Mathematics and Statistics 8(6): 637-644, 2020
DOI: 10.13189/ms.2020.080603 http://www.hrpub.org

Construction of Lorenz Curves Based on Empirical Distribution Laws of Economic Indicators

Aleksandr Bochkov[1,*], Dmitrii Pervukhin[2], Aleksandr Grafov[3], Veronika Nikitina[4]

[1]Department of Information and Computing Systems, Faculty of Automation and Intelligent Technologies, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
[2]Department of System Analysis and Management, Faculty of Economics, St. Petersburg Mining University, St. Petersburg, Russia
[3]Department of Economic Security, Faculty of Business, Customs and Economic Security, St. Petersburg State University of Economics, St. Petersburg, Russia
[4]Department of Neurology and Manual Therapy, Faculty of Postgraduate Studies, Academician I.P. Pavlov First St. Petersburg State Medical University, St. Petersburg, Russia

Received September 9, 2020; Revised October 12, 2020; Accepted October 30, 2020

Cite This Paper in the following Citation Styles

(a): *[1] Aleksandr Bochkov, Dmitrii Pervukhin, Aleksandr Grafov, Veronika Nikitina , "Construction of Lorenz Curves Based on Empirical Distribution Laws of Economic Indicators," Mathematics and Statistics, Vol. 8, No. 6, pp. 637 - 644, 2020. DOI: 10.13189/ms.2020.080603.*

(b): *Aleksandr Bochkov, Dmitrii Pervukhin, Aleksandr Grafov, Veronika Nikitina (2020). Construction of Lorenz Curves Based on Empirical Distribution Laws of Economic Indicators. Mathematics and Statistics, 8(6), 637 - 644. DOI: 10.13189/ms.2020.080603.*

Abstract The quality of construction of Lorenz curves depends on the features of the information used. As a rule, information is represented by a sample of values of the studied indicator, which is checked for unevenness. Economic indicators of income and cost, and features of their samples are considered. The feature of the cost economic indicator associated with the presence in the sample of its values of the clot is highlighted (the concentration of values on a small segment of the entire range of sample). It is shown that the established order of constructing empirical laws based on such samples does not give the desired effect when constructing Lorenz curves due to the loss of information content of the sample in the places of the clot. The purpose of this article is to improve the quality of the Lorenz curve by increasing the information content of the sample with a clot by applying the clustering procedure when constructing an empirical law. A step-by-step clustering procedure is proposed for dividing the entire range of sample into intervals to construct an empirical distribution law, which is an element of the novelty of this study. A specific example shows how to improve the quality of building a Lorenz curve using this procedure. In addition, it is shown that Lorenz curves for economic indicators can be constructed directly on the basis of the empirical distribution law and at the same time take into account its features.

Keywords Lorenz Curve, Empirical Distribution Law, Clustering, Histogram, Economic Indicators, Information Content of the Sample

1. Introduction

The issue of constructing Lorenz curves is relevant; this is confirmed by the fact that they are used in various fields related to economics, biology, medicine, technology, etc. Lorenz curves are interesting because they provide an analysis of economic inequality, uneven distribution of income, processes of stratification of mixtures, uneven costs of research work, etc. Researchers are attracted by the visibility of the deviation of the uneven distribution of the considered indicator from its uniform distribution and how it will change after the control effect on this indicator.

There are known works related to the construction of empirical laws and Lorenz curves [1–26], in which the construction was carried out in different interpretations in accordance with the features of the object of study. In General, following the classical view of the process of

constructing Lorenz curves, the following steps can be formed. At the first stage, a set of values of the unevenness indicator is formed. The features of the sample of values are associated with the nature of the indicator under consideration – biological indicator, economic, technical, and chemical, etc. The distribution of random variables over the entire range of sample is analyzed. At the second stage, a histogram of the indicator is constructed (an empirical law), on the basis of which a theoretical distribution law is selected according to the consent criterion. The third stage is related to getting on the theoretical law functions Lorenz, reflecting the uneven distribution of the index relative to the line of equal distribution. The conclusion is made about the use of one or another type of Lorenz curve and the criterion reflecting unevenness. The fourth stage is associated with a visual representation of the Lorenz curve, which characterizes the uneven distribution of the indicator under consideration, on the graph. By varying the parameters of the Lorenz curve, interpreting the results of the study and forming the necessary conclusions.

The work [1] is aimed at obtaining all possible types of Lorenz curves and automating the calculation of their parameters using the developed *lorenz* program, which at the first stage of constructing the Lorenz curve fully supports the estimation of variance for complex samples, and at subsequent stages evaluates the Lorenz curves and concentration curves for specific data. This means that calculation data is always available for Lorenz curves, and the results are displayed as a graph only when necessary. Calculated dependencies on the construction of the Lorenz curve based on a sample of values of a specific economic indicator are presented. This is valuable when constructing a Lorenz curve based on simulation modeling, when there is a specific law of distribution of an economic indicator. Then, using the algorithm for simulating random variables according to the selected distribution law, presented, for example, in [2], it is possible to obtain the calculated values of the Lorenz curve and plot it. However, this method of constructing the Lorentz curve is quite complex and requires additional research, especially if the original sample of random variables has specific differences, such as the presence clots of random variables.

In [3] reflects all the stages of constructing Lorenz curves, the implementation of which requires a fairly representative sample of random variables, which is not always found in real conditions of economic indicator research, and the need for time spent on the selection of the theoretical distribution associated with additional research. At the same time, it is necessary to monitor the correspondence of the essence of the studied process of uneven economic indicators to the features of the selected distribution law. Getting the expression of the Lorenz function itself involves quite complex transformations, during which you can make an inaccuracy or just an error, which will distort the appearance of the Lorenz curve. For example, expressions for the Lorenz function are given in [4], but only for such laws as the uniform law and the lognormal law. Getting them does not cause much difficulty.

In [5, 6], attention is paid to the derivative of the Lorenz curve – the Gini coefficient and its behavior in assessing the uneven distribution of economic indicators. The analysis of empirical distributions of income of the world's countries over several years has been carried out and it is shown that the Gini indices are centered on the value of 33.33% corresponding to the Gini index of uniform distribution and that the Lorenz curves of these distributions are consistent with the Lorenz curves of lognormal distributions. This corresponds to research in [7], where the Gini coefficient of income inequality in the rural region is 35%. In fact, the third stage of constructing the Lorenz curve is well considered.

In [8,9] presents the results of studies directly with Lorenz curves, how their appearance affects the original empirical law. However, nothing is said about such an economic indicator as the cost, the unevenness of its distribution. There are no examples of empirical laws of such an indicator and their corresponding Lorenz curves.

In [10], a method is given for constructing the Lorenz curve based on the empirical law of distribution, assuming a uniform density in each group interval, which leads to an overestimation of the total average income. It is shown that the average absolute errors of the Gini index estimation obtained by two methods using group values of averages are significantly less than by the method based on a histogram. This allows you to find the average values when dividing the histogram intervals into equal parts and continue the construction process further. But there are no examples of constructing a Lorenz curve for the case when there are samples with a concentration of random variables at a small interval 8–10 times smaller than the span of the entire sample. In addition, the processes of constructing Lorentz curves based on trigonometric functions, a new generalized Weibull-Pareto distribution, a four-parameter distribution called a non-standard generalized power distribution, and other distribution laws [11-15]. All of them take into account the features of the source data, which are considered in specific examples. However, there are no examples of such samples (samples with a concentration of random variables over a small interval). Meanwhile, when constructing the Lorenz curve, different situations arise related to the information conditions of the study. One of these situations is a strong spread of random variables. In official theory, such quantities are statistically recognized as anomalous and are simply discarded, and then the distribution laws are constructed without them. An example of this situation is a sample of the cost of research works (R & d) taken from the R & d register for 2019 [16]. Funding for research related to strategic studies of economic development can be allocated funds that are tens of times higher than the cost of conventional local research,

which is the majority in the register. Therefore, it is necessary that such situations must be taken into account when constructing the Lorenz curve. In addition, it should be noted that the selection of the theoretical distribution law is based primarily on the construction of an empirical histogram, the type and information content of which strongly depends on the number and length of the histogram intervals. The situation is complicated by the fact that for economic indicators, for example, income, the cost of household work, a common feature is the asymmetry of distribution and the presence of so-called clots of random variables in the sample. The term clot is used by researchers at the conversational level, so in the future we will understand the clot as the final range of possible values of the studied parameter, in which they are located with a probability close to one (for example, from 0.7 to 0.90), and the final interval itself is 8–10 times smaller than the entire range of sample. The construction of an empirical distribution law for a sample with such a clot is very problematic, since it is known that the use of large intervals in the construction of an empirical law is associated with the loss of information [4]. This loss of information will affect the quality of the Lorenz curve.

Thus, there is a contradiction between the presence of samples of random variables with clots in the economy and the lack of methods and procedures for preserving the information content of the sample in specific applications, for example, when constructing Lorenz curves. With this in mind, the refined goal of this study is to develop a procedure for constructing an empirical law for samples of random variables with clots in order to preserve informative content of the sample when constructing the Lorenz curve.

In [17-19], when constructing an empirical law, the question of forming intervals using the so-called binning procedure is considered, which is associated with a decrease in the size of the histogram intervals. However, when changing the length of the interval, there is some uncertainty in finding the degree of decrease (increase), in addition, the problem of non-parametric statistics was solved, which distracted attention and increased the time for research. For example, the Sturges' formula, which gives the dependence of the number of intervals on the sample size, is currently outdated for many reasons (see [20]). The formulas of Scott [17], Freedman and Diaconis [21] are mainly used when calculating the length of intervals depending on the number of measurements and sample variance, the number of measurements and the difference between the upper and lower quartile. In [18, 19, 22–24], the length of intervals is calculated using the risk function, Shannon entropy, and other calculation methods. However, it turns out that the length of the histogram intervals is usually the same for the entire range of sample, and this leads to a loss of information when constructing an empirical histogram exactly in the places of a clot of random variables.

2. Materials and Methods

It seems that the best way out of this situation, taking into account the selected features of the known methods for constructing Lorenz curves, as well as techniques for choosing the length of intervals when constructing an empirical histogram, is to build an empirical distribution law for the initial sample and obtain the Lorenz curve itself on its basis. At the same time, an empirical histogram for sampling random variables must be constructed using the clustering procedure. This procedure involves ordering random variables in the places of the clot in relatively uniform intervals by their number. This will help better account for the clot information and improve the quality of the resulting Lorenz curve. In addition, you can adjust the degree of reduction in the size of the interval with a clot by setting how many times to reduce the interval and determining the number of clustering stages.

The algorithm for forming intervals for constructing an empirical histogram in this case will be as follows.

1. Defining the first cluster: splitting the entire set of random variables in the sample by the selected number of intervals. As a rule, at the first stage, the length of each interval (or interval step) is selected the same

$$l0 = \frac{x_{\max} - x_{\min}}{N}, \tag{1}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of random variables in the sample, and N is the number of intervals.

Researchers usually (for small samples, up to 100 values) choose 8–10 intervals. This is most likely due to the researcher's ability to store information about the number of random variables in each interval in RAM. You can use well–known formulas. Thus, N intervals are formed in the first cluster, $N_{1kl}=N$.

2. If the sample of random variables has a clot, then additional clusters must be formed. If a clot falls in one interval, two clusters are formed, a cluster that covers the interval with the clot, and a cluster that includes all the other intervals (the first cluster has one interval, and the second has N–1 intervals). The selected interval is divided into n_1 subintervals (or parts) depending on the number of random variables that fall within the selected interval. The length of each subinterval, provided they are equal, is calculated using the expression

$$l1 = \frac{l0}{n_1}. \tag{2}$$

Thus, two clusters are formed, in the first n_1 intervals, in the second N–1 intervals.

3. The total number of intervals of the entire set of random variables in the sample is determined, taking into account the two clusters formed

$$N_{2kl} = n_1 + (N-1). \quad (3)$$

4. If the problem of removing the clot is not solved in the first cluster, a third cluster is formed. The subinterval in the first cluster is divided into n_2 parts, depending on whether the clot is removed or not. The length of each part is determined by the expression

$$l2 = \frac{l1}{n_2}. \quad (4)$$

Three clusters were formed: in the first n_2 intervals, in the second (n_1– 1) intervals, in the third – (N–1).

5. The total number of intervals of the entire set of random variables for three clusters is determined

$$N_{3kl} = n_2 + (n_1 - 1) + (N-1) \text{ etc.} \quad (5)$$

Thus, the implemented two–stage procedure of clustering intervals: zero level – formed a single cluster that contains N intervals; the first stage is formed by two clusters in the first n_1 subintervals (or parts), in the second (N–1) intervals; the second step generates three cluster n_2 in the first parts, the second (n_1–1) and the third (N–1).

As can be seen from the above algorithm, to smooth out clots of random variables, the range of the entire sample of random variables is divided into intervals of different lengths, reflected by the dependencies (1), (2), (4). This algorithm can be used locally to eliminate several clots of random variables in complex samples used to construct Lorenz curves.

Confirmation that the Lorenz curve constructed according to the empirical distribution law, is actually no different from the Lorenz curve constructed according to the theoretical law can be illustrated by constructing the Lorenz curve for one of the variants of the forecast distribution of per capita income of Russia's population until 2025 (see example 1). The construction of the Lorenz curve for a sample containing a clot using the clustering procedure can be shown using the example of source data from the R & d register for 2019 (see example 2).

Expressions for the Lorenz curve function.

It is known [3-5] that the Lorenz curve for the continuous case has the following form

$$L(\alpha) = \frac{\int_0^{x_\alpha} xf(x)\,dx}{\int_0^{\infty} xf(x)\,dx}, \quad (6)$$

where x_α is the α quantile of a random variable X, $f(x)$ is the probability density of the distribution of a random variable X, $\alpha=F(x_\alpha)$, $F(x_\alpha)$ is the distribution function of a random variable X.

For a discrete case corresponding to an empirical law,

$$L_i(\alpha) = \frac{\sum_{j=1}^{i} x_j \cdot p_j}{\sum_{i=1}^{N} x_i \cdot p_i}, \quad (7)$$

where N is the number of intervals of the empirical histograms of the random variable X, x_i is the mathematical expectation of a random variable for the i-th interval of the histogram, p_i is the probability that the random variable in the i-th interval, $\alpha=F(x_{\alpha i})$ is a value of the distribution function at the point $x_{\alpha i}$, $x_{\alpha i}$ – α quantile of the random variable X for the right border of the i-th interval.

Expressions (6) and (7) show that the value of the Lorenz curves at point x_α is the ratio of the mathematical expectation of random variables for the interval [0, x_α] to the mathematical expectation of the entire set of random variables that the empirical histogram is based on.

The order of construction of the Lorenz curve according to the empirical distribution law is as follows.

1. Select a set of source data, check their reliability, representativeness of the sample, and so on.
2. Intervals are formed for constructing an empirical histogram, parameters are calculated for it, and the histogram itself is constructed. If there is a strong concentration of random variables in the places of clots, the clustering procedure is used when constructing an empirical histogram to increase the information content of the sample and the quality of the Lorenz curve graph.
3. The right boundary points of the intervals are fixed, and the values of the points of the Lorenz curve are calculated for them using expression (7). The Lorenz curve is plotted from the calculated points.

Adhering to this order of construction of the Lorenz curve, it is necessary to say that it is not necessary to count the average value for each interval of the histogram for the available sample; you can use the group average, i.e. the middle of the interval. It is taken into account that the values in the interval are evenly distributed. In [10] it is shown that for the construction of Lorenz curves, this assumption improves the overall picture of curve smoothing.

Example 1. Based on the original data [25] obtained a variant of the forecast distribution of the Russian population by average per capita money income until 2025 (see table 1). It is necessary to construct an empirical histogram of the distribution of income of the population, choose a theoretical distribution law and construct Lorenz curves accordingly. Make a conclusion about the identity of the curves.

Table 1. Initial data on the distribution of the average per capita income of the population of the Russian Federation until 2025 (variant of the forecast)

l	0; 1	1; 2	2; 3	3; 4	4; 5	5; 6
p_i	0.06	0.123	0.245	0.21	0.123	0.088
x_i	0.5	1.5	2.5	3.5	4.5	5.5
l	6; 7	7; 8	8; 9	9;10	10;11	11;12
p_i	0.07	0.042	0.018	0.0105	0.007	0.0035
x_i	6.5	7.5	8.5	9.5	10.5	11.5

In Fig. 1, an empirical histogram is constructed based on the initial data, and a theoretical distribution law – the lognormal law-is selected. A characteristic feature of economic indicators is reflected-the asymmetry of distribution.

The lognormal law is represented as:

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right],\ \mu=ln\left(\mu^{*2}/\sqrt{\mu^{*2}+\sigma^{*2}}\right),$$

$$\sigma^2=ln\left((\mu^{*2}+\sigma^{*2})/\mu^{*2}\right).$$

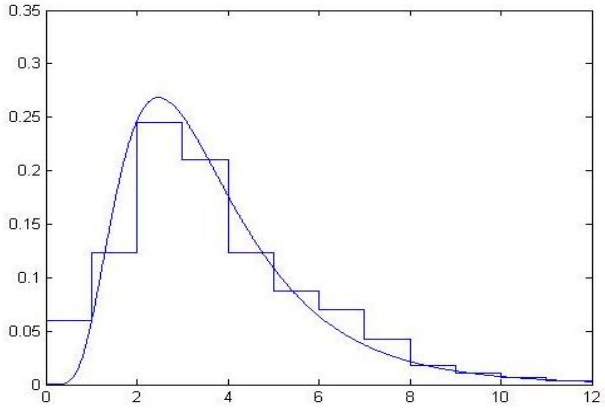

Figure 1. Graphs of the empirical histogram and probability density for the lognormal law (μ=1.181, σ=0.523)

In accordance with the table 1 μ^*=3.736, σ^{*2}=4.3933, then μ=1.1812, σ^2=0.2737. It can be seen, even without checking by the consent criterion, that the approximation is quite acceptable. Using dependencies (6) and (7), Lorenz curves are constructed. For the empirical distribution law at x=1, the value of a point on the Lorenz curve

$$L(\alpha)=L(F(1))=\frac{x_1p_1}{\sum_{i=1}^{12}x_ip_i}=\frac{0{,}03}{3{,}736}=0{,}008,$$

where x_i, p_i are taken from table 1,

x=2, $L(\alpha)=L(F(2))=\frac{x_1p_1+x_2p_2}{\sum_{i=1}^{12}x_ip_i}=\frac{0{,}2145}{3{,}736}=0{,}0574$ etc.

For clarity, see table 2 the calculated points of the Lorenz curves for the empirical law $L_э(p)$ and the lognormal distribution $L_т(p)$ are summarized. We can conclude that the Lorenz curves are identical.

Example 2. Based on the initial data in the R & d register [16], construct an empirical histogram of the distribution of the cost of research work and obtain the Lorenz curve, which reflects the uneven distribution of the cost of work.

The total number of entries in the register is M=112. Allocated N=10 intervals long $l0$ =3 million rubles – [0; 3], [3; 6], [6; 9], ... , [27; 30]. For each i–th interval, the following parameters are calculated: n_i – the number of random variables that fall into the i–th interval, p_i – the probability of a random variable falling into the i–th interval, p_i= n_i /M, x_i – the average value for the i–th interval, h_i =p_i/$l0$, i=1,…, N. Guided by these parameters in the table 3 the results of intermediate calculations for the construction of the Lorenz curve are presented. Figure 2 shows the initial empirical law of R & d cost distribution (left side of the figure) and the corresponding Lorenz curve (right side of the figure). It should be noted that the zero stage of clustering is implemented when constructing an empirical histogram.

Table 2. Calculated points $L_т$ (p) and $L_э$ (p) for the right borders of 12 intervals

$L_т(p)$	0.002756	0.074	0.254	0.459	0.632	0.759
$L_э(p)$	0.0080	0.0574	0.2214	0.4181	0.5662	0.6958
$L_т(p)$	0.847	0.906	0.945	0.971	0.989	1
$L_э(p)$	0.8176	0.9019	0.9429	0.9696	0.9892	1

Table 3. The results of the intermediate calculations (zero-tier clustering)

i	Interval	h_i	x_i	n_i	p_i	x_i p_i	$\Sigma^i_{j=1}=x_j$ p_j	$L_i(\alpha)$	α
1	0; 3	0.279762	1.5	94	0.8393	1.2589	1.2589	0.4159	0.8393
2	3; 6	0.017857	4.5	6	0.0536	0.2411	1.5000	0.4956	0.8929
3	6; 9	0.014881	7.5	5	0.0446	0.3348	1.8348	0.6062	0.9375
4	9; 12	0.005952	10.5	2	0.0179	0.1875	2.0223	0.6681	0.9554
5	12; 15	0.002976	13.5	1	0.0089	0.1205	2.1429	0.7080	0.9643
6	15; 18	0	16.5	0	0.0000	0.0000	2.1429	0.7080	0.9643
7	18; 21	0.002976	19.5	1	0.0089	0.1741	2.3170	0.7655	0.9732
8	21; 24	0	22.5	0	0.0000	0.0000	2.3170	0.7655	0.9732
9	24; 27	0.005952	25.5	2	0.0179	0.4554	2.7723	0.9159	0.9911
10	27; 30	0.002976	28.5	1	0.0089	0.2545	3.0268	1.0000	1.0000
		Σ		112	1.0000	3.0268			

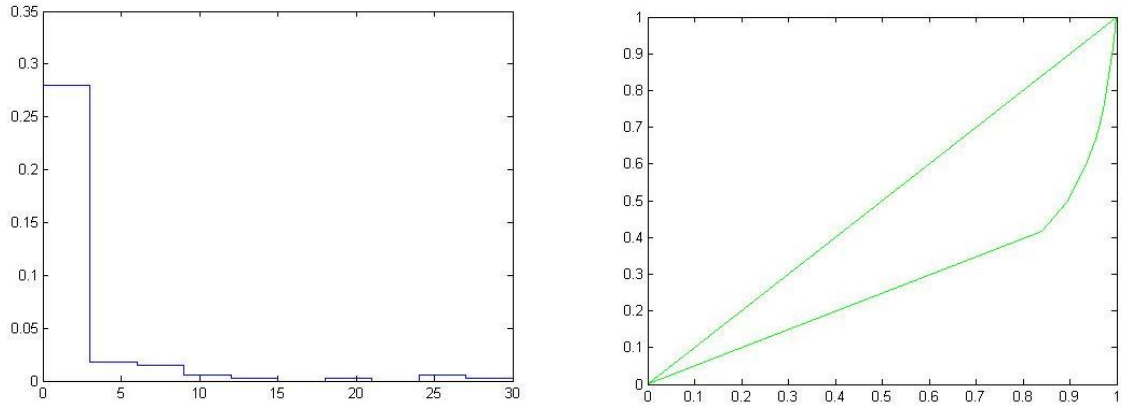

Figure 2. Graphs of the empirical R & d cost distribution law and the corresponding Lorenz curve (0-th stage of clustering's)

Table 4. The results of the intermediate calculations (first and second stage clustering)

i	Interval	h_i	x_i	n_i	p_i	$x_i\ p_i$	$\Sigma^{i}_{j=1} = x_j\ p_j$	$L_i(\alpha)$	α
1	0; 0.33	0.919912	0.165	34	0.303571	0.050089	0.050089	0.02039	0.303571
2	0.33; 0.66	0.189393	0.495	7	0.0625	0.030938	0.081027	0.032984	0.366071
3	0.66; 1	0.840335	0.83	32	0.285714	0.237143	0.318169	0.129517	0.651785
4	1; 2	0.098214	1.5	11	0.098214	0.147321	0.46549	0.189487	0.749999
5	2; 3	0.089286	2.5	10	0.089286	0.223215	0.688705	0.280351	0.839285
6	3; 6	0.017857	4.5	6	0.053571	0.24107	0.929775	0.378483	0.892856
7	6; 9	0.014881	7.5	5	0.044643	0.334823	1.264597	0.514779	0.937499
8	9; 12	0.005952	10.5	2	0.017857	0.187499	1.452096	0.591104	0.955356
9	12; 15	0.002976	13.5	1	0.008929	0.120542	1.572637	0.640173	0.964285
10	15; 18	0	16.5	0	0	0	1.572637	0.640173	0.964285
11	18; 21	0.002976	19.5	1	0.008929	0.174116	1.746753	0.71105	0.973214
12	21; 24	0	22.5	0	0	0	1.746753	0.71105	0.973214
13	24; 27	0.005952	25.5	2	0.017857	0.455354	2.202106	0.89641	0.991071
14	27; 30	0.002976	28.5	1	0.008929	0.254477	2.456583	1	1
		Σ		112	1.0000	2.456583			

In the first approximation, the empirical law can be approximated by an exponential distribution. However, this is highly controversial due to the presence of a significant clot of random variables. More research is needed.

Figure 2 shows that the Lorenz curve corresponding to an empirical histogram with a significant clot of random variables in the interval [0; 3] does not fully describe the process of uneven distribution of the R & d cost for 2019. Information stored in the place of the clot of random variables is not taken into account. In order to take this information into account, the proposed two–step clustering algorithm described above is used. In accordance with it, the zero clustering stage is implemented (one cluster with 10 intervals is considered), which did not disclose the information content of the sample due to too large clot in the first interval, i.e. information is lost. The following clustering steps are used to extract information. In this case, a zero–degree cluster is transformed: the first interval changes its step from three million rubles to one. It becomes not 10, but 12 intervals (the first cluster has 3 intervals, the second – 9). Thus, the first stage of clustering is implemented. Then the first cluster of the first stage is transformed: the first interval changes its step from 1 million rubles to 0.33 million rubles. It is no longer 12 intervals, but 14 (the first cluster is 3 intervals, the second 2, and the third 9). By setting the number of intervals at each stage of clustering, you can control the degree of reduction in the value of the histogram intervals at the location of the clot, which is a distinctive feature in the proposed algorithm.

Table 4 shows the results of intermediate calculations of the first and second stages of clustering for constructing the Lorenz curve. Using them, figure 3 shows the empirical distribution law and the corresponding Lorenz curve. It is clear that the situation has changed in a positive direction. The Lorenz curve describes much better the unevenness of the R & d cost distribution.

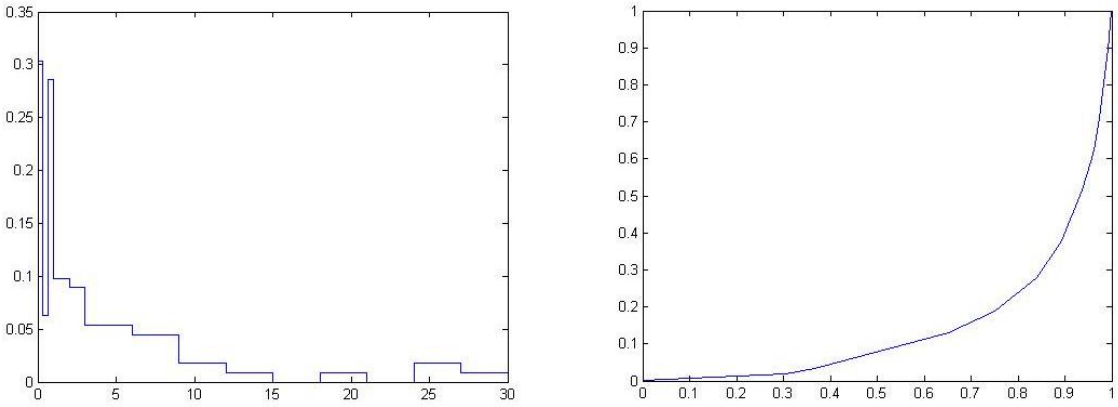

Figure 3. Graphs of the empirical law of distribution of the R & d cost and the corresponding Lorenz curve (first and second stage clustering)

3. Conclusions

When conducting research on the construction of Lorenz curves based on the empirical law of distribution of economic indicators, methods of descriptive statistics, the method of histograms, data processing technique – binning, designed to convert the initial values into ranges with small intervals (bins), methods of clustering theory were used. The construction of empirical distribution laws for economic indicators takes into account the features associated with asymmetry and the presence of clots of random variables in the initial sample. Features of random variable samples are currently observed and discussed, for example, in [26]. It is shown that when constructing Lorenz curves, it is possible to avoid the laborious procedure of selecting the theoretical distribution law and obtain Lorenz curves for complex samples with clots of random variables directly according to the empirical distribution laws.

Thus, we propose a rather original method for constructing the Lorenz curve, which significantly expands the range of applied problems for estimating the uneven distribution of various economic indicators. Specific examples show that the main features of the empirical laws of distribution of economic indicators are taken into account; the asymmetry is reflected in example 1, and the presence of clots in the sample under study in example 2.

It should be emphasized the novelty of the research constructing the Lorenz curve by the empirical distribution associated with application of the developed clustering procedure that enables us to vary the long of intervals for the histogram, which significantly extends the application of Lorenz curves for the samples of random variables with clots. At the same time, it is not necessary to carry out labor-intensive work on the selection of the theoretical distribution law and obtaining a specific expression for the Lorenz function.

REFERENCES

[1] B. Jann. Estimating Lorenz and concentration curves, The Stata Journal, vol.16, no.4, pp. 837–866, 2016.

[2] Bochkov A.P. Modeling random numbers by an arbitrary distribution law, Technico–tehnologicheskie problemy servisa, vol.49, no.3, pp. 23–27, 2019, Online available from https://unecon.ru/sites/default/files/num49.pdf

[3] F. Cowell. Measurement of Inequality. In: A. B. Atkinson and F. Bourguignon, Eds., Handbook of Income Distribution, pp. 87–166, 2000.

[4] Yves Tille, Matti Langel. Histogram-Based Interpolation of the Lorenz Curve and Gini Index for Grouped Data, The American Statistician, vol. 66, no. 4, pp. 225–231, 2012.

[5] Alfred Ultsch, Jörn Lötsch. A data science based standardized Gini index as a Lorenz dominance preserving measure of the inequality of distributions. PLOS ONE | https://doi.org/10.1371/journal.pone.0181572 August 10, pp.1–15, 2017.

[6] Kwasi A. Darkwah, Ezekiel N. N. Nortey and Anani Lotsi. Estimation of the Gini coefficient for the lognormal distribution of income using the Lorenz curve, Springer Plus, vol. 5, issue 1, pp. 1–9, 2016.

[7] Zaheer Ahmad Qureshi, Khuram Nawaz Sadozai. Income Inequality Estimation of Rural Region of KP Province: A Gini Coefficient Approach, Journal of Agriculture, vol. 32, issue 4, pp. 394–398, 2016.

[8] Gengsheng Qin, Baoying Yang, Nelly E. Belinga-Hall. Empirical likelihood–based inferences for the Lorenz curve, Annals of the Institute of Statistical Mathematics February, vol.65, issue 1, pp. 1–21, 2013.

[9] ZuXiang Wang, Russell Smyth. A Piecewise Method for Estimating the Lorenz Curve, Economics Letters, vol.129 (C), pp. 45–48, 2015.

[10] Merritt Lyon, Li C. Cheung, Joseph L. Gastwirth. The Advantages of Using Group Means in Estimating the Lorenz Curve and Gini Index from Grouped Data, The American Statistician, vol. 70, no. 1, pp. 25–32, 2016.

[11] Ding Heng, Li Yan–lai, Xiong Sheng–hua, Chen Zhen–song. Constructions of Lorenz curves based on trigonometric functions, Application Research of Computers / Jisuanji Yingyong Yanjiu, vol. 31, issue 11, pp.3273–3280, 2014.

[12] Ahmed Z. Afify, Haitham M. Yousof, Nadeem Shafique Butt. Hamedani G. G. *Pakistan*. The Transmuted Weibull–Pareto Distribution, Journal of Statistics, vol. 32, issue 3, pp.183–206, 2016.

[13] Amal Hassan, Elsayed Ahmed Elsherpieny, Rokaya Elmorsy. Odd Generalized Exponential Power Function Distribution: Properties and Applications. Gazi University Journal of Science, vol.32, pp. 351–370, 2019.

[14] ZuXiang Wang, Yew–Kwang Ng, Russell Smyth. A general method for creating Lorenz curves. Review of Income & Wealth, vol. 57, issue 3, pp.561–582, 2011.

[15] Nurulkamal Masseran, Lai Hoi Yee, Muhammad Aslam Mohd Safari, Kamarulzaman Ibrahim. Power Law Behavior and Tail Modeling on Low Income Distribution, Mathematics and Statistics, vol. 7, no. 3, pp. 70–77, 2019. DOI: 10.13189/ms.2019.070303

[16] Register of research and development 2019 on the website. Pdf [Electronic resource]. – Mode of access: https://www.rea.ru/ru/org/managements/orgnirupr/Documents/%D0%A0%D0%B5%D0%B5%D1%81%D1%82%D1%80%20%D0%9D%D0%98%D0%A0%202019%20%D0%BD%D0%B0%20%D1%81%D0%B0%D0%B9%D1%82.pdf

[17] Scott D.W. On optimal and data–based histograms, Biometrika, vol. 66, pp. 605–610, 1979.

[18] Taylor C. Akaike's information criterion and the histogram, Biometrika, vol. 74, pp. 636–639, 1987.

[19] Sturges H. The choice of a class–interval, J. Amer. Statist. Assoc., vol. 21, pp. 65–66, 1926.

[20] Rob J Hyndman. The problem with Sturges' rule for constructing histograms, 5 July 1995, Online available from https://robjhyndman.com/papers/sturges.pdf

[21] Freedman D. and Diaconis P. On this histogram as a density estimator: L2 theory, Zeit. Wahr. ver. Geb., vol. 57, pp.453–476, 1981.

[22] Sauerbrei S. Lorenz Curves, Size Classification, and Dimensions of Bubble Size Distributions, Entropy, vol.12, pp. 1–13, 2010.

[23] Joanna M. Landmesser, Arkadiusz J. Orłowski. Measuring and Explaining Income Inequalities in Poland: An Estimation of Lorenz Curves using Hazard Function Approach, Acta Physica Polonica, vol. 133, issue 6, pp.1445–1449, 2018.

[24] Sordo, Miguel; Navarro, Jorge; Sarabia, José Distorted Lorenz curves: models and comparisons, Social Choice & Welfare, vol. 42, issue 4, pp. 761–780, 2014.

[25] Federal state statistics service. Official statistics. Population. Standard of living. Distribution of the population by the amount of per capita monetary income. Updated 06.05.2019, urov_31g.doc. [Electronic resource]. – Mode of access: http://www.rosstat.gov.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/level/#

[26] Gorban I.I. Fundamentals of the mathematical statistics of probability theory, Mathematical Engineering, issue 9783319607795, pp.61–72, 2018.