UDK 311.1 DOI: 10.30977/ETK.2225-2304.2024.43.144

THE METHOD OF DETERMINING THE OPTIMAL NUMBER OF GROUPS IN THE STUDY OF SOCIO-ECONOMIC PHENOMENA

SHIRINOV B., PhD in Economics, Associate Professor.
E-mail: revan622@mail.ru, ORCİD: 0000-0003-4872-4035
SALAMOVA I., Master's student.
Azerbaijan University of Architecture and Construction, Ayna Sultanova st., 11, Baku, Azerbaijan.

Abstract. *It is common knowledge that the separation of socio-economic event units into homogeneous groups by important characteristics is called grouping in statistics. Grouping is one of the most efficient methods of statistical data processing. Grouping is the most important stage in the study of large-scale social phenomena. In the socio-economic study of statistical data and average, relative, etc. it is necessary to use grouping in the calculation of final indicators. Namely, with the help of correct and convenient grouping, the complexity of social life events can be expressed and reflected with statistical quantities. Therefore, the method of grouping together with the method of summative statistical indicators is a tool for the correct understanding of socio-economic events and processes. Grouping is a manifestation of implementing analysis and synthesis, and at this time, the following issues are resolved: 1) Determination of socio-economic types; 2) Studying the structure of socio-economic events; 3) Studying the relationship between organized groups.*

In problem solving with the grouping method, the following forms of grouping are mainly used: typical, structural, analytical. By means of typical grouping, the most important statistical problem is solved: socio-economic types are determined on aggregate units, that is, separate groups qualitatively different from the aggregate are created. When grouping according to this form, the correct determination of the group sign has a special role. Based on the analysis of the content of the studied event, the basis of the grouping (that is, the main feature in the grouping) is determined. Structural groupings reflect the development of constituent parts of socio-economic events and processes or their structure according to one or another characteristic. Analytical groupings are used to study the interrelationships between phenomena and the various signs affecting them. Through such groupings, it is possible to determine the cause and effect factors influencing the development of the studied phenomenon and process. Each of the aggregate units has an individual characteristic in its development. For this reason, the absolute level of the investigated symptom is different. Those units are grouped by variation (varying) from each other according to the level of research characteristics. Such groups are called sequences of numbers. The number of units (volume) or specific weight relative to the total can be given in the series of numbers; a quality indicator that cannot be counted can also group information. According to the author, the correct determination of the optimal number of groups in the statistical study of socio-economic phenomena will give more effective results.

Key words: *grouping, number of groups, quantitative and qualitative characteristics, interval, aggregate, optimal number of groups, equal interval, unequal interval.*

Introduction. The study of socio-economic phenomena necessitates the application of robust methodologies for grouping, which are pivotal for organizing and analyzing diverse datasets. Two fundamental challenges arise in this context: the selection of group characteristics and the determination of the optimal number of groups. Any grouping in the study of socio-economic phenomena is based on two complex issues: the selection of group characteristics and the determination of the number of groups under study.

Selection of a group characteristic, i.e. gathering the studied units in one group according to the selected characteristic, is an important and complex issue of grouping theory and statistical operations. In order to form the basis of the group, the most important and decisive traits should be selected.

Moreover, determining the optimal number of groups presents another significant challenge. It is essential to strike a balance between having an adequate number of groups to capture the diversity within the dataset and avoiding excessive fragmentation that may obscure insights. Achieving this balance requires careful consideration of various factors, including the nature of the data, the research objectives, and the intended analytical methods. This task demands the application of appropriate methodologies and criteria to ensure the integrity and interpretability of the results.

The crux of the issue lies in devising effective strategies for both selecting group characteristics and determining the number of groups, thereby facilitating robust and insightful analyses of socio-economic phenomena. By addressing these issues effectively, researchers can unlock valuable insights into the complex dynamics of society and the economy.

Analysis of recent researches and publications. The study of socio-economic phenomena involves complex processes of grouping, primarily focusing on two fundamental aspects: selecting group characteristics and determining the number of groups. Understanding these processes requires a thorough analysis of existing literature and methodologies.

The selection of group characteristics involves identifying key traits that effectively categorize units under study. This process is crucial for ensuring the relevance and accuracy of groupings. Various studies emphasize the importance of correctly determining thresholds for qualitative characteristics to assign units to appropriate groups based on factors such as social status, gender, qualification, and education [1, 2].

For instance, statistical indicators published by national agencies often employ grouping based on demographic factors like age to provide insights into population distribution [3].

Moreover, the literature highlights the significance of considering fluctuation in quantitative characteristics when structuring groups. Factors such as the nature of the studied phenomenon and the intended analysis dictate the choice between equal or unequal intervals in grouping [5]. While equal intervals are suitable for uniform distributions, unequal intervals, such as gradually increasing or decreasing intervals, offer better representation of economic events with varying significance across different scales [6].

Determining the optimal number of groups is essential for characterizing qualitative diversity within a population. Studies explore various methodologies for this purpose, including the use of logarithmic formulas and root mean square distance calculations [7]. Stercess' formula, which considers the logarithm of the total population size, is commonly employed to determine the number of groups, with adjustments made based on the nature of the data and distribution [6].

Additionally, grouping based on multiple characteristics, known as combined grouping, presents challenges in terms of increasing complexity and interpretation. While age grouping is deemed significant in analyzing socio-economic phenomena, the literature cautions against overwhelming analyses with too many grouping features, as clarity and interpretation become more challenging [4].

Practical applications of grouping methodologies are evident in various sectors, including industry analysis and regional development planning. Studies showcase how grouping techniques facilitate the analysis of economic trends, forecast sales volumes, and inform policy decisions [6-8]. Furthermore, recommendations emphasize the need for a balanced approach in determining group characteristics and sizes, considering both statistical relevance and practical interpretability [3].

The literature provides valuable insights into the complexities of grouping in socio-economic studies. By synthesizing existing methodologies and practical applications, this article aims to contribute to the development of robust techniques for optimizing group selection and determining the number of groups, thereby enhancing the effectiveness of statistical analyses in understanding socio-economic phenomena.

Group signs can be quantitative or qualitative signs.

When grouping according to quality characteristics (including: social status, gender, qualification, education, etc.), it is necessary to correctly determine the thresholds of the selected factor so that it is known to which group the grouped units belong.

For example, in the set of statistical indicators published by the State Statistics Committee of Azerbaijan every year, the following grouping of the working-age population is given: total population, including the number of people under working age, working age and over working age.

The fluctuation of the quantitative sign used as the basis for grouping should be taken into account. Because the number of groups of the same gender depends on the fluctuation of the sign and the number of units included in the observation. Each group should reflect the typical character of the units included in the group.

When obtaining groups that most adequately reflect reality, it is necessary to be guided by the essence of the studied phenomenon. In this case, the intervals may be uneven. Thus, gradually increasing or decreasing intervals can be used when studying economic events. For example, according to the number of employees, construction enterprises can be divided into the following groups: up to 50 people, 20-50-100, 100 and more people. This is explained by the fact that quantitative changes in the size of the attribute have different meanings in the lower and upper groups according to the size of the attribute: a change in the number of employees of 25 people is significant for small construction enterprises, but not for large ones. Thus, the size of the interval is determined in advance based on the tasks and nature of the event.

Groupings with equal intervals are used when the degree of variation is insignificant and the distribution is almost uniform (for example, when grouping workers of the same profession by wages, the output of a certain product by productivity). For groups with equal intervals, the value of the interval is calculated using a certain formula.

Thus, in all cases, groups should be structured so that the groups they form correspond as completely as possible to reality.

Intervals between groups should be determined in the grouping according to the quantitative sign. They can be equal or unequal. Unequal intervals gradually increase and decrease in turn. Group intervals can be open or closed. Equally spaced groups are used when the sign change within the set is of equal size. At this time, the interval limit is calculated by the following formula:

$$d = \frac{R}{n}$$

$$R = x_{\max} - x_{\min}$$

$$d = \frac{x_{\max} - x_{\min}}{n}$$

Where,

d - stands for the amount of the intervals;

$x_{\max}$ - for the high value of the sign;

$x_{\min}$ - for the low value of the sign;

R - means the difference between the maximum and the minimum value of the sign;

n - shows the number of groups.

The formulation of the objectives of the article. The objective of this article is to develop methods for determining the optimal number of groups and grouping characteristics in the study of socio-economic phenomena. The article aims to propose methodologies for selecting grouping characteristics and determining the number of groups, taking into account various aspects of socio-economic data. The primary focus is on the process of selecting the optimal number of groups, including an analysis of various methods for determining the number of groups, their advantages and disadvantages. Additionally, the article seeks to develop practical recommendations for organizing grouping, ensuring the most effective use of statistical methods in analyzing socio-economic phenomena.

Statement of the main material of the research. In the study of socio-economic phenomena, the classification of data into appropriate groups is paramount for insightful analysis. The optimal number of groups serves as a fundamental parameter, delineating qualitative diversity within a population. This section elucidates the methodologies employed to ascertain the ideal number of groups, paving the way for subsequent analytical computations.

Determınıng the optımal number of groups.

The optimal number of groups is chosen to characterize the qualitative diversity in the total population. In many cases, the optimal number of groups is determined by the following formula developed by the American statistician Stercess:

$$n = 1 + 3.32 \lg N \qquad \text{or,}$$

$$n = 1.44 \ln N + 1$$

Where, n is the number of groups,
N – the number of units in the set,
lg – decimal logarithm,
ln – indicates the natural logarithm.

Suppose that the maximum amount of annual expenses of 50 construction companies was 300,000 manats and the minimum amount was 90,000 manats (figures are conditional). Based on the given information, it is necessary to select the optimal number of groups.

To determine the number of groups, since the number of aggregate units in our example, that is, the number of construction companies, is 50: (using the logarithmic table) since the decimal logarithm is equal to lg 50 = 1.699,

$n = 1 + 3.32 \lg N = 1+3.32 \lg 50 = 1 + 3.32*1{,}699 = 1 + 5.64 = 6.64$ or (using the logarithmic table) since the natural logarithm is ln 50 = 3.912,
$n = 1.44 \ln N + 1 = 1.44 \ln 50 + 1 = 1.44*3.912 + 1= 6.63$.

In both cases, the number of groups in the sample can be a whole number of 6 or 7.

Let's organize the groups by taking the number of groups in this example (6.63 and 6.64 are rounded closer to 7) by 7. First we need to find the discontinuity quantity (d).

$$d = \frac{R}{n} = \frac{300-90}{7} = 30.$$

Then we make the 1st group by adding 30 to the minimum (low) value of the sign, i.e. 90. To form the 2nd group, the upper boundary of the 1st group (120) is brought to the lower boundary of the 2nd group and a break quantity of 30 is applied to it, to form the 3rd group, the upper boundary of the 2nd group (150) is brought to the 3rd group is brought to its lower limit and a break quantity of 30 is applied to it, to make the 4th group the upper limit of the 3rd group (180) is brought to the lower limit of the 4th group and a break amount of 30 is applied to it, to make the 5th group the 4th the upper border of the group (210) is brought to the lower border of the 5th group and a break amount of 30 is added to it, to form the 6th group, the upper border of the 5th group (240) is brought to the lower border of the 6th group and a break amount of 30 is applied to it is reached, to form the 7th group, the upper limit of the 6th group (270) is brought to the lower limit of the 7th group, and the break quantity 30 is applied to it, and the following seven groups are formed.

I-group 90 - 120 (90 + 30)
II-group 120 - 150 (120 + 30)
III-group 150 - 180 (150 + 30)
IV-group 180 - 210 (180 + 30)
V-group 210 - 240 (210 + 30)
VI-group 240 - 270 (240 + 30)
VII-group 270 - 300 (270 + 30)

If the resulting group does not meet the requirements of the analysis, then the group can be rearranged. Efforts should not be made to form a large number of groups, since in such cases the differences between the groups are negligible.

If the result is not an integer, it is usually rounded. This formula is applied only when the distribution of population units is close to normal due to a certain characteristic and when uniform intervals are used in groups.

Methods used to determıne the number of groups.

One of the methods used to determine the number of groups is the use of the root mean square distance (σ). At this time, it is assumed that the range of change of the indicator is equal. The total is divided into 12 groups when the interval quantity is equal to 0.5 σ, 9 groups when it is equal to (2/3) σ, and 6 groups when it is equal to σ.

If the fluctuation of sign between aggregate units is too large, grouping with uneven intervals should be used (uneven intervals are mainly used for quantities that increase very rapidly). The use of such intervals is related to the fact that the small difference between the indicators in the lower groups is of great importance, while in the other groups this difference is not significant.

If the size of the sign varies greatly, then in the grouping, intervals corresponding to the logarithm of the size of the sign can be accepted.

When defining groups, the quantity of the boundary of each group must be specified. If the group sign is discrete, the boundary of the (intersecting) groups must be specified.

It is relatively easy to organize grouping on discrete traits. For example, in the grouping of workers according to tariff rates, as many groups as the number of tariff rates (5-6) are formed.

For example: The given information on the distribution of workers in the enterprise by salary (table 1), (numbers are conditional) can be shown as an example of grouping by discrete characteristics.

Table 1

Grouping by discrete trait

Wages of workers according to tariff rates	number of workers (people)
I	20
II	25
III	35
IV	40
V	15
Total	135

Source: The table was compiled by the author.

If the grouping is done by one characteristic, such grouping is called simple grouping (table 2) (numbers are conditional).

Table 2

2018-2022 average annual number of employees working in the industry of the Republic, in thousand people

Years	The average annual list of employees, thousand people
2018	101,9
2019	98,1
2020	96,7
2021	104,0
2022	108,5

Source: The table was compiled by the author.

From the information in Table 2, it is clear that the average annual number of employees working in the industry of the republic increased by 6.6 thousand people in 2022 compared to 2018.

Grouping based on two or more characteristics is called combined grouping (table 3) (numbers are conditional).

Table 3

The average annual list number of employees working in the industry of the Republic in 2018-2022 and the average monthly nominal salary

Years	The average annual list of employees, thousand people	Average monthly nominal salary, manat
2018	101,9	626,9
2019	98,1	677,7
2020	96,7	812,9
2021	104,0	783,3
2022	108,5	698,7

Source: The table was compiled by the author.

In the analysis of complex socio-economic phenomena, age grouping is of great importance. At the same time, it should be taken into account that the number of groups increases rapidly due to the increase in the number of grouping features in cluster grouping. Therefore, taking more than two or three grouping signs in complex grouping can complicate the analysis of the result, make their clear reading and understanding difficult.

Based on the data of Table 4, let's calculate by applying the grouping method (numbers are conditional):

1) number of groups;

2) groups of the size of insurance companies according to the average annual volume of the charter capital;

3) the number of insurance companies in the groups and the specific weight of the groups.

Solution:

1) To find the number of groups;

(using a logarithmic table) decimal logarithm Since lg 15 = 1.176,

$n = 1 + 3.32 \lg N = 1 + 3.32 \lg 15 = 1 + 3.9 = 4.9$ in total.

will be or (using a logarithmic table) the natural logarithm

Since ln 15 = 2.708,

$n = 1.44 \ln 15 + 1 = 1.44 * 2.708 + 1 = 4.9$ in total.

In both cases, the number of groups in the sample (4,9) can be 4 or 5. So the number of groups will be 4 or 5.

Table 4

The average annual volume of authorized capital of insurance companies

Insurance company serial number	The average annual volume of authorized capital of insurance companies is mln. man.
1	15,5
2	15,9
3	16,7
4	16,1
5	15,2
6	16,8
7	10,6
8	18,9
9	15,1
10	19,0
11	14,8
12	17,5
13	18,6
14	14,6
15	15,6
Total	**240,9**

Source: The table was compiled by the author.

2) Let's determine the groups of the size of insurance companies according to the average annual volume of the charter capital;

We take the average annual volume of the authorized capital of the company as a group indicator. According to this sign, in order to divide the average annual volume of the authorized capital of the company into 5 groups with equal intervals (as 4.9 is rounded closer to 5), let's determine the volume of the interval with the following formula:

$$d = \frac{x_{\max} - x_{\min}}{n}$$

$$d = \frac{19 - 10{,}6}{5} = 1{,}68$$

So, the volume (quantity) of the gap between the groups is 1.68 mln. manat.

Let's make groups of insurance companies according to the volume (quantity) level of the break of the average annual volume of the authorized capital of insurance companies. Let's organize the following groups based on that break quantity:

Group I 10.6 – 12.28 (10.6 + 1.68 = 12.28)
Group II 12.28 – 13.96 (12.28 + 1.68 = 13.96)
Group III 13.96 – 15.64 (13.96 + 1.68 = 15.64)
Group IV 15.64 – 17.32 (15.64 + 1.68 = 17.32)
Group V 17.32 – 19 (17.32 + 1.68 = 19)

3) Let's determine the number of insurance companies in the groups and the specific weight of the groups.

We compile the groups of the volume of insurance companies, the number of organizations in the groups and the specific weight of the groups (table 5) according to the average annual volume of the authorized capital:

Table 5

Size groups of insurance companies according to the average annual volume of authorized capital

Groups of the average annual volume of the authorized capital of insurance companies, mln. man.	Number of insurance companies	Specific weight of insurance companies in the relevant group, in %
I qroup 10.6 – 12,28	1	6.6
II qroup 12,28– 13.96	-	0.0
III qroup 13.96 – 15.64	6	40,0
IV qroup 15.64 – 17.32	4	26.7
V qroup 17.32 – 19	4	26.7
Total	15	100 %

Source: The table was compiled by the author.

As can be seen from the data of Table 5, the specific weight of the average annual volume of the authorized capital of insurance companies is higher in the III group (40 %).

Conclusion. The number of groups and the size of the interval are determined based on the purpose of the study, the value of the studied characteristic and many other factors. The number of groups and the size of the interval are related: the larger the number of groups, the smaller the interval, and vice versa. The number of groups depends on the number of units in the studied population and the level of variation of the grouping characteristic.

The smaller the volume of the aggregate, the smaller the number of formed groups. Since the number of groups will be small, it is not possible to create a large number of groups with a small total number. The indicators calculated for such groups will not correctly characterize the studied population. The number of groups should meet the requirement of the large numbers law and should be optimal, i.e. each group should include a sufficient number of population units. But in some cases, small groups are also interesting. For example, new, advanced phenomena that are not yet widespread appear in a small number of facts, but require detailed study. One of the tasks of statistics is to identify and study these facts. Thus, the issue of the number of units in the groups cannot be formally approached, it is necessary to know the essence of the studied phenomenon.

References

1. Shirinov, B. H., & Mammadova, F. A. (2022). Statistics (for Bachelor level) Textbook. Baku.

2. Huseynova, A. M. (2008). General theory of statistics [Teaching materials]. Baku.

3. Yagubov, S. M., & Mammadov, A. C. (2010). Social-Economic Statistics Textbook. Baku.

4. Yusifov, E., Sarkarli, A., Byba, V., & Kushnirova, T. (2020). Analysis of the Role and Place of the Building Materials Industry in the Development of Azerbaijan's Economy. In *ICBI 2020: Proceedings of the 3rd International Conference on Building Innovations*, 809–819.

5. Ruhangiz, A. T., & Farid, A. Z. (2023). Main directions of accelerating regional socio-economic development in the digital era. *Agora International Journal of Economical Sciences*, 17(2), 170-180.

6. Yusifov, E. M. (2021). Level of implementation and development directions of innovations in the field of construction and investment in Azerbaijan. In *Building Innovations– 2021:* IV International Ukrainian-Azerbaijani Scientific-Practical Conference, National University "Poltava Polytechnic", (May 20–21, Poltava, Ukraine), 323-327.

7. Shirinov, B., & Nataliia-Mahas. (2022). Forecasting Sales Volume in Construction Companies. In *ICBI 2022: Proceedings of the 4th International Conference on Building Innovations*, 825–830.

8. Shirinov, B. H. (2020). The role of sales forecasting in organizing the production process of enterprises. *Azerbaijan University of Architecture and Construction Economics and Management Scientific-Practical Journal of Construction*, Baku.

МЕТОД ВИЗНАЧЕННЯ ОПТИМАЛЬНОЇ КІЛЬКОСТІ ГРУП У ДОСЛІДЖЕННІ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ЯВИЩ

ШІРІНОВ Б. Г., кандидат економічних наук, доцент.
E-mail: revan622@mail.ru, ORCİD 0000-0003-4872-4035
САЛАМОВА І. І., студентка магістратури.
Університет архітектури та будівництва Азербайджану, вул. Айна Султанова, 11, м. Баку, Азербайджан.

Анотація. *Як відомо, поділ одиниць соціально-економічних подій на однорідні групи за важливими ознаками в статистиці називають групуванням. Групування є одним із найефективніших методів статистичної обробки даних. Групування є найважливішим етапом дослідження масштабних суспільних явищ.*

При соціально-економічному вивченні статистичних даних і середніх, відносних тощо необхідно застосовувати групування при розрахунку підсумкових показників, а саме: за допомогою правильного та зручного групування складність подій соціального життя може бути виражена та відображена статистичними величинами. Тому метод групування разом із методом підсумовуючих статистичних показників є інструментом правильного розуміння соціально-економічних подій і процесів.

Групування є проявом реалізації аналізу та синтезу, і при цьому вирішуються такі питання:

1) визначення соціально-економічних типів;

2) вивчення структури соціально-економічних подій;

3) вивчення взаємовідносин між організованими групами.

При розв'язуванні задач методом групування в основному використовуються такі форми групування: типове, структурне, аналітичне.

За допомогою типового групування вирішується найважливіша статистична задача: на агрегатних одиницях визначаються соціально-економічні типи, тобто створюються окремі групи, якісно відмінні від сукупності. При групуванні за цією формою особливу роль відіграє правильне визначення ознаки групи. На основі аналізу змісту досліджуваної події визначається основа угруповання (тобто основна ознака в угрупованні).

Структурні угруповання відображають розвиток складових частин соціально-економічних явищ і процесів або їх структуру за тією чи іншою ознакою.

Аналітичні групування використовуються для вивчення взаємозв'язків між явищами і різними ознаками, що впливають на них. За допомогою таких групувань можна визначити причинно-наслідкові фактори, що впливають на розвиток досліджуваного явища та процесу.

Кожна з агрегатних одиниць має індивідуальну характеристику у своєму розвитку. З цієї причини абсолютний рівень досліджуваного симптому різний. Ці підрозділи згруповані за варіацією (відмінністю) один від одного відповідно до рівня досліджуваних характеристик. Такі групи називаються послідовностями чисел. Кількість одиниць (об'єм) або питома вага відносно загальної кількості може бути наведена в ряду чисел; також можна згрупувати за показником якості, який не піддається підрахунку.

На думку автора, правильне визначення оптимальної кількості груп при статистичному вивченні соціально-економічних явищ дасть більш ефективні результати.

Ключові слова: *групування, кількість груп, кількісна та якісна характеристика, інтервал, сукупність, оптимальна кількість груп, рівний інтервал, нерівний інтервал.*