid development and significant progress in this new field. For modeling complex systems using probabilistic distributions, the construction of a probabilistic model of physical processes is one of the most difficult problems in the analysis of observed data. When analyzing physical processes, information methods based on the assessment of the moments of statistical distributions are often used. [1, 2, 3]

The possibility of applying the methods of moments is limited by the problem of assessing the correspondence of the found moments of the distribution to the real sample of values. The point is that the moment estimates are sensitive to random outliers caused by powerful external influences. In this case, the distribution forms obtained for the distributions of random variables based on high-order moments do not allow obtaining a fair distribution form for the studied sample of results.

When analyzing random time sequences, approximations are constructed, the purpose of which is to identify the deterministic properties and states of complex systems. It is possible the use of statistical series makes to approximate the sorted sequence in the vicinity of the estimation point of the center of the sample array, if these series are built on the basis of distribution moments. We can construct an
approximation if the distribution function is infinitely differentiable in the vicinity of the point. The differentiability of distributions imposes restrictions on the applying area of the statistical power series that it is considered in paper.

The purpose of this paper is to study the possibility of constructing an approximation of a sorted array of values using a statistical series which the coefficients are determined using the central moments of the n-th order.

2. Statistical series for the approximation of sorted time sequences

A data sequence is a convenient tool for studying the characteristics of the distribution of the output parameter of a complex system if it is ordered in ascending or descending order. In order to find the form of data distribution in a complex system, high-order moments are often used. Expression for the connection of distribution moments with distribution parameters we can be found in various reference books on probability [2, 4, 5]. The formula for calculating the \( \mu_s(Y) \) central moment of s order from the \( y_i \) values of the \( Y \) array it is as:

\[
\mu_s(Y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - m)^s.
\]  

(1)

Where \( m \) is the arithmetic mean of the \( Y \) array, \( n \) is the number of values in the array.

For the approximation of the data array that it is ordered in ascending or descending order, the author in [6] built a statistical series by analyzing the properties of the distribution moments. The statistical series was used to approximate the data that was obtained over time of \( \tau_c \). The statistical series of moments is given by

\[
y_i = Me(Y) + \sigma(Y) \frac{\tau_{u/2} - \tau_i}{\Delta \tau \cdot 1!} \left( 1 + \sum_{s=3}^{4} \mu_s(Y) \frac{\left( \tau_{u/2} - \tau_i \right)^{s-2}}{\Delta \tau^{s-2} (s-1)!} \right).
\]

(2)

Where \( Me(Y) \) and \( \sigma(Y) \) are median and standard deviation for the \( Y \) array, \( \tau_{u/2} \) is time interval that it is equal to half period of the cycle registration of \( \tau_c \), \( \Delta \tau \) is variable distribution parameter:

\[
\Delta \tau = k \frac{\tau_c}{2\sqrt{3}}
\]

Where \( k \) is the coefficient of variation. This allows you to change the slope of the approximation at the point of the median position, if array values is ordered in descending or ascending order.

In statistics, such distribution shape coefficients as \( Sk(Y) \) skewness and \( Ex(Y) \) kurtosis to select a model are often used [5, 7]. The standard deviation \( \sigma(Y) \) characterizes the scale of the distribution uncertainty interval. The vector of distribution parameters is given using the central moments (1) as

\[
\begin{bmatrix} \sigma(Y) & Sk(Y) & Ex(Y) \end{bmatrix}^T = \begin{bmatrix} \sqrt{\mu_2(Y)} & \frac{\mu_4(Y)}{(\mu_2(Y))^2} & \frac{\mu_6(Y)}{(\mu_2(Y))^3} \end{bmatrix}^T.
\]

(3)

If approximation (2) is constructed with restriction of the expansion to the central moment of the 4-th order, then the truncated statistical series of values of the array \([Y]\) of the output parameter of the stochastic system is given as [6, 8]

\[
y_i = Me(Y) + \sigma(Y) \frac{\tau_{u/2} - \tau_i}{\Delta \tau 1!} + Sk(Y) \frac{\left( \tau_{u/2} - \tau_i \right)^2}{\Delta \tau^2 2!} + Ex(Y) \frac{\left( \tau_{u/2} - \tau_i \right)^3}{\Delta \tau^3 3!} \right).
\]

(4)

Statistical series of (2) and (4) are written for decreasing ordering of data array values.

A typical example of approximating the sorted sequence of values of an array of a symmetrically distributed random variable is shown in figure 1, a. The following notation is used there. Number 1 is a
sorted sequence of an array values of a normally distributed random variable. Number 2 is approximation if it is used the (2) statistical series of moments. It is show dotted. Number 3 is approximation if it is used a (4) truncated statistical series. Numbers 4 and 5 are show reduced errors. It is obtained at approximation using the statistical series of moments and the truncated statistical series respectively. Figure 1, b illustrates the position of the histogram of the distribution of a random variable that is observed in the time period $\tau$.

![Figure 1. Approximation of a sorted sequence of values for an array of normally distributed of random variable.](image)

As follows from consideration of figure 1, a, if the sorted sequence of values of the array $Y$ is limited to quantiles of 6% and 94%, then the reduced error of approximation of the truncated statistical series does not exceed 2%. In figure 1 the quantiles of 6% and 94% are denoted as $\tau_6$ and $\tau_94$. For the statistical series of moments (2), the region of approximation is limited to quantiles of 3% and 97%. The approximation is constructed for a normally distributed random variable. The coefficient of variation is equal 3.1.

### 3. Statistical Series for Approximating Sorted Arrays of Distributions

It is often the analyzed data array does not depend on time. In this case, the $\tau$ parameter characterizes the position of the value with number $i$ for the data array that it is sorted in ascending order. This explains the peculiarity of the sorted series if we use the normalized value of the parameter $\tau'$ such that the limiting value of $\tau_i'$ is 1. In this case, the (2) statistical series of moments and the (4) truncated statistical series are given as

\[
y_i = \text{Me}(Y) + \sigma(Y) \cdot v \left( \tau_i' - 0.5 \right) \left[ 1 + \sum_{r=3}^{s} \frac{\mu_r(Y)}{(\mu_2(Y))^{r/2}} \cdot \frac{v^{r-2}(\tau_i' - 0.5)^{r-2}}{(s-1)!} \right],
\]  

(5)

\[
y_i = \text{Me}(Y) + \sigma(Y) \left( \tau_i' - 0.5 \right) v + \text{Sk}(Y) \left( \tau_i' - 0.5 \right)^2 v^2 + \text{Ex}(Y) \left( \tau_i' - 0.5 \right)^3 v^3.
\]  

(6)
Where \( \nu \) is the coefficient of variation equal to \( 2\sqrt{3k^{-1}} \).

If the normalized parameter \( \tau/ \) is determined through the values of \( y \), then the dependence \( \tau'(y) \) is the probability of the distribution of the values of the random variable \( Y \), that it is given by

\[
\tau'(y) = F(y, Me, \alpha, Sk, Ex).
\]

(7)

Figure 2 shows an example of an approximation of beta distributions for various \( \alpha, \beta \) shape parameters. The beta distribution is given as

\[
f(y, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}I_{[0,1]}(x)}{B(\alpha, \beta)}.
\]

(8)

Where \( \alpha, \beta \) are shape parameters (\( \alpha \) and \( \beta > 1 \)); \( I_{[0,1]}(x) \) is an indicator function, which guarantees a nonzero probability in the range of values from 0 to 1; \( B(\alpha, \beta) \) is a beta function:

\[
B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.
\]

(9)

The beta distribution describes a curves family that are nonzero only on the interval \([0,1]\). In the figure, the following designations are given. Numbers 1, 2, 3 and 4 are sequences of values arrays of sorted in ascending order at various \( \alpha, \beta \) shape parameters. Numbers 5, 6, 7 and 8 are approximations of the sorted sequence if it using the (2) statistical series of moments. Numbers 9, 10, 11 and 12 are approximations of the sorted sequence if it using the (4) truncated statistical series.

![Figure 2. Approximations of beta distribution at unipolar density.](image)

As follows from figure 2, statistical series can be used to approximate distributions in the vicinity of the median of distributions at unipolar distribution density (\( \alpha > 1 \) and \( \beta > 1 \)). For the range of values that limited by the 0.1 and 0.9% quantiles, the approximation error did not exceed 2%.

Since the terms of the series are smooth functions, the approximations are limited to the domain of the distribution function where derivatives of higher orders exist. It follows from the examples given that the statistical parameters of the distributions make it possible to effectively find the coefficients of the expansion of random functions in power series. In this case, the quality estimates of the obtained approximations for sorted samples of values are applicable to solving the problem of constructing...
estimates of the adequacy of the found distribution parameters when using them to establish the shape of distributions of a sample of random ECG values.

4. The use of a statistical series for the analysis of electrocardiographic information.
Solving the problems of developing tools for breakthrough research in the field of physiology and fundamental medicine is based on new methods of signal research [9, 10]. The increase in the information content of the recorded signals is achieved through the use of various processing methods so spectral, correlation, fractal and other methods of automatic analysis of biological signals [11].

From the point of view of an outside observer, who is able to control only the beginning and end of the heartbeat cycles, the obtained electrocardiographic (ECG) measurements during one cycle represent a sample of random values, which contains information about the nature of the process that it is implicitly expressed in the form of a statistical distribution curve of the sample values. The application of statistical analysis methods to a sample of random ECG measurements obtained in one cycle of the heart, allows you to extract information about the features of the entire process and as its individual components.

A typical ECG signal is the number 1 in figure 3 that was recorded during one heartbeat. This is given as a solid curve. In figure 3 number 2 is a sorted sequence of values that was recorded during the cardio cycle. The curve with the number 3 illustrates the use of the (4) truncated statistical series to approximate the sorted sequence. This is given as a dotted curve. The approximation of a sorted series of values is considered as a regression of the behavior of a random ECG signal, which characterizes the variability of the signal during one cycle.

![Figure 3. Approximations for typical ECG signal of one period.](image)

For the array of ECG signal values in figure 3, the following characteristics were obtained. The \( \tau_c \) cardio cycle duration was 797 ms. The \( Me \) median and the \( \sigma_U \) rms spread are 66 mV and 218.8 mV, respectively. To construct a statistical sequence, distribution coefficients such as the \( As \) asymmetry and the \( Ex \) kurtosis were calculated, which are equal to 2.31 and 10.82, respectively. The coefficient of variation was 2.25. In this case, the \( \Delta \tau \) variable time interval of uncertainty was 634 ms.

A change in the signal activity in characteristic time intervals will lead to a redistribution of the interval durations, that will change both the shape and the statistical characteristics of the distributions array of ECG values of the cardiac cycle signal. In this case, information about both the composition of
the array of ECG values and its change is contained in its parameters, among which the $\mu_s(U)$ central moments of the $s$-th order for the $U$ array values of ECG signal should be distinguished. The assessment of the central moments allows one to judge the shape of the distribution histogram and as a consequence the state of the heart.

Investigating the forms of distributions of the values of the entire signal or its individual time segments, it is possible to solve the problem of establishing the shape and duration of the segments themselves. In particular, it is known that with a decrease in the heart rate, the T-P time interval of the relative rest of the heart activity increases. Since during normal operation of the heart, the time interval of electrical activity with one contraction does not exceed 0.63 s, then with a heart rate of about 65 beats per minute the T-P time interval will be 0.3 s. Therefore, 1/3 of all values of the arrays fall on a close to constant value of the TP voltage of the ECG interval. All these values located in one or two columns of the histogram, that it is causing a high level of estimation of the distribution kurtosis.

Reference
[1] Vadzinskiy R N 2001 Probability Distributions Handbook (Saint Petersburg: Publishing house Nauka) pp 295
[2] Cramer H 1946 Mathematical Methods of Statistics (Princeton: Princeton University Press)
[3] Grubbstrom R W and Tang O 2006 The moments and central moments of a compound distribution European Journal of Operational Research 170 106-19
[4] Gubarev V V 1992 Probabilistic Models: A Handbook (Novosibirsk: Novosibirsk Electrotechnical Institute Press) 2 pp 198
[5] Kobzar A I 2012 Applied Mathematical Statistics. For engineers and scientific workers (Moscow: FIZMATLIT) pp 816
[6] Polosin V G 2014 Way of the statistical analysis of electrocardiographic information Izvestia SFedU Technical science 10 120-6
[7] Cain M K, Zhang Z and Yuan K H 2017 Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation Behav. Res 49 1716-35
[8] Polosin V G 2017 Stochastic monitoring system heart of electrophysiological characteristics Journal of New Medical Technologies 3 47-53
[9] Zaichenco K V 2013 Elektronic and information tools for physiological and medical research Achievements of Modern Radioelectronics 2 107-11
[10] Zaichenco K V 2015 Information system for the signals readout and processing from the sensors of the living systems bioelectric activity Sensors & Systems 2 52-6
[11] Gulyaev Yu V and Zaichenco K V 2013 Ultra-high-resolution electrocardiography. Tasks. Problems. Perspectives Biomedical radio electronics 9 5-15