https://doi.org/10.33271/nvngu/2024-4/114

Yu. Golovko,
orcid.org/0000-0001-6081-8072,
O. Sdvyzhkova*,
orcid.org/0000-0001-6322-7526

Dnipro University of Technology, Dnipro, Ukraine
* Corresponding author e-mail: sdvyzhkova.o.o@nmu.one

CUMULATIVE TRIANGLE FOR VISUAL ANALYSIS OF EMPIRICAL DATA

Purpose. The development of a graphical object for visual analysis that allows for simultaneous evaluation of both general characteristics and details of the empirical data distribution.

Methodology. Justification of the feasibility and sequence of creating the cumulative triangle, as well as proving its properties, was carried out using geometric constructions, generalization, and lattice functions. The construction of the cumulative triangle was implemented in the "Matlab" software. Samples of random variables with known distribution laws were obtained using a pseudo-random number generator. Previously calculated dependencies of the spectral power density of seismic-acoustic noise-like signals were used as empirical data.

Findings. A folded cumulative function of the n-th order was introduced as a generalization of the known folded cumulative function. Using the folded cumulative functions, a geometric object that is the cumulative triangle, was designed to visualize the empirical distribution function. Lines dividing the triangle into flat curvilinear quadrilaterals are plotted on each triangle. It is shown that the face area can be used as a characteristic of the random variable concentration near the abscissa of the face upper node, and the difference in the areas of the face left and right parts provides for assessing the asymmetry of the distribution over the interval covering the face.

Originality. A new graphical object for visual analysis of empirical data distribution is proposed. It is shown how, relying on its appearance, conclusions can be drawn both regarding the characteristics of the entire sample and individual intervals of the distribution function.

Practical value. The cumulative triangle can be a useful addition to graphical visualization tools. Its use allows for simultaneous detailing and generalization of the properties of experimentally obtained data at different scale levels, which is particularly valuable when data have complicated and variable distributions.

Keywords: *visualization, empirical data, distribution function, folded cumulative function, power spectrum*

Introduction. Graphical representation of data is undeniably valuable in most cases. Visualization in an easily comprehensible form can enhance overall understanding of information, illustrate expected patterns, and reveal anomalies or incorrect prior assumptions. Utilizing graphical formats can occur at both the initial and final stages of data analysis and is always used when presenting research results and demonstrating the significance of conducted studies. This is particularly true when visualizing empirical distribution functions.

Most commonly, frequency histograms are used for this purpose. In the case of continuous values, the line connecting the midpoints of the upper bases of the histogram bars can be considered as an estimate of the distribution density. Such a procedure is simple to perform even with a large number of observations and is typically used to determine the closeness of an experimental distribution to one of the known generalized distributions and to compare various samples. Since histograms use the combination of data, their shape depends significantly on the width of the intervals (bins) into which the range of the obtained values is divided. Despite the fact that the method has been in use for over 200 years, the issue of choosing an interval remains a matter of controversy [1]. The agreement is that various interval widths must be used in computations and that the optimal interval width depends on the previously unknown distribution of the studied value [2]. Therefore, histograms remain subjective objects that may not reflect important details or, on the contrary, give an excessively detailed picture, where generalized properties are lost.

One of the reasons for the widespread use of histograms is the possibility of constructing them manually. The addition of computational tools allows one to directly find a smooth empirical function of the distribution density. To do this, the method of kernel density estimation (KDE-Kernel Density Estimation) is employed [3, 4]. Usage of the method involves choosing a kernel function (typical functions: uniform, triangular, quadratic, normal [5]) and bandwidth. The bandwidth affects the resulting density graph in a similar way to the width of the interval when constructing a histogram. A small band clutters the chart with small fluctuations and hides the main trends. A wide band does not make it possible to detect small features of the distribution. The kernel function also affects the type of dependence obtained, introducing various artifacts [5, 6]. Especially unexpected results can occur if there are long "tails" in the distribution [7]. Thus, the method of nuclear density estimation also significantly depends on the subjective choice of parameters.

Disadvantages of histograms and the method of nuclear density estimation are primarily caused by the necessity to group data (either explicitly or through the kernel function). The cumulative function does not have such a drawback [8]. The construction of its graph does not require the addition of any numerical parameters or assumptions. Each observation can be clearly plotted on a function graph, which simultaneously allows a visual estimate of the sample size. At the same time, relying on the graph of the cumulative function, the interpretation of the random variable features and the comparison of different data sets are complicated by the visual proximity of the curves for different distributions.

Generalized information of the cumulative function can be visualized using a box diagram ("box-and-whiskers plot") [9]. In the classic form, the box is bounded by the first and third quartiles and has a mark for the second quartile (median). "Whiskers" indicate the limits of a statistically significant part of the sample and in most cases are close to the I^{st} and 99^{th} percentiles.

The box diagram allows you to visually determine the median, assess symmetry, and establish the presence of outliers in the values of the studied data. The obvious disadvantages of the diagram are the lack of display of details and gaps in the data, as well as the impossibility to estimate the sample size from the diagram. In order to overcome these shortcomings, various modifications are proposed [10, 11], which complicate the diagram shape, but do not eliminate the generalization defect. Therefore, it is most appropriate to use a box plot when tracking or comparing data that have the same or similar type of distribution.

All the information present in the cumulative function is stored in the folded empirical cumulative function, the use of which has spread since the article by K. Monti (1995). The difference of the latter is only in the graphically displayed data, but this significantly simplifies their interpretation. From the graph, you can easily determine the median, assess the symmetry of the distribution, track outliers, and detect gaps in the sample. In [12] it is shown that the area under the graph of the folded cumulative function is equal to the average absolute deviation of the sample values from the median, that is, the area under the graph can be used to estimate the variance of a random variable. It is also important that all the data is reflected on the graph of the function. Therefore, it is not surprising that this method of visualization is the most common in medical research [13, 14], where a specific patient can stand behind each specimen of the sample.

As a disadvantage, it is necessary to point out the difficulty in comparing several distributions, especially when it is necessary to observe changes over time. In such situations, a box plot may be more useful. In case of an unknown type of distribution in advance, preference should be given to the folded cumulative function. Even though all the data are clearly present on the graph, it is easy to make a visual assessment only in relation to the entire distribution as a whole. Peculiarities in the distribution on some separate segments affect the appearance of the function, but in order to interpret them, it is necessary to carry out auxiliary calculations and, possibly, graphic constructions.

The research **purpose** is development of a graphical object that simultaneously displays both the general characteristics and the details of empirical data distribution. At the same time, a generalization of the folded cumulative function is used.

Theoretical basis. If $F(x)$ is a cumulative distribution function of a random variable X, then folded distribution function (mountain plot) is defined as

$$\widehat{F}(x)=\begin{cases}F(x), & F(x)\le\dfrac{1}{2}\\ 1-F(x), & F(x)>\dfrac{1}{2}\end{cases}. \qquad (1)$$

Graph $\widehat{F}(x)$ visualizes the distribution most clearly, displaying its symmetry and the median of the distribution.

Based on the same cumulative function, let us introduce a function

$$\widehat{F}_2(x)=\begin{cases}F(x), & F(x)\le\dfrac{1}{4}\\ \dfrac{1}{2}-F(x), & \dfrac{1}{4}<F(x)\le\dfrac{1}{2}\\ F(x)-\dfrac{1}{2}, & \dfrac{1}{2}<F(x)\le\dfrac{3}{4}\\ 1-F(x), & \dfrac{3}{4}<F(x)\le 1\end{cases}, \qquad (2)$$

which we will call the folded cumulative function of the second order (mountain graph of the second order). The graph of this function has two extremes (two peaks, two mountains). The abscissas of the extremes are the quantiles of the levels $0,\frac{1}{4},\frac{1}{2},\frac{3}{4},1$. Fig. 2 visualizes the deviation with respect to the symmetry of both the entire distribution and its two parts separately: the left part, where $F(x)\le\frac{1}{2}$, and the right part, where $F(x)>\frac{1}{2}$.

To identify the details of the distribution, we will introduce a folded cumulative function of the k^{th} order (mountain graph of the k^{th} order), which we will denote as $\widehat{F}_k(x)$.

Let us assume that $F(x)$ is a monotonic nondecreasing function for the interval $[a, b]$ and $F(a) = 0$, $F(b) = 1$.

This function could be considered "theoretical". Graphs of functions will be constructed in a rectangular coordinate system xOy, while putting the values of cumulative $F(x)$ and folded cumulative $\widehat{F}_k(x)$ distribution functions along the Oy axis.

Let us denote as x_i quantiles of the levels $\frac{i-1}{2^k}$, $i=\overline{1,2^k+1}$ cumulative function $F(x)$. Then there are equalities

$$F(x_i)=\frac{i-1}{2^k},\quad i=\overline{1,2^k+1}. \qquad (3)$$

Graph $\widehat{F}_k(x)$ is limited by points with coordinates $(a, 0)$ and $(b, 0)$; it has 2^{k-1} maximums ("mountain peaks") located at points

$$\left(x_i,\frac{1}{2^k}\right),\quad i=2,4,\ldots,2^k,$$

and $2^{k-1} + 1$ minimums ("intermountain troughs" together with the extreme points), which are located in points $(x_i, 0)$, $i = 1, 3, \ldots, 2^k + 1$ (odd indices are applied for minimums; even indices are applied for maximums).

For convenience, let us designate the abscissas of the maximums as x_i^+ and minimums as x_i^-

$$x_i^+=x_{2i},\ i=\overline{1,2^{k-1}};\quad x_i^-=x_{2i-1},\ i=\overline{1,2^{k-1}+1}.$$

"Peak" number i $(i=\overline{1,2^{k-1}})$ has maximum at the point with abscissa x_i^+ and is limited by the points with abscissas x_i^- and x_{i+1}^- (Fig. 1, *a*).

With the entered designations the curve equation at the intervals of growth $[x_i^-,x_i^+]=[x_{2i-1},x_{2i}]$, $i=\overline{1,2^{k-1}}$ looks like

$$y=\widehat{F}_k(x)=F(x)-F(x_i^-)=F(x)-F(x_{2i-1})=F(x)-\frac{2i-2}{2^k},$$

and at the intervals of decline $[x_i^+,x_{i+1}^-]=[x_{2i},x_{2i+1}]$, $i=\overline{1,2^{k-1}}$ we have

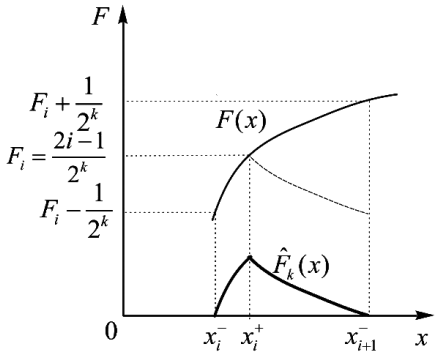
$$y=\widehat{F}_k(x)=-F(x)+F(x_{i+1}^-)=-F(x)+F(x_{2i+1})=-F(x)+\frac{2i}{2^k}.$$

Let us compile a general expression for the folded cumulative function of the k^{th} order

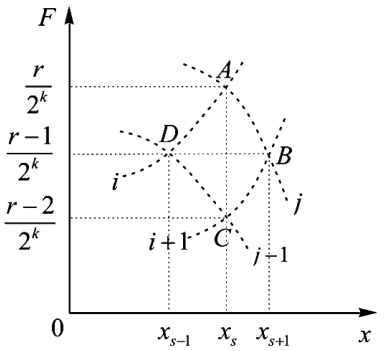
$$\widehat{F}_k(x)=\sum_{i=1}^{2^{k-1}}\left\{\left[F(x)-\frac{2i-2}{2^k}\right]\cdot Box\left(F(x),\frac{2i-2}{2^k},\frac{2i-1}{2^k}\right)+\left[\frac{2i}{2^k}-F(x)\right]\cdot Box\left(F(x),\frac{2i-1}{2^k},\frac{2i}{2^k}\right)\right\}, \qquad (4)$$

where $Box(y, a, b) = H[y - a] - H[y - b]$ – "Boxcar" function, $H(x)$ – the Heaviside function.

Based on the same data and, accordingly, the same cumulative function, it is possible to obtain folded functions of different

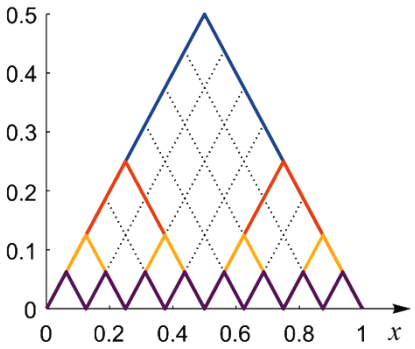


a

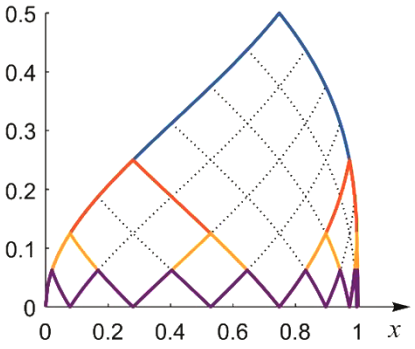


b

Fig. 1. Construction scheme of the graph section $\widehat{F}_k(x)$ *with* $x\in[x_i^-,x_{i+1}^+]$ *(a) and diagram of a triangle face (b)*



a



b

Fig. 2. Cumulative triangles for uniform distribution (a) and beta distribution with parameters $\alpha = 0.5$, $\beta = 0.3$ *(b)*

orders. Graph at $k > 1$ has a saw-looking shape. Let us construct a graph of the cumulative function of the 1st order $\widehat{F}_1(x)$ in the same coordinates and extend the lines of growth and decline of the graph $\widehat{F}_k(x)$ to the intersection with the curves of the graph $\widehat{F}_1(x)$. The constructed lines form a grid. The intersection points of the grid lines will be called nodes, the segments of lines between adjacent nodes are called edges, and the area of the plane bounded by adjacent lines is called faces.

Folded cumulative functions of different orders illustrate the distribution of a random variable at different scales. That is why for ease of visual perception it makes sense in addition to graphs $\widehat{F}_k(x)$ and $\widehat{F}_1(x)$ also highlight graphs $\widehat{F}_2(x),\ldots,\widehat{F}_{k-1}(x)$, that are superimposed on part of the constructed grid. The graphical object obtained in this way will be called a cumulative triangle of the k^{th} order (k – triangle), according to the highest order of the function used, and marked as $\Delta\widehat{F}_k$.

When depicting a triangle, graphs $\widehat{F}_1(x),\ldots,\widehat{F}_k(x)$, and a grid between them are first built. In this case, a dashed black line is used. Next, the lines of the graphs $\widehat{F}_1(x),\widehat{F}_2(x),\ldots,\widehat{F}_k(x)$ are drawn sequentially with a change in color. Thus, the graph of the next function covers part of the previous one. Only the graph of the folded cumulative function of the higher order $\widehat{F}_k(x)$ is fully in its own color.

As an example, Fig. 2 shows cumulative triangles of the 4th order for uniform and beta distributions on the finite interval [0, 1].

Let us consider the lines $y = y(x)$ forming the grid in the coordinate system. All the lines come from points with coordinates $(x_i^-,0)=(x_{2i-1},0), i=\overline{1,2^{k-1}+1}$. Two lines emerge from each point in the middle of the interval (a, b). One line is parallel to the rising part of the graph $\widehat{F}_1(x)$, and the second one is parallel to the falling part of the same graph. We will use the symbols "u" and "d" for y, respectively, and in addition identify the family of lines of the first type (increasing) with indices $i=\overline{1,2^{k-1}}$, and the second (decreasing) $j=\overline{1,2^{k-1}}$. Note that the lines of different families with the same numbers cross the Ox axis at different points, namely

$$(x_i^-,0)=(x_{2i-1},0), i=\overline{1,2^{k-1}};\quad (x_{j+1}^-,0)=(x_{2j+1},0), j=\overline{1,2^{k-1}}.$$

The equations of the lines look like

$$y=y_u(x,i)=F(x)-\frac{2i-2}{2^k}, i=\overline{1,2^{k-1}}; \tag{5}$$

$$y=y_d(x,j)=-F(x)+\frac{2j}{2^k}, j=\overline{1,2^{k-1}}. \tag{6}$$

The range of values x for the function $y = y_u(x, i)$ (5) is limited on one side by the point x_{2i-1} on the Ox axis from which the line begins, and on the other side, by the abscissa of the point of intersection with line $y = y_d(x, 2^{k-1})$, which are the falling part of $\widehat{F}_1(x)$. From the latter we can state the following

$$F(x)-\frac{2i-2}{2^k}\le -F(x)+1.$$

Whence, according to (3), we can write, that $x\le x_{2^{k-1}+i}$.

Thus, equation (5) should be supplemented by the condition $x\in\left[x_{2i-1},x_{2^{k-1}+i}\right]$.

Similarly, the condition for equation (6) is the following $x\in\left[x_{j+1},x_{2j+1}\right]$.

Any node lies at the intersection of grid lines and its coordinates $x_* = x_*(i, j)$, $y_* = y_*(i, j)$ are determined by a pair of numbers (i, j). From (5) and (6) in the node (i, j) we have

$$y_*=\frac{j-i+1}{2^k}=\frac{r}{2^k};\quad F(x_*)=\frac{j+i-1}{2^k}=\frac{s-1}{2^k}.$$

From the latter, considering (3), one can obtain that $x_* = x_{i+j}$.

Thus, the node (i, j) on the plane xOy is located at the point $\left(x_{i+j}, \frac{j-i+1}{2^k}\right)=\left(x_s, \frac{r}{2^k}\right)$. It means that all nodes are located only at the intersections of the horizontal $y=\frac{r}{2^k}, r=\overline{0,1,2^{k-1}}$ and vertical $x=x_s, s=\overline{1,2^k}$ lines. Horizontal lines correspond to quantile levels from 0 to $\frac{1}{2}$ with a constant step $\frac{1}{2^k}$. Vertical lines pass only through the points on the abscissa axis, which are equal to the values x_i of the quantiles of levels $\frac{i-1}{2^k}$, $i=\overline{1,2^k+1}$.

If each node is identified by the quantile level (it coincides with the coordinate y) and the quantile value, then the transition from any node along one edge to the adjacent one leads to changes in both specified identifiers for a given value graph.

Any face has four nodes (Fig 1, *b*). Let us mark them as

$$A(i,j),\ B(i+1,j),\ C(i+1,j-1),\ D(i,j-1).$$

In the system xOy, these nodes have coordinates

$$A\left(x_s, \frac{r}{2^k}\right), B\left(x_{s+1}, \frac{r-1}{2^k}\right), C\left(x_s, \frac{r-2}{2^k}\right), D\left(x_{s-1}, \frac{r-1}{2^k}\right),$$

and the edges are described by the equations

$$AB: y=y_d(x,j)=-F(x)+\frac{2j}{2^k},\quad x\in\left[x_s, x_{s+1}\right];$$

$$BC: y=y_u(x,i+1)=F(x)-\frac{2i}{2^k},\quad x\in\left[x_s, x_{s+1}\right];$$

$$CD: y=y_d(x,j-1)=-F(x)+\frac{2j-2}{2^k},\quad x\in\left[x_{s-1}, x_s\right];$$

$$DA: y=y_u(x,i)=F(x)-\frac{2i-2}{2^k},\quad x\in\left[x_{s-1}, x_s\right].$$

The face is symmetrical with respect to the diagonal BD, lying on the straight line $y=\frac{r-1}{2^k}=\frac{j-i}{2^k}$. Another diagonal AC lies on a straight line $x = x_s$. The diagonal AC divides the face into two parts. Considering the symmetry with respect to the horizontal diagonal, we find the areas of the left S_1 and right S_2 parts of the face

$$S_1=2\int_{x_{s-1}}^{x_s}\left[F(x)-\frac{2i-2}{2^k}-\frac{j-i}{2^k}\right]dx=$$

$$=2\left[\int_{x_{s-1}}^{x_s}F(x)dx-\frac{i+j-2}{2^k}(x_s-x_{s-1})\right];$$

$$S_2=2\int_{x_s}^{x_{s+1}}\left[-F(x)+\frac{2j}{2^k}-\frac{j-i}{2^k}\right]dx=$$

$$=2\left[-\int_{x_s}^{x_{s+1}}F(x)dx+\frac{i+j}{2^k}(x_{s+1}-x_s)\right].$$

Using the relation

$$\int_a^b G(x)dx=(b-a)G(a)+\int_a^b(b-x)G'(x)dx=$$

$$=(b-a)G(b)+\int_a^b(a-x)G'(x)dx,$$

and taking into account (3), one can obtain

$$S_1=2\left[(x_s-x_{s-1})F(x_{s-1})+\int_{x_{s-1}}^{x_s}(x_s-x)F'(x)dx-\frac{s-2}{2^k}(x_s-x_{s-1})\right]=$$

$$=2\int_{x_{s-1}}^{x_s}(x_s-x)F'(x)dx;$$

$$S_2=2\left[-(x_{s+1}-x_s)F(x_{s+1})-\int_{x_s}^{x_{s+1}}(x_s-x)F'(x)dx+\frac{s}{2^k}(x_{s+1}-x_s)\right]=$$

$$=2\int_{x_s}^{x_{s+1}}(x-x_s)F'(x)dx.$$

Then

$$S_s=S_1+S_2=2\int_{x_{s-1}}^{x_{s+1}}\left|x-x_s\right|F'(x)dx; \tag{7}$$

$$\Delta S_s=S_2-S_1=2\left[\int_{x_s}^{x_{s+1}}(x-x_s)F'(x)dx-\int_{x_{s-1}}^{x_s}(x_s-x)F'(x)dx\right]=$$

$$=\frac{2}{2^k}\left[\frac{1}{F(x_{s+1})-F(x_{s+1})}\int_{x_{s-1}}^{x_{s+1}}xF'(x)dx-x_s\right]. \tag{8}$$

For a random value in the interval $[x_{s-1}, x_{s+1}]$ the value x_s is a median, and $\frac{1}{F(x_{s+1})-F(x_{s+1})}\int_{x_{s-1}}^{x_{s+1}}xF'(x)dx$ is a mean value. Therefore, it follows from (7 and 8) that for values from the interval $[x_{s-1}, x_{s+1}]$ with accuracy up to the constant factors for the given triangle (with given k) can be stated: the face area is equal to the mean of a random variable deviation from the median; the difference in the areas of the left and right parts of the face is equal to the difference between the mean and median.

Thus, the area of the face $\dot{S}_s$ can be used as a characteristic of the concentration of random variable values near the abscissa of the upper node x_s: the smaller is the area, the higher is the concentration. The sign of the Pearson asymmetry coefficient is determined by the sign of the difference between the mean and median [15]. Hence, the difference in areas ΔS_s can be used to estimate the asymmetry of the distribution on the interval $[x_{s-1}, x_{s+1}]$. If the difference is zero, then the distribution is symmetric. If the difference is greater than zero, then the distribution has positive asymmetry, that is, the density of the distribution is shifted to the left border of the range. In the opposite case, when the difference is less than zero, the asymmetry is negative and the density of the distribution is shifted to the right border.

Both values S_s and ΔS_s are easily visually evaluated and compared at different S.

Similarly, we can use the amalgamation of faces and draw conclusions about larger intervals. If the vertex of the constituent face is at the node (x_s, y_*), extreme left and right nodes have coordinates $\left(x_{s-m}, y_*-\frac{m}{2^k}\right)$ and $\left(x_{s+m}, y_*-\frac{m}{2^k}\right)$ respectively, and the bottom node has coordinates $\left(x_s, y_*-\frac{m}{2^{k-1}}\right)$, then conclusions about the distribution properties can be made relative to the interval (x_{s-m}, x_{s+m}) and the point x_s. It is easier to do this along the edges formed by folded cumulative functions of a lower order and depicted in the appropriate color (Fig. 2).

Application to sampling of a modulated random variable. When experimentally investigated, the cumulative function is calculated from a set of data $z_{01}, z_{02}, z_{03}, \ldots, z_{N_0}$. In most cases, all elements in the sample are considered equal, while the values of different elements may coincide. The elements are arranged in the order of increasing their values. An element whose value is equal to the previous one is deleted. Thus, the data set $z_1 < z_2 < z_3 < \ldots < z_N$, where $N \le N_0$ is obtained. Each value observed in the initial set is assigned a value $\frac{1}{N_0+1}$ once, and each value observed n times is assigned a value $\frac{n}{N_0+1}$. Dividing by $N_0 + 1$, and not by N_0 ensures the exact calculation of the median, and in addition, eliminates the contradictory equality of the cumulative function of unity at the experimentally observed farthest point z_N.

Denoting the constructed numerical sequence $f_1, f_2, f_3, ..., f_N$, the empirical cumulative function can be defined as

$$F(x)=\sum_{z_i\le x} f_i, \quad x\in(-\infty,\infty). \tag{9}$$

Graph (9) has a stepped form. Assuming that N is not too small, instead of (9) we will use a lattice function with a variable discreteness interval

$$F(z_i)=\sum_{j=1}^{i} f_j, \quad i=\overline{1,N}, \tag{10}$$

and form an envelope from segments of linear interpolation between adjacent nodes. This function is continuous and non-decreasing in the interval $[a, b] = [z_1, z_N]$, which corresponds to the requirements for the theoretical cumulative function considered above. The differences are that the quantiles of the levels of multiples $\frac{1}{2^k}$ do not coincide with the discrete z_i ones and the conditions $F(a) = 0$, $F(b) = 1$ are not fulfilled at the ends of the interval.

To solve the first of the mentioned problems, auxiliary counts were introduced into the lattice function (10) at the points that are quantiles of the required level, found along the interpolation segments. These counts were used when constructing the lines of the graph, but, unlike the points related to the data, they were not clearly marked on the graph.

The second of the mentioned differences requires some change in the calculation relationships used above.

For the sake of generalization, let us allow some arbitrary deviations at the ends of the domain of the cumulative function, that is, we assume that

$$F(a) = F_a \ge 0; \quad F(b) = F_b \le 1.$$

It should be noted that the same conditions must be taken into account in the case of the theoretical cumulative function, when it is defined in the infinite interval. Instead of (4) now we use

$$\widehat{F_k}(x)=\sum_{i=1}^{2^{k-1}}\left\{\left[F(x)-\phi_i^-\right]\cdot Box\left(F(x),\phi_i^-,\frac{2i-1}{2^k}\right)+\right.$$
$$\left.+\left[\phi_{i+2}^- - F(x)\right]\cdot Box\left(F(x),\frac{2i-1}{2^k},\phi_{i+2}^-\right)\right\},$$

where

$$\phi_i^- = \begin{cases} F_a, & i=1 \\ F_b, & i=2^{k-1}+1. \\ \dfrac{2i-2}{2^k} \end{cases}$$

To satisfy the requirement of obtaining 2^{k-1} maxima ("peaks") the following restrictions should be put on the highest order $\hat{F}_k(x)$

$$0\le F_a<\frac{1}{2^k}, \quad \frac{2^k-1}{2^k}<F(b)\le 1.$$

From the latter, the maximum possible order value is

$$k_{\max}=\lceil \min[-\log_2 F_a, -\log_2(1-F_b)]\rceil =$$
$$=-\lfloor \max[\log_2 F_a, \log_2(1-F_b)]\rfloor - 1.$$

All other relationships and the methodology of constructing a triangle remain the same.

Fig. 3 shows cumulative triangles constructed from a data set obtained as a mixture of an equal number of elements taken from two normally distributed random variables with means equal to 1 and 2, and variances equal to 1 and 0.2, respectively. The sample consisted of 60 elements. For the triangle shown in Fig. 3, *a*, only the first 20 elements of the sample were used, and for the triangle in Fig. 3, *b* all 60 elements were involved. The difference in images illustrates the selective variability. An increase in the sample size naturally leads to an increase in the range of values and, at the same time, to a stabilization of the overall shape of the triangle. Zones of concentration of values are quite simply determined, especially if, in addition to the

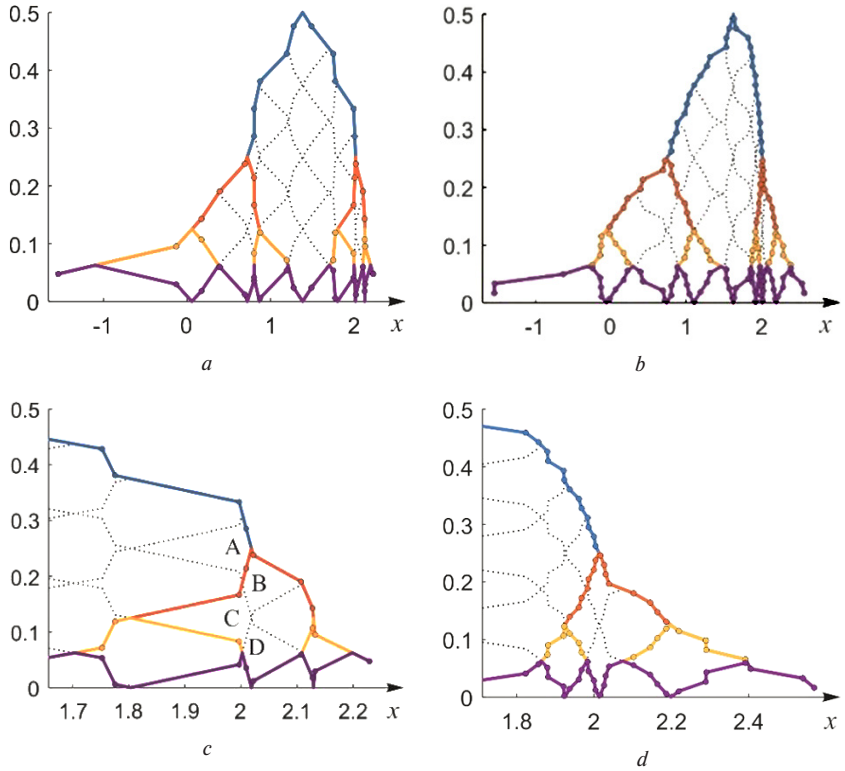


Fig. 3. Cumulative triangles: (a, b) and their parts (c, d), which are constructed from a simulated sample of a random variable of 60 elements: a, c – involving the first 20 values; b, d – involving all sample

reduction of the face area, the asymmetry of the neighboring faces is taken into account. Thus, Figs. 3, *c*, *d* show enlarged parts of triangles in the same zone of values x obtained with different sample volumes. In the case of a larger sample (Fig. 3, *d*), to determine the zone of increased concentration of a random variable, it is sufficient to involve only the area of faces. The smallest area refers to $x \approx 2$. In the case of a smaller sample (Fig. 3, *c*), the faces that are tangent to the vertical line $x = 2$ are too large and the concentration of values is manifested in the asymmetry of the neighboring faces. Only those adjacent faces whose top node is close to $x = 2$ are taken into account. In the figure, these faces are marked with letters. In the faces to the left (A, C), the area of the right part is much smaller than the area of the left part, which means that the distribution on the interval covering this face has a negative asymmetry and the density of the distribution is shifted to the right border, i. e. to $x \approx 2$. In the faces on the right (B, D), the area of the left part is much smaller than the area of the right one, accordingly, this face has positive symmetry and the density of the distribution is shifted to the left border, i. e. again to $x \approx 2$. Thus, based on the constructed cumulative triangle, it is possible to draw conclusions regarding the features of the empirical data distribution.

Results. ***Application to empirical data.*** An additional convenience of using the cumulative triangle is the simplicity of the calculation and visualization algorithm. The graphs presented in this article are obtained as the results of a simple code written in the Matlab software of "The MathWorks" company.

The proposed approach to visualization of experimental data, in particular, was applied to the analysis of seismoacoustic signals used to assess the stability of mining excavations under the impact of working destructive mechanisms. Predictive estimation of the excavation state is based on the power spectral density of the registered acoustic signals [16, 17]. Fig. 4 shows the dependence of the unilateral power spectral density G on the signal frequency f. The spectra are computed based on the signals obtained in the same excavation, but at different time intervals in segments of different durations: signal No. 1 matches to duration of 38 s, signal No. 2 matches to duration of 21 s. The different duration of the signals is due to the unregulated time of the rock-destroying mashine action, which is the source of the generated oscillations. Full spectra occupy a much larger frequency range than shown in Fig. 4, namely from 0 to 5,512.5 Hz (Nyquist frequency). Counts G at frequencies not shown in the graph have much smaller values, but the number of such counts is significant and must be taken into account when calculating the cumulative function. The spectra shape indicates the noise-like nature of the signal with a rather complex spectral structure, and individual outliers (extrema) of the spectral density are variable and strongly depend on the parameters of the spectral evaluation procedure [18]. The use of a cumulative triangle to represent the distribution of signal power simplifies the systematization and analysis of signals, at least in the initial stages of research. Fig. 5 shows cumulative triangles of the fifth order for the same signals, whose spectra are shown in Fig. 4.

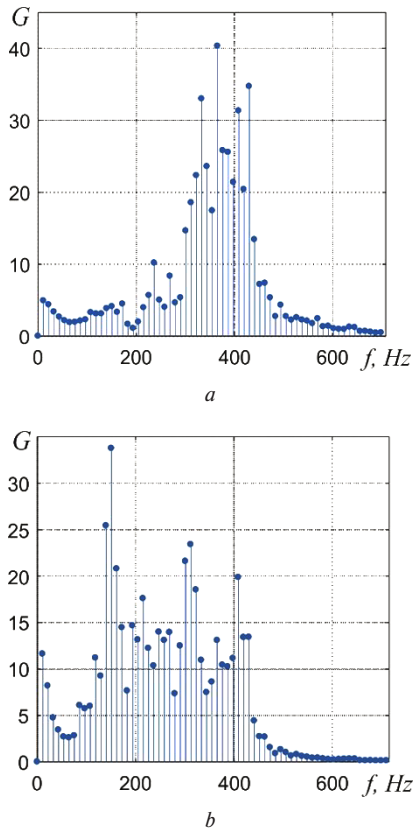


Fig. 4. Spectral densities of mine seismoacoustic signals power:
a – signal No. 1; b – signal No. 2

The comparison of triangles obviously indicates not only a change in the frequency interval and medians of the power distribution, but also changes in the structure of the power distribution. Moreover, the features of changes can be observed at different scale levels based on the area and shape of faces with different sizes.

Conclusions.

1. Folded cumulative function of the k^{th} order $\widehat{F}_k(x)$ is introduced as a generalization of the commonly known folded cumulative function.

2. A new geometric object called the cumulative triangle is proposed for visualizing the empirical distribution function. Outside, the triangle is limited by the graph $\widehat{F}_1(x)$ and has in its field the graphs of several folded cumulative functions with increasing orders $\widehat{F}_2(x), \ldots, \widehat{F}_k(x)$. Based on $\widehat{F}_k(x)$, an auxiliary grid of lines is applied to the triangle, which divides the area of the triangle into curvilinear quadrilaterals (faces).

3. It is shown that:

- the area of the face is equal to the average value of the modulus of deviations of the random variable from the median and, thus, can be used as a characteristic of the concentration of the random variable values near the abscissa of the upper end of the face;

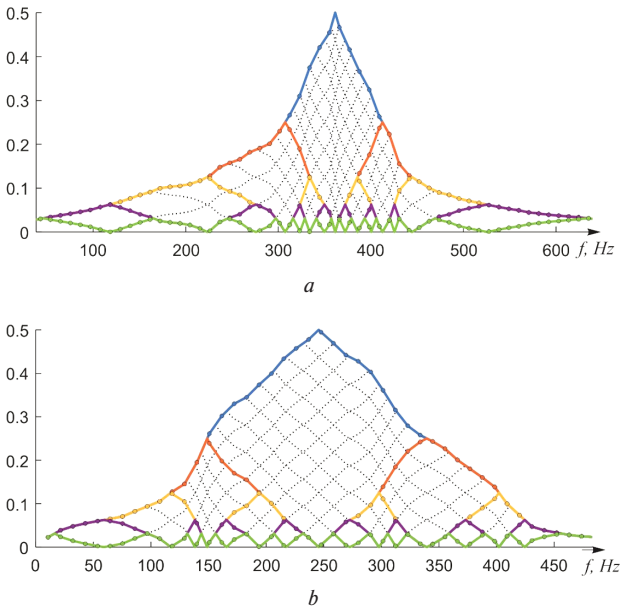


Fig. 5. Cumulative triangles of power distribution of mine seismoacoustic signals:
a – signal No. 1; b – signal No. 2

- the difference in the areas of the left and right parts of the face is equal to the difference between the mean and the median and, thus, can be used to estimate the asymmetry of the distribution on the interval, which is the projection of the face on the axis of random variable;

4. As examples, cumulative triangles are developed for:

- samples of random numbers generated by the random number generator;

- spectral power densities of seismo-acoustic signals, which were registered during the observed state of the mining excavation under conditions of working rock-crushing units.

5. Shortcomings. Developing a cumulative triangle is more difficult than constructing a histogram or box plot. If there is a significant number of samples, it may be more appropriate to use other types of visualization to compare their distributions. The redundancy of the geometric image, manifested in the duplication of faces in the vertical direction should be noted.

At the same time, the cumulative triangle makes it possible to simultaneously detail and generalize the properties of experimentally obtained data at different scale levels. Therefore, it should be preferred when studying data with complicated and varying distributions.

References.

1. Wilke, C. O. (2019). Fundamentals of Data Visualization. *O'Reilly Media*. Retrieved from https://data.vk.edu.ee/powerbi/opikud/Fundamentals_of_Data_Visualization.pdf.

2. Scott, D. W. (2010). Scott's rule. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*. https://doi.org/10.1002/wics.103.

3. Chen, Y. C. (2017). A tutorial on kernel density estimation and recent advances. *Biostat. Epidemiol, 1*, 161–187. Retrieved from https://arxiv.org/pdf/1704.03924.pdf.

4. Weglarczyk, S. (2018). Kernel density estimation and its application. *In ITM Web of Conferences; EDP Sciences: Les Ulis, France, 23,* 00037. https://doi.org/10.1051/itmconf/20182300037.

5. Scott, D. W. (2018). Kernel density estimation. *Wiley StatsRef: Statistics Reference Online*, 1-7. https://doi.org/10.1002/9781118445112.stat07186.pub2.

6. Koutsoyiannis, D. (2022). Replacing Histogram with Smooth Empirical Probability Density Function Estimated by K-Moments. *Sci, 4*, 50. https://doi.org/10.3390/sci4040050.

7. Karczewski, M., & Michalski, A. (2022). A data-driven kernel estimator of the density function. *Journal of Statistical Computation and Simulation*, *92*(17), 3529-3541. https://doi.org/10.1080/00949655.2022.2072503.

8. Park, K. I. (2018). Basic Mathematical Preliminaries. In *Fundamentals of Probability and Stochastic Processes with Applications to Communications. Springer.* Cham. https://doi.org/10.1007/978-3-319-68075-0_2.

9. Weaver, K. F., Morales, V. C., Dunn, S. L., Godde, K., & Weaver, P. F. (2017). *An Introduction to Statistical Analysis in Research: With Applications in the Biological and Life Sciences*. Germany: Wiley.

10. Marmolejo-Ramos, F., & Tian, S. (2010). The shifting boxplot. A boxplot based on essential summary statistics around the mean. *International Journal of Psychological Research, 3*(1), 37-45. https://doi.org/10.21500/20112084.823.

11. Wickham, H., & Stryjewski, L. (2011). *40 years of boxplots*. Retrieved from https://vita.had.co.nz/papers/boxplots.pdf.

12. Xue, J.-H., & Titterington, D. M. (2011). The p-folded cumulative distribution function and the mean absolute deviation from the p-quantile. *Statistics and Probability Letters, 81,* 1179-1182. https://doi.org/10.1016/j.spl.2011.03.014.

13. Olshaker, H., Buhbut, O., Achiron, A., & Dotan, G. (2021). Comparison of keratometry data using handheld and table-mounted instruments in healthy adults. *International Ophthalmology, 41*(1). https://doi.org/10.1007/s10792-021-01909-8.

14. Stokar, J., Leibowitz, D., Durst, R., Shaham, D., & Zwas, D. (2019). Echocardiography overestimates LV mass in the elderly as compared to cardiac CT. *PLoS ONE, 14*(10), e0224104. https://doi.org/10.1371/journal.pone.0224104.

15. Weisstein, E. W. (2024, April 25). Pearson's Skewness Coefficients. *From MathWorld--A Wolfram Web Resource*. Retrieved from https://mathworld.wolfram.com/PearsonsSkewnessCoefficients.html.

16. Sdvyzhkova, O., Golovko, Yu., Dubytska, M., & Klymenko, D. (2016). Studying a crack initiation in terms of elastic oscillations in stress strain rock mass. *Mining of Mineral Deposits. Dnepr: National Mining University (Dnepr, Ukraine), 10*(2), 72-77. https://doi.org/10.15407/mining10.02.072.

17. Golovko, Yu. (2017). Estimation of seismoacoustic signal spectral parameters under the current prediction of gasodynamic phenomena in mines. *Heotekhnichna mekhanika, 134*, 141-154.

18. Golovko, Yu. M. (2023). Spectral estimation of a broadband time-limited noise signal. *Matematychne modeliuvannia*, *2*(49), 86-97. https://doi.org/10.31319/2519-8106.2(49)2023.292638.

Кумулятивний трикутник для візуального аналізу емпіричних даних

Ю. М. Головко, *О. О. Сдвижкова*[*]

Національний технічний університет «Дніпровська політехніка», м. Дніпро, Україна

[*] Автор-кореспондент e-mail: sdvyzhkova.o.o@nmu.one

Мета. Розробка графічного об'єкту для візуального аналізу, що давав би можливість одночасно оцінювати як загальні характеристики, так і деталі розподілу емпіричних даних.

Методика. Обгрунтування доцільності й послідовності створення кумулятивного трикутника, а також доведення його властивостей виконувалось із залученням геометричних побудов, узагальнених і решітчастих функцій. Побудова кумулятивного трикутника здійснювалася програмно у середовищі «Matlab». Вибірки випадкових величин з відомими законами розподілу отримувалися з використанням генератора псевдовипадкових чисел. У якості емпіричних даних використані попередньо обчислені залежності спектральної щільності потужності сейсмоакустичних шумоподібних сигналів.

Результати. Уведена згорнена кумулятивна функція *k*-го порядку як узагальнення відомої згорненої кумулятивної функції. Використовуючи згорнені кумулятивні функції, побудовано геометричний об'єкт – кумулятивний трикутник, призначений для візуалізації емпіричної функції розподілу. На трикутник наносяться лінії, що розбивають його на плоскі криволінійні чотирикутники. Показано, що площа грані може використовуватися як характеристика концентрації значень випадкової величини біля абсциси верхнього вузла грані, а різниця площ лівої та правої частин грані дає оцінку асиметрії розподілу на проміжку, що покриває грань.

Наукова новизна. Запропоновано новий графічний об'єкт для візуального аналізу розподілу емпіричних даних. Показано, яким чином, спираючись на його вид, можна робити висновки як відносно характеристик усієї вибірки, так і окремих проміжків функції розподілу.

Практична значимість. Кумулятивний трикутник може бути корисним доповненням до графічних засобів візуалізації. Його використання дає можливість візуальної одночасної деталізації та узагальнення властивостей експериментально отриманих даних на різних масштабних рівнях, що є особливо цінним, коли дані мають ускладнені й мінливі розподіли.

Ключові слова: *візуалізація, аналіз даних, функція розподілу, згорнута кумулятивна функція, спектр потужності*

The manuscript was submitted 30.03.24.