UDC 336.774.3

GROUPING AND CLASSIFICATION FEATURES OF FINANCIAL DATA CERTAIN TYPES

D.M. Rahel
PhD in Economics, Associate Professor of the Department of Economics of the Belarusian State University of Informatics and Radioelectronics
ragel@bsuir.by

D. M. Rahel

In 2000 he graduated from the Faculty of Economics of the Belarusian State University of Informatics and Radioelectronics with a degree in Economic Informatics. In 2016 he graduated from the postgraduate course of the Academy of Management under the President of the Republic of Belarus. In 2018 he defended his PhD thesis. Research interests: data mining in marketing, process modeling, data analysis, statistical forecasting, macroeconomics.

Annotation. The article presents an approach to analyzing the distribution of various types of financial data. An approach is proposed that allows analyzing a set of similar data based on constructing a distribution series and analyzing values in different intervals of the series. The distribution of values will allow us to draw conclusions about the normality of the activities of the counterparties under study and, taking this into account, conduct further data analysis.

Keywords: data analysis, financial data, big data, data array, forecasting, analytics, distribution series, Sturges method, data distribution, financial characteristics, predictive analytics, financial analysis.

Based on the results of a sample study of 100 similar enterprises, data on the volume of their assets as of 09.01.2023 was obtained (Table 1). In order to operate with these data or draw initial conclusions regarding the business entities under consideration, it is necessary to carry out an initial classification in order to divide them into groups for further research. As the main solution for classifying the obtained data, you can use the construction of an interval variation series using the Sturges method.

Table 1. Volume of assets of 100 surveyed enterprises

Initial data as of 09.01.2023, million rubles.									
0,86	0,77	0,54	2,22	2,7	2,26	2,996	2,29	2,25	2,98
0,97	4,14	1,77	2,31	2,42	2,07	2,41	6,55	2,29	2,44
0,74	2,11	3,61	2,99	2,64	2,32	2,28	2,25	2,2	2,93
2,97	2,51	2,94	2,1	2,61	2,45	2,44	2,19	2,37	3,44
3,55	2,92	4,1	3,12	2,93	2,55	4,64	3,34	3,12	3,22
2,34	1,97	0,91	4,34	2,92	2,91	5,55	4,44	4,65	5,15
2,71	1,86	0,95	2,9	2,65	2,98	4,36	6,41	6,1	5,16
2,94	2,94	0,76	2,99	2,44	2,39	4,32	5,33	4,33	3,34
1,89	2,14	0,61	1,33	2,5	2,42	3,37	5,59	4,11	4,92
1,45	2,89	1,12	1,56	2,38	2,94	3,02	5,91	4,69	4,2

To carry out a general classification and interval distribution of the obtained values, it is necessary to understand the width of the sample values, and then divide it into intervals that will allow drawing conclusions regarding the distribution of values and, on the basis of this, making a selection for further research. To do this, for the data array under consideration, we will determine the minimum value to determine the general lower limit of the entire considered interval, in this case Xmin = 0.54 million rubles. and maximum, as the upper limit for the considered set of values: Xmax = 6.55 million rubles.

Taking this into account, the range of variation in the data of the sample under consideration will be R=6.01. For further grouping and understanding the nature of the distribution, it is necessary to divide this segment into intervals. To split and determine the number of intervals, you can use the Sturges formula (1).

$$H = R/(1+log_2 n) = R/(1+3{,}322lgn) \quad (1)$$

where n is the sample size, in our case, we are talking about 100 observations.

In the case of our sample of values, we obtain the duration of the interval H = 0.7863, which indicates 8 intervals to which the values of our sample for the assets of the companies under consideration will be assigned.

Taking into account the obtained value of the interval length, we obtain the following set of intervals according to the proposed method of partitioning and determining their number (Table 2).

Table 2. Distribution of intervals for classification of the data

Interval sequence number	Upper line	Bottom line
1	0,147	1,326
2	1,326	2,113
3	2,113	2,899
4	2,899	3,685
5	3,685	4,471
6	4,471	5,258
7	5,258	6,044
8	6,044	6,83

The population is divided into intervals until the beginning of the next one is equal to or greater than the maximum value in the population of analyzed data. Taking this into account, in our case we got 8 intervals.

After ranking the values, we will determine how the values are distributed over the previously established intervals. Taking this into account, we will construct an interval variation series of the distribution of asset volumes of the business entities under consideration based on a number of mandatory characteristics (Table 3).

Table 3. Interval variation series of distribution of volumes of enterprise assets as of 09.01.2023, million rubles.

Intervals	Frequency	Cumulative frequency	Relative frequency (frequency / n)	Relative cumulative frequency (cumulative frequency / n)	Frequency / interval duration
0,147 – 1,326	10	10	0,1	0,1	0,13
1.326 - 2.113	10	20	0,1	0,2	0,13
2.113 - 2.899	32	52	0,32	0,52	0,41
2.899 - 3.685	26	78	0,26	0,78	0,33
3.685 - 4.471	9	87	0,09	0,87	0,11
4.471 - 5.258	6	93	0,06	0,93	0,076
5.258 - 6.044	4	97	0,04	0,97	0,051
6.044 - 6.83	3	100	0,03	1	0,038
TOTAL	100		1		

Based on the obtained values, for a more visual presentation of the results, we will construct a graph that will display the distribution of enterprises by intervals (Figure 1).

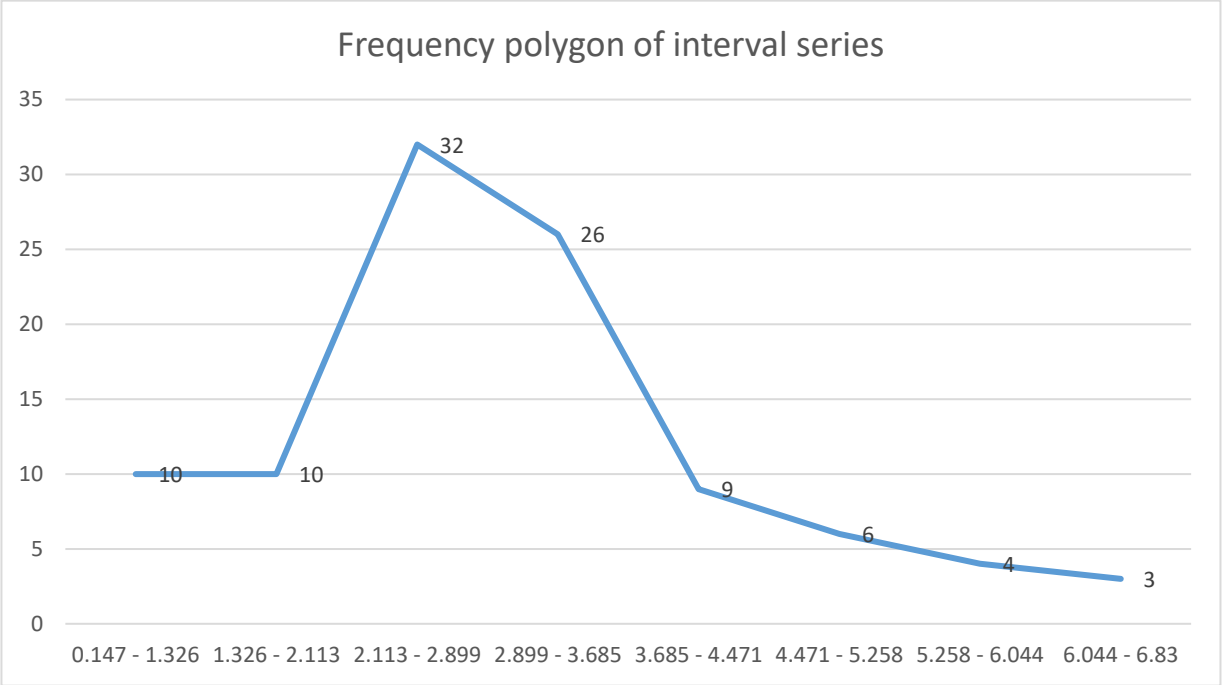

Figure 1. Final frequency distribution of the considered sample of values

Based on the results obtained, it can be noted that the largest number of enterprises under consideration belong to the average range of values, 32 enterprises in the range from 2.113 million rubles to 2.899 million rubles. and 26 enterprises in the range from 2.899 million rubles. up to 3.685 million rubles. The interval of maximum values includes three companies in the sample; the minimum interval includes 10 enterprises of the sample under consideration. 48 enterprises have

assets greater than the average value of 2.93 million rubles, which indicates a stable situation in the sample for the parameter under consideration.

References

[1] Мыльников Л.А. Статистические методы интеллектуального анализа данных. – БХВ-Петербург, 2021. – 119 с.

[2] Data Science and Big Data Analytics. A Step by Step Guide to learn Data Science from Scratch with Python Machine Learning and Big Data / Andrew Park. – Published by Andrew Park, 2021. – 124 p.

[3] Nicholson W.L. Exploring Data Analysis. – Nobel Press, 2012. – 421 p.

ОСОБЕННОСТИ ГРУППИРОВКИ И КЛАССИФИКАЦИИ НЕКОТОРЫХ ТИПОВ ФИНАНСОВЫХ ДАННЫХ

Д.М. Рагель
к.э.н., доцент кафедры экономики
Белорусского государственного
университета информатики и
радиоэлектроники

Аннотация. В статье изложен подход к анализу характера распределения различных типов финансовых данных. Предлагается подход, при котором можно проанализировать совокупность однотипных данных на основании построения ряда распределения и анализа значений в различных интервалах ряда. Характер распределения значений позволит сделать выводы о нормальности деятельности изучаемых контрагентов и, с учетом этого, провести дальнейший анализ данных.

Ключевые слова: анализ данных, финансовые данные, большие данные, массив данных, прогнозирование, аналитика, ряд распределения, правило Стерджеса, распределение данных, финансовые характеристики, прогнозная аналитика, финансовый анализ.