На рис. 2 построена гистограмма распределения объектов по группам по данным примера 1. По оси абсцисс расположены значения изучаемого признака, по оси ординат отношение частот к длинам интервалов, т. е. плотность распределения единиц.

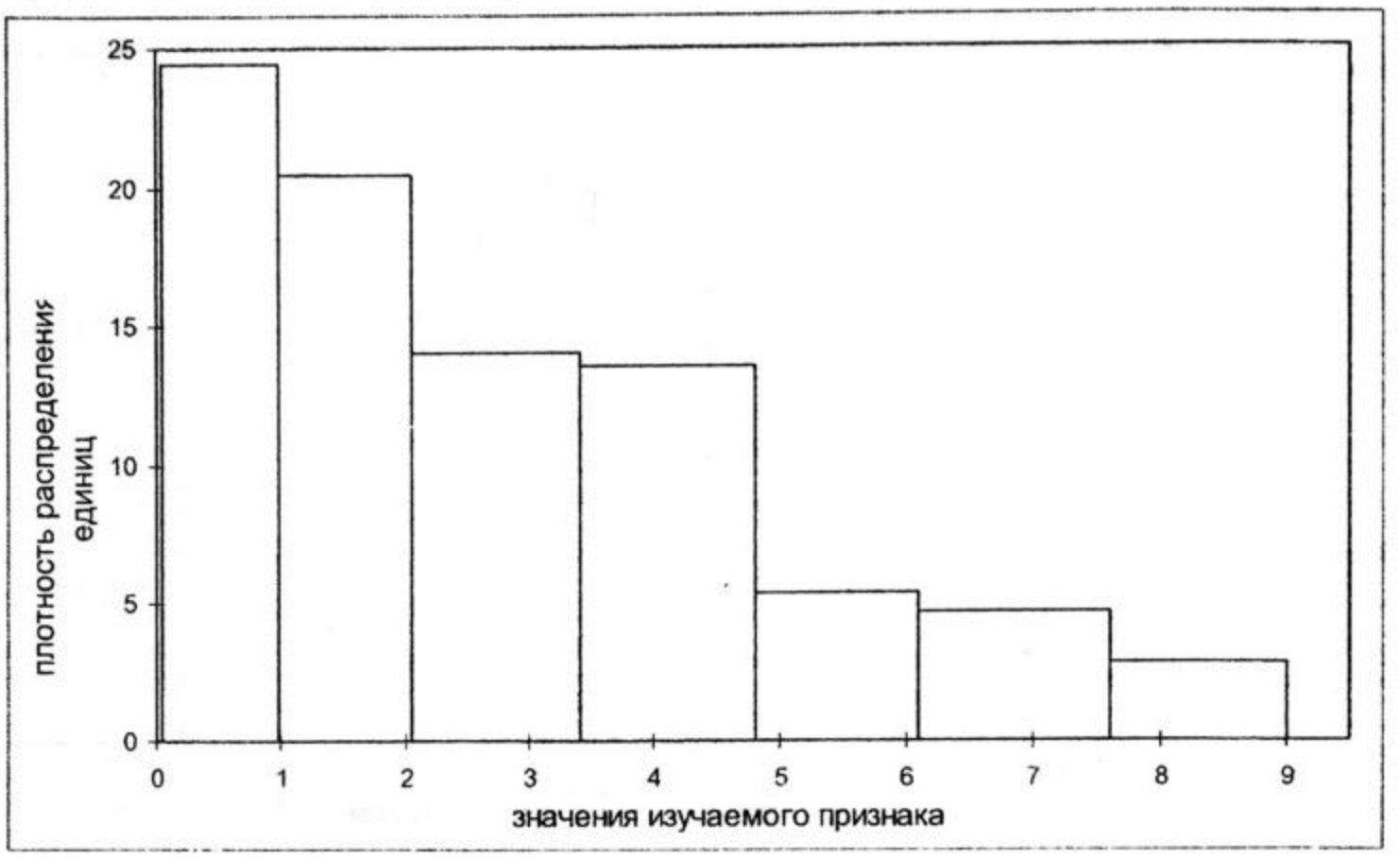

Рис. 2. Гистограмма распределения изучаемого признака (пример 2).

Из рисунка 2 также следует вывод о том, что чем больше плотность единиц изучаемого признака, тем меньше должна быть длина интервала. При этом межгрупповая дисперсия принимает максимальное значение.

Очевидно, что рассмотренную в работе модификацию метода перебора при распределении n объектов по m группам, можно применять и в случае значительно большего числа объектов.

ЛИТЕРАТУРА

1. Бардасов С. А. К вопросу о методе группировок в экономической статистике // Вестник Тюменского государственного университета. 2000, № 1. С. 141-146.

2. Ефимова М. Р., Петрова Е. В., Румянцев В. Н. Общая теория статистики. М.: ИНФРА-М, 1996. 416 с.

3. Елисеева И. И., Юзбашев М. М. Общая теория статистики. М.: Финансы и статистика, 1996. 368 с.

II. ПОСТРОЕНИЕ АНАЛИТИЧЕСКОЙ ГРУППИРОВКИ МЕТОДОМ МАКСИМИЗАЦИИ МЕЖГРУППОВОЙ ДИСПЕРСИИ РЕЗУЛЬТАТИВНОГО ПРИЗНАКА

АННОТАЦИЯ. Рассматривается аналитическая группировка, когда группы образуются по максимальному значению межгрупповой дисперсии результативного признака.

The grouping is considered, when the groups are formed according to the maximal value of intergroup dispersion of a resultative attribute.

Пусть имеется n объектов, которые характеризуются численными значениями факторного признака $x_1, x_2, \ldots, x_n$, которым соответствуют значения $y_1, y_2, \ldots, y_n$ результативного признака. Метод аналитических группировок является одним из традиционных приемов статистического изучения связей при изучении социально-экономических явлений. Взаимосвязь проявляется в том, что с возрастанием значения факторного признака x, который положен в основу группировки, систематически возрастают или убывают групповые средние значения результативного признака $\bar{y}_j$ (j - номер группы, $j = 1,2,\ldots,m$). Отметим, что в работе рассматривается случай, когда признаки имеют количественное выражение. Как отмечается в [1], выбор числа групп и границ интервалов — центральная проблема, так как этим обусловливается объективность характеристик связи. Поэтому вопрос о числе групп, варианте группировки — это вопрос о доказательности метода аналитических группировок. В решении этого вопроса исходят из противоречивых требований. С одной стороны, интервалы должны быть значительны по своим размерам, т. е. такими, чтобы изменения в величине фактора x обусловили существенное различие в величинах $\bar{y}_j$. С другой стороны, величина интервала не должна быть чрезвычайно большой, так как в каждой группе единицы совокупности должны быть однородными по x_{ij} (i - номер единицы совокупности, $i = 1,\ldots,n$; j – номер группы по признаку x, $j = 1,2,\ldots,m$). К тому же различия $\bar{y}_j$ можно объяснить влиянием x только при достаточно большом количестве элементов в группах n_j, когда в значениях $\bar{y}_j$ уравновешивается действие прочих факторов.

В работе [1] указывается, что влияние степени дробления совокупности на заключения о характере связи и на ее количественные определения приводит к выводу о необходимости выполнения группировки в нескольких вариантах. Следовательно, алгоритм построения аналитической группировки должен основываться на переборе вариантов группировки по числу групп (от максимального, при котором все $n_j = 1$, до минимального, $m = 3$) при разных способах образования групп (с равными интервалами или равнонаполненных). Отметим, что это обстоятельство обычно не упоминается в учебной литературе. В [1] также приведены примеры, когда при делении совокупности на различное число групп можно по результатам группировки сделать различные выводы о виде связи между признаками.

Таблица 1

Значения факторного признака x и соответствующие им значения результативного признака y. Данные условные

x	y	x	y	x	y	x	y	x	y
1,00	0,73	3,00	3,47	3,66	3,67	4,50	4,16	5,64	4,01
1,20	1,00	3,03	3,34	3,70	3,53	4,55	3,92	5,71	4,34
1,40	1,56	3,06	3,22	3,73	3,57	4,60	3,92	5,77	4,14
1,60	1,95	3,10	3,24	3,76	3,74	4,65	4,06	5,84	3,92
1,80	2,06	3,13	3,40	3,80	3,99	4,70	4,08	5,90	4,06
2,00	2,57	3,16	3,61	3,83	4,07	4,75	4,04	6,00	4,03
2,06	2,42	3,20	3,77	3,86	4,02	4,80	4,14	6,10	4,20
2,13	2,64	3,23	3,72	3,90	3,83	4,85	4,24	6,20	3,81
2,19	2,73	3,26	3,59	3,94	3,72	4,90	4,10	6,30	4,10
2,26	2,79	3,30	3,44	3,98	3,77	4,99	3,90	6,40	4,02
2,32	2,99	3,33	3,44	4,00	3,84	5,00	3,94	6,50	3,85
2,39	2,90	3,37	3,56	4,05	3,94	5,06	4,29	6,60	3,91

Продолжение табл. 1

x	y	x	y	x	y	x	y	x	y
2,45	2,72	3,40	3,64	4,10	3,90	5,12	4,28	6,70	3,99
2,52	3,07	3,43	3,67	4,15	3,95	5,19	3,97	6,80	3,77
2,58	3,33	3,46	3,66	4,20	4,09	5,25	4,04	6,90	3,74
2,64	3,12	3,50	3,66	4,25	4,07	5,32	4,14	7,00	3,96
2,71	3,04	3,53	3,72	4,30	3,84	5,38	4,08	7,20	3,72
2,77	3,24	3,56	3,82	4,35	3,76	5,45	4,23	7,50	3,33
2,84	3,27	3,60	3,88	4,40	4,02	5,51	4,20	7,60	3,74
2,90	3,35	3,63	3,81	4,45	4,26	5,58	3,87	7,90	3,30

Все вышеизложенные проблемы метода аналитической группировки могут быть связаны с тем, что группы образуются или с равными интервалами или равнонаполненные. Так как при таком делении на группы не учитывается характер связи между признаками.

Будем делить изучаемую совокупность на заданное число групп таким образом, чтобы длины групповых интервалов по возможности соответствовали характеру связи между признаками. Используем для этого эмпирическое корреляционное отношение.

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}},$$

где σ^2 — общая дисперсия результативного признака, δ^2 — межгрупповая дисперсия результативного признака. В соответствии с этим группы будем образовывать по величине факторного признака, но таким образом, чтобы при этом величина межгрупповой дисперсии была максимальной. Очевидно, что в общем случае длины групповых интервалов будут различны и группы не будут равнонаполненными. Лучшее разбиение, которое дает максимальное значение межгрупповой дисперсии результативного признака, получаем методом перебора элементов. Образуем всевозможные разбиения на группы по факторному признаку. Для каждого разбиения рассчитываем межгрупповую дисперсию результативного признака. Выбираем то разбиение, которое соответствует максимальному значению результативного признака. Для расчетов использовался персональный компьютер, программа написана на языке Turbo Pascal 7. 0. Отметим, что с увеличением числа групп и числа объектов быстро растет число возможных вариантов группировки и соответственно растет время счета. Таким образом, группировка методом полного перебора в общем случае неосуществима. Поэтому была разработана специальная процедура перебора элементов, которая подробно изложена в статье I.

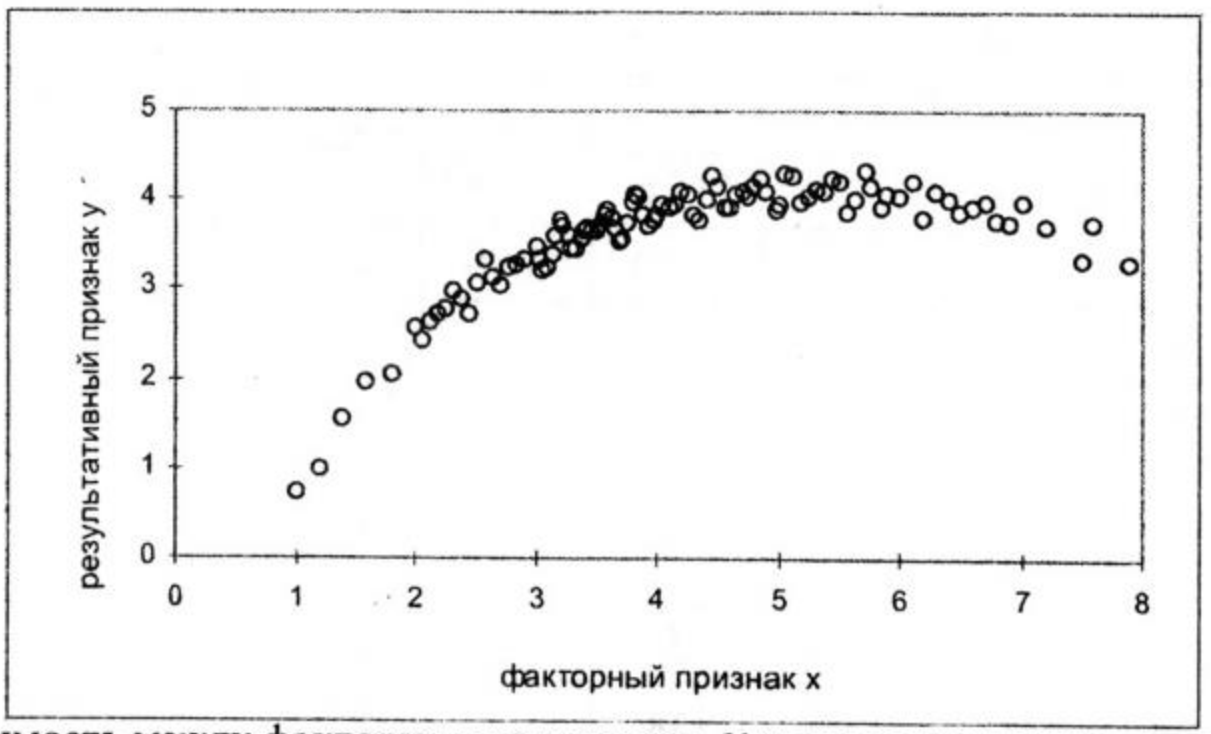

Рис. 1. Зависимость между факторным признаком x и результативным признаком y. Данные условные.

Для расчетов было взято 100 значений факторного признака, которым соответствовало 100 значений результативного признака. Исходные данные приведены в табл. 1 и на рис. 1. Как видно из рис. 1 между признаками имеется нелинейная зависимость

Первоначально разобьем изучаемую совокупность, например, на 4 группы. То есть возьмем заведомо заниженное значение для числа групп. Для сравнения с методом, рассматриваемым в данной работе, будем использовать метод равных интервалов.

Результаты группировки приведены в табл. 2 и 3.

Таблица 2

Результаты группировки по факторному признаку (4 группы) при максимуме межгрупповой дисперсии результативного признака

x	n_j	$\overline{x}_j$	$\overline{y}_j$
1,0-1,5	3	1,20	1,10
1,5-2,48	10	2,08	2,56
2,48-3,52	23	3,06	3,40
3,52-7,9	64	5,09	3,94
Итого	100	4,21	3,60

Таблица 3

Результаты группировки по факторному признаку (4 группы) при равенстве групповых интервалов

x	n_j	$\overline{x}_j$	$\overline{y}_j$
1,00-2,73	17	2,07	2,45
2,73-4,45	43	3,61	3,70
4,45-6,18	27	5,23	4,09
6,18-7,90	13	6,89	3,79
Итого	100	4,21	3,60

Групповые средние значения факторного и результативного признаков приведены на рис. 2. Кружки соответствуют группировке по максимуму межгрупповой дисперсии результативного признака, а треугольники — равным интервалам.

По результатам группировки для данного примера можно сделать следующие выводы. Группировка методом максимума межгрупповой дисперсии результативного признака приводит к тому, что чем быстрее изменяется результативный признак под влиянием факторного признака, тем меньше длины соответствующих групповых интервалов. Следовательно, этот метод дает более подробную характеристику, когда зависимость между признаками выступает более отчетливо. И, наоборот, в случае небольшого числа групп метод равных интервалов подробнее характеризует связь между признаками, когда с ростом факторного признака результативный признак изменяется незначительно.

По-видимому следует считать, что влияние факторов, которые не учитываются, проявляется тем значительнее, чем меньше изменение результативного признака под влиянием факторного. Группы по методу максимума межгрупповой дисперсии образуются так, что чем быстрее изменение результативного признака, тем меньше элементов в группе. И, наоборот, чем медленнее изменяется результативный признак, тем больше элементов в группе. Можно предположить, что данный метод позволяет автоматически уменьшить влияние других (не известных или не рассматриваемых) факторов.

Рис. 2. Зависимость групповых средних значений результативного признака от групповых средних значений факторного признака (4 группы).
Ряд 1 — группировка при максимуме межгрупповой дисперсии результативного признака,
ряд 2 — группировка при равенстве групповых интервалов.

Разобьем теперь изучаемую совокупность на семь групп. Результаты группировки приведены в табл. 4 и 5.

Таблица 4

Результаты группировки по факторному признаку (7 групп) при максимуме межгрупповой дисперсии результативного признака

x	n_j	$\overline{x}_j$	$\overline{y}_j$
1,00-1,30	2	1,10	0,86
1,30-1,90	3	1,60	1,86
1,90-2,55	8	2,23	2,72
2,55-2,15	12	2,86	3,26
2,15-3,78	19	3,46	3,66
3,78-7,10	52	5,12	4,01
7,10-7,90	4	7,55	3,52
Итого	100	4,21	3,60

Таблица 5

Результаты группировки по факторному признаку (7 групп) при равенстве групповых интервалов

x	n_j	$\overline{x}_j$	$\overline{y}_j$
1,00-1,99	5	1,40	1,46
1,99-2,97	15	2,45	2,95
2,97-3,96	29	3,46	3,65
3,96-4,94	20	4,43	4,00
4,94-5,93	16	5,42	4,09
5,93-6,91	10	6,45	3,94
6,91-7,90	5	7,44	3,61
Итого	100	4,21	3,60

Групповые средние значения факторного и результативного признаков в этом случае приведены на рис. 3. Кружки соответствуют группировке по максимуму межгрупповой дисперсии результативного признака, а треугольники — равным интервалам.

В этом случае подтверждается та же тенденция, что и при делении на четыре группы. Отметим, что для данного примера при делении совокупности на семь групп оба метода дают достаточно хорошие результаты.

Рис. 3. Зависимость групповых средних значений результативного признака от групповых средних значений факторного признака.
Ряд 1 — группировка при максимуме межгрупповой дисперсии результативного признака,
ряд 2 — группировка при равенстве групповых интервалов.

Для дальнейшего сравнения методов группировки по максимуму межгрупповой дисперсии и равными интервалами предполагается рассмотреть случай, когда значения факторного признака сильно несимметричны относительно среднего значения. Так как именно в этом случае образование равных интервалов по факторному признаку может приводить к ошибочным выводам о характере связи между признаками.

ЛИТЕРАТУРА

1. Елисеева И. И., Рукавишников В. О. Группировка, корреляция, распознавание образов. М.: Статистика, 1977, 144 с.

2. Бардасов С. А. Группировка методом скользящего перебора. //Сб. Экономико- управленческие аспекты деятельности предприятий Тюменского региона, изд-во Санкт-Петербургского университета экономики и финансов. 2000. С. 22-27.