ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение высшего профессионального образования «Уральский государственный университет им. А.М. Горького»

ИОНЦ «Бизнес информатика»

Институт управления и предпринимательства

Статистические методы анализа рынков

Учебное пособие

Подпись руководителя ИОНЦ
Дата

Екатеринбург
2008

Печатается по решению Ученого совета
Института управления и предпринимательства
Уральского государственного университета
им. А. М. Горького от 2008 г.

Рецензенты:

Кафедра мультимедиатехнологий УГТУ – УПИ, зав. кафедрой доктор физ.-мат. наук, профессор А. Н. Красовский.

Ведущий научный сотрудник Института математики и механики УрО РАН, доктор физ.-мат. наук, профессор В. Д. Мазуров.

Близоруков М. Г.

Б40 Статистические методы анализа рынка: Учебно-метод. пособие / Близоруков М. Г. – Екатеринбург: Ин-т управления и предпринимательства Урал. гос. ун-та, 2008. – 75 с.

ISBN

В учебно-методическом пособии рассматриваются некоторые методы многомерного анализа – наиболее действенного количественного инструмента исследования социально-экономических явлений, описываемых большим числом характеристик. Приведены математические основания кластерного, факторного, дисперсионного анализа, метода ранжирования многомерных величин и их компьютерная реализация в рамках известных пакетов прикладных программ (в первую очередь в SPSS, «Statistica» и «Stadia»).

Издание адресовано студентам управленческих специальностей, в том числе бизнес – информатика, прикладная математика в экономике, математические методы в экономике, специалистам в области экономики и управления, слушателям программы MBA, а также может оказать помощь менеджерам, использующим ERP системы, обрабатывающим статистическую информации.

ISBN

ОГЛАВЛЕНИЕ

Введение…… 4
Глава 1. Метод многомерной средней…… 5
1.1. Исходные данные. Постановка задачи…….... 5
1.2. Алгоритм реализации…….... 6
1.3. Реализация метода многомерной средней средствами «Excel»..……... 7
Глава 2. Дисперсионный анализ……..... 9
2.1. Однофакторный дисперсионный анализ. Постановка задачи…….…… 9
2.2. Алгоритм реализации……... 10
2.3. Однофакторный дисперсионный анализ средствами «Excel»…… 12
2.4. Реализация однофакторного дисперсионного анализа в пакете «Stadia»..……..... 15
2.5. «ANOVA» – реализация дисперсионного анализа в пакете «Statistica»……..... 18
2.6. Особенности реализации дисперсионного анализа в пакете SPSS……..... 20
2.7. Многофакторный дисперсионный анализ…… 21
2.8. Одна задача применения двухфакторного дисперсионного анализа…….. 23
2.9. Компьютерная реализация двухфакторного дисперсионного анализа……... 24
Глава 3. Кластерный анализ……..…….. 30
3.1. Постановка задачи……... 31
3.2. Подготовка данных для кластеризации…….…… 33
3.3. Иерархический кластерный анализ. Алгоритм последовательной кластеризации… 34
3.4. Графическое представление результатов кластеризации……..... 36
3.5. Определение числа кластеров…….…… 37
3.6. Методы минимальной дисперсии…….……... 38
3.6.1. Метод полных связей, или метод «дальнего соседа» (Complete linkage, Furthest neighbor)……... 39
3.6.2. Метод одиночной связи, или метод «ближайшего соседа» (Single linkage, Nearest neighbor)…….. 40
3.6.3. Метод невзвешенного попарного среднего (Unweighed pair-group average, Between-groups linkage)……..... 40
3.6.4. Метод Варда (Ward's method)…… 41
3.6.5. Невзвешенный центроидный метод (Unweighed pair-group centroid, Centroid clustering)…….…..... 42
3.6.6. Метод взвешенного попарного среднего, или метод минимальной связи (Weighted pair-group average, Within-groups linkage)……..……..…… 42
3.6.7. Метод медиан, или взвешенный центроидный метод (Weighted pair-group centroid Median clustering)……..….. 43
3.7. Реализация кластерного анализа в пакете SPSS……..…… 43
3.8. Особенности реализации кластерного анализа в пакете «Statistica»..... 48
3.9. Пример. Расчет загрузки персонала с учетом специализации……..… 55
Глава 4. Факторный анализ……..…… 58
4.1. Постановка задачи. Линейная модель факторного анализа…….…..… 58
4.2. Дисперсия, коэффициенты корреляции признаков и их составляющие…..…….. 60
4.3. Алгоритмы решения задачи факторного анализа. Метод главных компонент…….. 61
4.4. Пример. Выявление основных факторов, влияющих на объем продаж. Реализация факторного анализа в пакете «Statistica»…..……..... 62
4.5. Реализация факторного анализа в пакете SPSS……..……..……..… 68
Литература……..……..... 70
Приложение 1. Таблица критических значений числа разбиений $S(\alpha, \beta)$……. 71
Приложение 2. Таблица F-распределения для уровня значимости $\alpha = 0{,}10$……. 72
Приложение 3. Измерения и шкалы……... 73
Приложение 4. Оценка качества и оформления новой марки сигарет. Результаты опроса …… 75

ВВЕДЕНИЕ

При системном анализе явлений исследователь довольно часто сталкивается с проблемой многомерности их описания. Это типично для статистической обработки информации многих задач в области психологии [4], техники [21, 22], экономики [8] и др. – например, при сегментировании рынка, прогнозировании конъюнктуры рынка отдельных товаров, построении типологии объектов по достаточно большому числу признаков и во многих других случаях.

Методы многомерного анализа – наиболее действенный количественный инструмент исследования социально-экономических явлений, описываемых большим числом характеристик. К ним относятся: метод ранжирования многомерных величин, кластерный анализ, таксономия, факторный анализ. Формально дисперсионный анализ нельзя трактовать как элемент этой группы методов, однако, по существу, в нем заложена сходная с перечисленными выше методами идеология [18]. Кластерный анализ наиболее ярко отражает черты многомерного анализа в классификации, факторный анализ – в исследовании связей. Дисперсионный анализ ориентирован на выявление зависимости изменения числовых характеристик наблюдаемой величины от некоторого качественного показателя. Этот метод часто называют методом планирования эксперимента [27].

В современных условиях обработка больших числовых массивов, которые поставляет статистическое исследование в любой области знания, немыслима без использования компьютера. Естественно, существует и значительное количество программных продуктов, специализированных для решения таких задач. Достаточно назвать такие отечественные пакеты, как «Stadia» и «Olymp»; зарубежные – SAS, «Statgraphics», SPSS, «Statistica», «S-Plus» и другие [9]. Отметим, что здесь упоминаются только профессиональные и универсальные пакеты, перечень специализированных программ занял бы существенно больше места.

Методам многомерного анализа посвящено большое количество литературы. Это касается и математического обоснования методов (см., например, [10, 14, 20, 26]), и их компьютерной реализации (например, [2, 3, 7, 9, 19, 24]).

В настоящем пособии в компактной форме изложены математические основания методов многомерного анализа, приводятся их разновидности и особенности применения. Также описана пошаговая реализация многомерных методов в рамках пакетов SPSS, «Statistica» и «Stadia». Выбор именно этих программных продуктов связан, во-первых, с их популярностью среди специалистов в области маркетинга, менеджмента, психологии и техники, а во-вторых, с доступностью этих пакетов. Все расчеты проведены в версиях: SPSS 11.5, «Statistica 6.0» и «Stadia 6.2».

Издание адресовано студентам математико-механического и экономического факультетов, специалистам в области экономики и управления, слушателям программы MBA, а также может оказать помощь всем тем, кто занимается обработкой статистической информации многомерных объектов.

Глава 1. МЕТОД МНОГОМЕРНОЙ СРЕДНЕЙ

В практической деятельности довольно часто приходится сталкиваться с проблемой рационального выбора (наилучшего с точки зрения субъекта, принимающего решение) одного или нескольких объектов из множества аналогичных. При этом, как правило, объективная (числовая) оценка, имеющая статистический характер, базируется на ряде наблюдений о характере поведения исследуемого множества этих объектов. Если каждый из элементов такого множества характеризуется одной числовой характеристикой, то проблема выбора очевидна – если, конечно, лицо, принимающее решение, знает, что хочет. Достаточно расположить элементы в порядке возрастания характеристики, отражающей его привлекательность (например, товары одинаковой цены, но с разными потребительскими качествами).

Однако, если в расчет принимается не одна, а две, три или более характеристик сравниваемых объектов, проблема становится куда более сложной. Прямое сравнение здесь невозможно, ибо у разных объектов есть сильные и слабые стороны. Как правило, если первый из однотипных объектов превосходит второй по одной характеристике, то уступает ему по другой. Например, DVD-рекордер BENQ DW 1620, стоимостью $70, – скоростной и достаточно бесшумный аппарат, но он не записывает DVD-RAM, а универсальный рекордер LG GSA-4120B, стоимостью $65, – не самый быстрый и немного шумноват.

Задача еще более осложняется, если объекты характеризуются показателями, измеренными в различающихся единицах, а иногда и шкалах (см. Приложение 3 настоящего пособия). Конечно, в ряде случаев можно довериться интуитивным решениям, основанным на эмоциональных мотивах. Так обычно и происходит в повседневной жизни. Но ситуация меняется, когда речь заходит о принципиальных решениях: выбор банка для открытия счета фирмы, покупка дорогостоящего промышленного оборудования, выбор сегмента рынка для развития бизнеса. В решении таких задач желательно чтобы эмоциональные мотивы нашли опору в объективной числовой оценке. Другими словами, требуется некоторый аппарат, позволяющий сравнивать объекты, характеризующиеся целым набором признаков (многомерные объекты), с учетом всех этих признаков.

Метод многомерной средней позволяет ранжировать многомерные объекты, и в большинстве случаев разделять их на группы (сегментировать), то есть решает сформулированную выше задачу [11]. Это наиболее простой из рассматриваемых здесь и тем не менее очень действенный метод обработки результатов наблюдений над многомерными величинами.

1.1. Исходные данные. Постановка задачи

Исходными данными в задачах многомерного анализа является набор векторов I_i, $i = 1, 2, \ldots, k$, выбранных в качестве множества объектов, подлежащих ранжированию. Как правило, это достаточно однородный массив. Он состоит из k различных векторов (объектов) одинаковой размерности (каждый объект охарактеризован по заданному набору n различных признаков):

$$I_1 = \begin{pmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,n} \end{pmatrix}, \; I_2 = \begin{pmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{pmatrix}, \; \ldots, I_k = \begin{pmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,n} \end{pmatrix}.$$

Таким образом, $x_{j,l}$ есть характеристика j-го объекта по l-му признаку.

Требуется провести ранжирование объектов, охарактеризованных по большому числу разнородных признаков. После ранжирования возможно проведение группировки (например, разбиением на три категории: высшая, средняя, низшая), учитывающей вклад значений по всем признакам.

1.2. Алгоритм реализации

Для реализации алгоритма удобно формализовать исходные данные в виде таблицы (табл. 1.2.1):

Таблица 1.2.1

Исходные данные

Объекты	Характеристики признаков			
	x_1	x_2	…	x_n
I_1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,n}$
I_2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,n}$
...	...	...	...	...
I_k	$x_{k,1}$	$x_{k,2}$	…	$x_{k,n}$
	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_n$

Каждая из строк этой таблицы содержит информацию об одном из объектов по всем признакам (разнородные данные с разными единицами измерения), а каждый столбец – информацию по всем объектам по одному признаку (однородные данные с одинаковой единицей измерений).

Среднее арифметическое, вычисленное по каждому из столбцов ($\bar{x}_j$, j = 1, 2, …, n), дает среднее значение признака по всей группе объектов:

$$\bar{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{i,j}.$$

Для приведения данных к виду, допускающему сравнение, требуется провести их нормирование по каждому из столбцов делением на соответствующее среднее по признаку значение:

$$y_{i,j} = \frac{x_{i,j}}{\bar{x}_j}.$$

Результатом такого нормирования являются сопоставимые безразмерные значения $y_{i,j}$, характеризующие признаки объектов. Если все рассматриваемые объекты достаточно однородны, то полученные в результате нормирования величины будут не только лишены размерности, но и будут представлять собой набор чисел, близких к единице.

Действительно, величина $y_{i,j}$ показывает, во сколько раз j-й показатель, подсчитанный для i-го объекта, превосходит соответствующее среднее значение этого признака по всему множеству рассматриваемых объектов.

После этой процедуры каждый объект может быть охарактеризован по всем нормированным признакам средним значением – α_i, то есть одним числом. Теперь возможно ранжирование по принципу «чем больше, тем лучше» – или, наоборот, в зависимости от направленности показателей (см. табл. 1.2.2).

Таблица 1.2.2

Ранжирование методом многомерной средней

Объекты	Характеристики признаков				Нормированные характеристики				Многомерные средние	Ранги
	x_1	x_2	…	x_n	y_1	y_2	…	y_n		
I_1	$x_{1,1}$	$x_{1,2}$	…	$x_{1,n}$	$y_{1,1}$	$y_{1,2}$	…	$y_{1,n}$	$\alpha_1 = \frac{1}{n}\sum_{i=1}^{n} y_{1,i}$	A_1
I_2	$x_{2,1}$	$x_{2,2}$	…	$x_{2,n}$	$y_{2,1}$	$y_{2,2}$	…	$y_{2,n}$	$\alpha_2 = \frac{1}{n}\sum_{i=1}^{n} y_{2,i}$	A_2
…	…	…	…	…	…	…	…	…	…	…
I_k	$x_{k,1}$	$x_{k,2}$	…	$x_{k,n}$	$y_{k,1}$	$y_{k,2}$	…	$y_{k,n}$	$\alpha_k = \frac{1}{n}\sum_{i=1}^{n} y_{k,i}$	A_k
	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_n$						

В табл. 1.2.2 через A_1, A_2, …, A_k обозначены рейтинги (ранги) объектов, присвоенные им на основании сравнения их обобщенных числовых характеристик α_1, α_2, …, α_k.

При реализации этого несложного алгоритма следует учитывать, что все характеристики признаков должны быть равнонаправленными (чем большее значение принимает характеристика, тем лучшее положение занимает объект по этому признаку). Если изначально это не так, то ситуацию легко исправить переходя к противоположным величинам, обратным характеристикам.

Так, например, можно выбрать в качестве характеристики признака, имеющего обратную направленность, величину, равную разности между максимальным значением и значением этого признака. Тем самым переходя к значению, показывающему, асколько данная характеристика лучше наихудшего значения. При наличии признаков разной значимости возможна модернизация метода с помощью введения весовых коэффициентов.

1.3. Реализация метода многомерной средней средствами «Excel»

В качестве примера рассмотрим задачу о выборе тарифа сотовой связи среди некоторых вариантов, предоставляемых компанией МТС (по состоянию на декабрь 2005 года) на территории Свердловской области.

Для характеристики различных тарифных планов выбраны следующие показатели: величина абонентской платы, количество оплаченных минут разговора в месяц, стоимость минуты исходящих звонков на мобильные телефоны внутри сети и на городские телефоны, стоимость SMS-сообщений.

Таблицу нормированных значений легко составить в программе «Excel». Однако при вычислениях следует обратить внимание на тот факт, что не все признаки являются однонаправленными. Большая их часть характеризует тариф по принципу «чем больше, тем хуже», за исключением признака «количество оплаченных минут разговора», где характеристика диаметрально противоположна. Элементам этого столбца присваивается значение равное количеству недоплаченных минут по сравнению с тарифным планом соответствующим максимальному значению оплаченных минут (здесь это «Бизнес 350»).

	A	B	C	D	E	F	G
1	Тарифные планы МТС						
2	№	Тариф	Аб. Плата	Опл.мин	Исх. Моб	Исх. ГТС	SMS
3	1	Оптима+50	7,07	300	0,13	0,14	0,05
4	2	Оптима-Вечер	2,35	350	0,13	0,15	0,05
5	3	Оптима+100	11,79	250	0,13	0,13	0,05
6	4	Оптима-Унив	2,35	350	0,13	0,13	0,05
7	5	Бизнес 250	22,42	100	0,11	0,11	0,05
8	6	Бизнес 350	46,02	0	0,12	0,12	0,05
9	7	Оптима Губерн	2,35	350	0,13	0,2	0,05
10	8	Джинс 007	0	350	0,07	0,31	0,06
11	9	Супер Джинс	0	350	0,24	0,24	0,06
12		Среднее значение.	10,48333333	266,666667	0,1322222	0,17	0,052222

Рис. 1.3.1. Расчет многомерной средней по некоторым тарифным планам МТС

Так, в рамках тарифного плана «Оптима +50» абоненту предоставляется 50 минут бесплатного разговора, что на 300 минут меньше, чем в рамках тарифного плана «Бизнес 350», поэтому в графе «Оплаченные минуты» указываем число 300.

После нормирования производится подсчет среднего значения по строкам. Затем легко произвести ранжирование, при этом тарифу с наименьшим средним значением присваивается первое место, а с наибольшим – последнее.

Тарифные планы МТС								
№	Тариф	Аб. Плата	Опл.мин	Исх. Моб	Исх. ГТС	SMS	Ср. знач.	Ранги
1	Оптима+50	0,674403816	1,125	0,9831933	0,82352941	0,957447	0,912715	3
2	Оптима-Вечер	0,224165342	1,3125	0,9831933	0,88235294	0,957447	0,871932	2
3	Оптима+100	1,124642289	0,9375	0,9831933	0,76470588	0,957447	0,953498	5
4	Оптима-Унив	0,224165342	1,3125	0,9831933	0,76470588	0,957447	0,848402	1
5	Бизнес 250	2,13863275	0,375	0,8319328	0,64705882	0,957447	0,990014	7
6	Бизнес 350	4,389825119	0	0,907563	0,70588235	0,957447	1,392143	9
7	Оптима Губерн	0,224165342	1,3125	0,9831933	1,17647059	0,957447	0,930755	4
8	Джинс 007	0	1,3125	0,5294118	1,82352941	1,148936	0,962875	6
9	Супер Джинс	0	1,3125	1,8151261	1,41176471	1,148936	1,137665	8

Рис. 1.3.2. Ранжированные тарифные планы МТС

При больших массивах данных ранжирование проще всего проводить с использованием функции **Сортировка**.

Полученный результат, конечно же, не является руководством к действию (кто-то практически не пользуется SMS-почтой, кто-то не разговаривает по телефону больше часа в месяц и т.д.), но он дает основания для рационального выбора. Кроме того, каждый может расставить здесь свои приоритеты с помощью весовых коэффициентов.

Как отмечалось выше, метод многомерной средней позволяет произвести разделение объектов на группы (в каком-то смысле, провести сегментацию). Так, в рассмотренном примере легко выделить три группы: выше среднего «Оптима-Универсал», «Оптима-Вечер» и «Оптима+50» (1-, 2- и 3-е места); средние: «Оптима-Губерния», «Оптима+100» и «Джинс 007» (4-, 5- и 6-е места);ниже среднего: «Бизнес 250», «Супер Джинс» и «Бизнес 350» (7-, 8- и 9-е места).

Глава 2. ДИСПЕРСИОННЫЙ АНАЛИЗ

Идея дисперсионного анализа, как и сам термин «дисперсия», принадлежит английскому статистику Р. Фишеру. Метод был разработан в 1920-х годах [1] и используется для определения степени влияния на изучаемый показатель некоторых факторов, в том числе и не поддающихся количественному измерению (достаточно, чтобы его можно было измерить хотя бы в шкале наименований). При исследовании зависимостей такого рода одной из наиболее простых является ситуация, когда можно указать один только фактор, который, возможно, влияет на конечный результат, и этот фактор может принимать лишь конечное число значений (уровней). В этом случае реализуется алгоритм однофакторного дисперсионного анализа. Если же на изменение показателя в равной степени могут оказывать влияние несколько факторов, то для установления степени их совместного влияния используется уже алгоритм двухфакторного или многофакторного анализа [3, 16 и др.].

2.1. Однофакторный дисперсионный анализ. Постановка задачи

Рассмотрим сначала проблему установления зависимости от одного фактора. Такие задачи весьма часто встречаются на практике. Типичный пример – сравнение эффективности нескольких различных способов действия, направленных на достижение одной цели (например, оценка эффективности работы компании при взаимодействии с различными поставщиками, результативности обучения на основании различных методик, успеха продвижения товара при использовании различных маркетинговых подходов). В пакетах прикладных программ часто конкретную реализацию фактора называют уровнем фактора или способом обработки, а значения измеряемого признака (то есть величину результата) – откликом.

Для сравнения влияния фактора на результат необходим определенный статистический материал. Обычно его получают следующим образом: каждый из k способов обработки применяют несколько раз (не обязательно одно и то же число раз) к исследуемому объекту и регистрируют результаты. Итогом подобных испытаний будет k выборок, вообще говоря, разных объемов. Таким образом, исходные данные представляют собой числовой массив (объема n) наблюдений случайной величины, зафиксированных при различных уровнях (k уровней фактора, $k < n$) внешнего фактора. При этом предполагается, что наблюдаемая величина распределена по нормальному закону. На первом уровне фактора n_1 наблюдений: $x_{1,1}$, $x_{1,2}$, …, x_{1,n_1}, на втором уровне фактора n_2 наблюдений: $x_{2,1}$, $x_{2,2}$, …, x_{2,n_2}, и так далее, на k-м уровне фактора n_k наблюдений: $x_{k,1}$, $x_{k,2}$, …, $x_{k,nk}$. Причем очевидно, $n_1 + n_2 + ... + n_k = n$. Наиболее удобным способом представления подобных данных является таблица вида 2.1.1.

Таблица 2.1.1

Исходные данные для однофакторного дисперсионного анализа

Уровень фактора	Значения выборки	Объем выборки
1	$x_{1,1}$, $x_{1,2}$, … x_{1,n_1}	n_1
2	$x_{2,1}$, $x_{2,2}$, … x_{2,n_2}	n_2
⋮	⋮	⋮
k	$x_{k,1}$, $x_{k,2}$, … $x_{k,nk}$	n_k

Требуется при заданной надежности определить, влияет рассмотренный внешний фактор на характер поведения наблюдаемой случайной величины или нет. При этом общая модель однофакторного дисперсионного анализа имеет вид:

$$x_{i,j} = \bar{x} + \alpha_i + \beta_{i,j}, \qquad (2.1.1)$$

где $\bar{x}$ – генеральная средняя (среднее арифметическое из всех наблюдаемых значений случайной величины вне зависимости от принадлежности к какой-либо группе уровня фактора); α_i – числовая характеристика степени влияния фактора на его i-м уровне; $\beta_{i,j}$ – случайная величина не подверженная влиянию фактора. Предполагается, что $\beta_{i,j}$ распределена по нормальному закону и имеет нулевое математическое ожидание ($\beta_{i,j} \sim N(0,\sigma)$, где σ – среднее квадратическое отклонение $\beta_{i,j}$).

Дисперсионный анализ является результатом применения общего статистического метода проверки гипотез [16] к сформулированной задаче. При этом в качестве гипотез рассматриваются следующие предположения: H_0 – фактор не оказывает существенного влияния на конечный результат (основная гипотеза); H_1 – фактор значимо влияет на конечный результат (конкурирующая гипотеза). Гипотеза H_0 эквивалентна предположению о том, что все α_i в (2.1.1) равны нулю.

2.2. Алгоритм реализации

Идея метода основана на сравнении величин дисперсий, порожденных влиянием внешнего фактора D_A (систематическая, или межгрупповая дисперсия) и дисперсии выборки, освобожденной от воздействия внешнего фактора D_R (остаточная, или внутригрупповая дисперсия). Систематическая дисперсия определяется как дисперсия между средними каждой из k групп, соответствующих различным уровням фактора. Эта дисперсия характеризует степень разброса усредненных по группам значений, ее величина зависит лишь от степени влияния фактора на изменение случайной величины. Остаточная дисперсия определяется как совокупная внутригрупповая дисперсия, причем в каждой группе наблюдаемые значения варьируются относительно своей групповой средней. Исчисление этой дисперсии исключает влияние рассматриваемого фактора, ее величина определяется лишь погрешностью измерений и влиянием других, неучтенных факторов.

Если систематическая дисперсия соизмерима с остаточной, то и влияние фактора нельзя признать значимым, ибо объясняемое им разнообразие поведения случайной величины соизмеримо с разнообразием, порожденным неточностью измерений. Другое дело, если объясненная наличием фактора часть дисперсии существенно больше, чем ее часть, порожденная случайными помехами. Решение о значимости воздействия внешнего фактора принимается на основании сравнения наблюдаемого и теоретического значений F-статистики Фишера – Снедекора.

Реализация алгоритма дисперсионного анализа связана с вычислением средних значений и вариаций как по каждой из групп:

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{j,i} \quad \text{и} \quad Q_j = \sum_{i=1}^{n_j}(x_{j,i} - \bar{x}_j)^2, \qquad j = 1, 2, \ldots, k,$$

так и по всей генеральной совокупности наблюдений:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{j,i} \quad \text{и} \quad Q = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{j,i} - \bar{x})^2 .$$

Следует отметить, что средняя из групповых средних обычно не совпадает с генеральной средней за исключением случая, когда совпадают объемы всех групп: $n_1 = n_2 = \ldots = n_k$.

Сумма всех Q_j называется остаточной (или внутригрупповой) вариацией

$$Q_R = \sum_{j=1}^{k} Q_j = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{j,i} - \bar{x}_j)^2 . \qquad (2.2.1)$$

Эта величина освобождена от влияния фактора, ибо варьируются здесь элементы каждой из групп вокруг групповой средней. Именно групповая средняя берет на себя всю степень воз-

действия фактора на случайную величину на каждом его уровне. Подстановкой (2.1.1) в (2.2.1) легко убедиться, что $Q_R = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \beta_{i,j}^{\ 2}$.

В отличие от остаточной, величина межгрупповой вариации, определяемой равенством

$$Q_A = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 , \qquad (2.2.2)$$

напрямую зависит от степени влияния фактора на случайную величину. Действительно, систематическая вариация получается варьированием групповых средних относительно генеральной. Подстановкой (2.1.1) в (2.2.2) нетрудно получить, что при неограниченном возрастании объема выборки $\lim\limits_{n\to\infty} Q_A = \sum\limits_{j=1}^{k} \alpha_j^2$.

Дальнейший анализ базируется на основном тождестве дисперсионного анализа [3]:

$$Q = Q_A + Q_R = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{j,i} - \bar{x}_j)^2 + \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 , \quad 2.2.3)$$

которое получается, если возвести в квадрат обе части очевидного равенства $x_{i,j} - \bar{x} = (\bar{x}_i - \bar{x}) + (x_{i,j} - \bar{x}_i)$, затем просуммировать результат по всем *i* и *j* и учесть, что

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{j,i} - \bar{x}_j)(\bar{x}_j - \bar{x}) = 0$$

в силу определения групповых и генеральной средних $\bar{x}_j$ и $\bar{x}$.

Внося результаты вычислений в табл. 2.2.1, являющуюся расширенным вариантом таблицы данных (табл. 2.1.1), получим:

Таблица 2.2.1

Результаты промежуточных расчетов однофакторного дисперсионного анализа

Уровень фактора	Значения выборки	Объем выборки	Средние	Вариации
1	$x_{1,1},\ x_{1,2},\ \dots\ x_{1,n_1}$	n_1	$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1,i}$	$Q_1 = \sum_{i=1}^{n_1} (x_{1,i} - \bar{x}_1)^2$
2	$x_{2,1},\ x_{2,2},\ \dots\ x_{2,n_2}$	n_2	$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2,i}$	$Q_2 = \sum_{i=1}^{n_2} (x_{2,i} - \bar{x}_2)^2$
⋮	⋮	⋮	⋮	⋮
k	$x_{k,1},\ x_{k,2},\ \dots\ x_{k,n_л}$	n_k	$\bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{k,i}$	$Q_k = \sum_{i=1}^{n_k} (x_{k,i} - \bar{x}_k)^2$
$x_{1,1},\ x_{1,2},\ \dots\ x_{1,n_1}, x_{k,1}, \dots, x_{k,2},\ \dots\ x_{k,n_л}$		*n*	$\bar{x} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{j,i}$	$Q = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{j,i} - \bar{x})^2$

Сумма всех, за исключением нижнего, элементов последнего справа столбца определит остаточную вариацию *Q*. Систематическая вариация может быть получена также на основании (2.2.3): $Q_A = Q - Q_R$.

Известно [16], что для выборки из значений нормально распределенной случайной величины, сумма квадратов $\sum\limits_{i=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ может быть представлена в виде произведения $\sigma_j^2 \chi^2$, где случайная величина χ^2 имеет распределение Пирсона с $n_i - 1$ степенями свободы, $D_j = \sigma_j^2$.

Поскольку данные в разных строках второго столбца таблицы 2.2.1 получены в результате независимых испытаний, объединенная сумма квадратов $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$ имеет распределение $D_R\chi^2$ с $N-k$ степенями свободы. Отсюда получаем оценку остаточной дисперсии:

$$D_R=\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 .$$

Согласно свойствам нормального распределения (предельные теоремы [21]), групповые средние $\bar{x}_1$, $\bar{x}_2$, …, $\bar{x}_k$ распределены по нормальному закону: $\bar{x}_i \sim N(a_i,\frac{\sigma^2}{n_i})$. Величина $Q_A=\sum_{j=1}^{k}n_j(\bar{x}_j-\bar{x})^2$ распределена по закону $\sigma^2\chi^2$ с $k-1$ степенью свободы. При этом Q_A и Q_R являются независимыми. В случае значительного превышения значения Q_A над значением Q_R следует признать, что рассматриваемый фактор оказывает влияние на исследуемую случайную величину. В противном случае – признать фактор несущественным для дальнейшего рассмотрения.

Таким образом, задача дисперсионного анализа формализуется как стандартная задача проверки гипотез [16, 27]. Основная гипотеза H_0 – рассматриваемый фактор не оказывает существенного влияния на изменение случайной величины (по существу, гипотеза об отсутствии статистической разницы между групповыми средними). Руководствуясь субъективными соображениями, задаем уровень значимости α. В качестве статистики метода выбирается частное от деления систематической дисперсии на остаточную:

$$F=\frac{D_A}{D_R}=\frac{\frac{1}{k-1}\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{\sigma})^2}{\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\tilde{o}_{ij}-\bar{o}_i)^2}. \qquad (2.2.4)$$

Известно, что если гипотеза H_0 верна, то эта статистика имеет распределение Фишера – Снедекора с ($k-1$, $N-k$) степенями свободы [16].

На основании (2.2.4) и имеющейся выборки определяем наблюдаемое значение статистики $F_{\text{набл.}}$ (подставляя данные табл. 2.2.1 в (2.2.4)). По таблицам распределения Фишера – Снедекора находим критическое значение статистики (таблица критических значений F-статистики при α = 0,10 приведена в Приложении 2) при заданном уровне значимости α и полученных степенях свободы $F_{\text{кр.}} = F_{1-\alpha}(N-k,\ k-1)$. Если $F_{\text{набл.}} < F_{\text{кр.}}$, гипотеза принимается (фактор не оказывает статистически значимого влияния на изменение случайной величины). В противном случае ($F_{\text{набл.}} > F_{\text{кр.}}$) следует признать значимость влияния фактора на изменение наблюдаемой случайной величины.

2.3. Однофакторный дисперсионный анализ средствами «Excel»

Наиболее простой способ провести дисперсионный анализ – использовать программы электронных таблиц «Excel» [5]. Для этого воспользуемся процедурой **Анализ данных** (**Сервис /Анализ данных**). Если в установленной конфигурации эта процедура отсутствует, то ее следует подключить через опцию **Сервис/Настройки**, поставив в появившемся меню «флажок» напротив строки **Пакет анализа**. Если такой настройки нет в списке, то потребуется переустановка всей системы.

Реализацию алгоритма однофакторного дисперсионного анализа проведем на решении модельного примера.

Пример. Отдел рекламы компании «Репа», специализирующейся на оптовой продаже семян высокоурожайных сортов сельскохозяйственных культур, разработал четыре типа различ-

ных этикеток для фасовки семян огурцов «Асанофф». Для выяснения вопроса о том, влияет ли внешний вид упаковки товара на объем его продажи, в четыре однотипных магазина были поставлены семена огурцов с разными этикетками (в первый магазин – с этикетками первого типа, во второй – с этикетками второго типа и т.д.). За первые четыре недели апреля были получены следующие данные об объемах реализованной продукции (табл. 2.3.1).

Таблица 2.3.1

Объемы реализации семян огурцов «Асанофф» (кол-во упаковок)

Период	Магазин			
	№ 1	**№ 2**	**№ 3**	**№ 4**
1-я неделя	140	150	148	150
2-я неделя	144	149	149	155
3-я неделя	142	152	146	154
4-я неделя	145	150	147	152

Следует ли считать значимым влияние вида этикетки на объем продаж?

Для решения задачи восстановим табл. 2.3.1 на первом листе книги «Excel». Далее в наборе настроек **Анализ данных** выберем строку **Однофакторный дисперсионный анализ** (рис. 2.3.1).

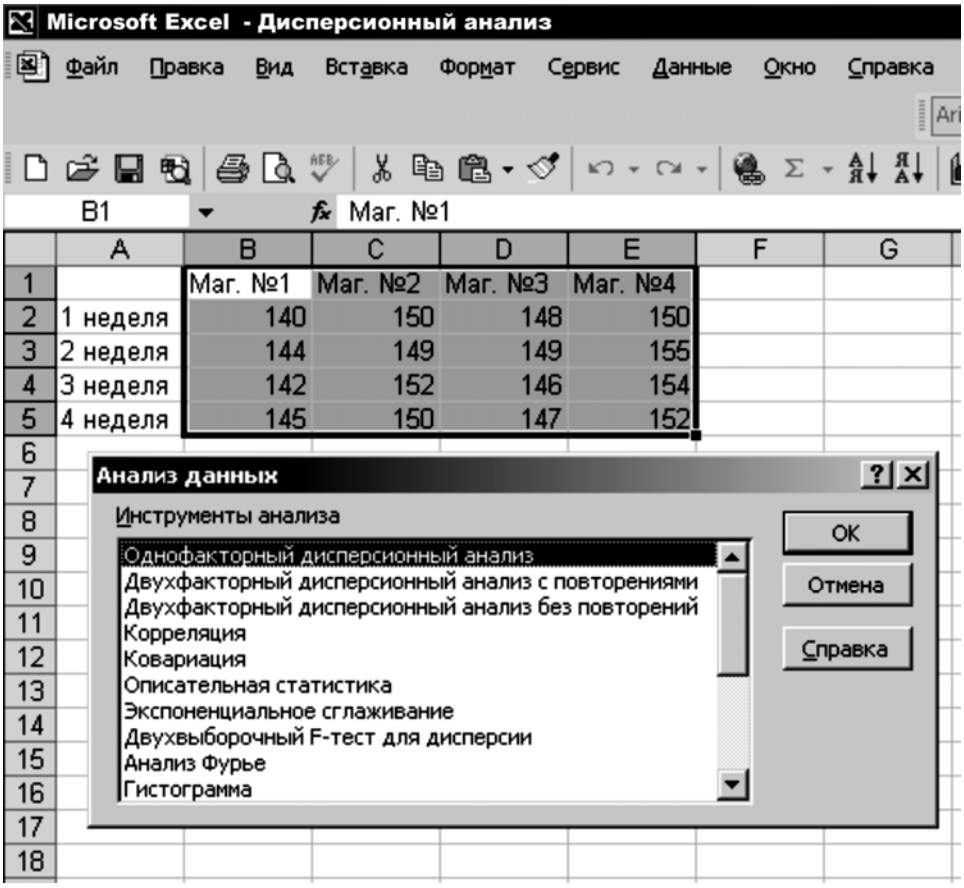


Рис. 2.3.1. Выбор процедуры однофакторного дисперсионного анализа в «Excel»

В появившемся окне требуется ввести данные задачи – **Входной интервал,** отметив способ группировки вводимой числовой информации: по столбцам или по строкам. В графе **Метки в первой строке** следует установить «флажок», если табличные данные вводятся вместе с заголовком (не числовыми характеристиками: «Маг. № 1, …»). Здесь же требуется задать уровень значимости в окне **Альфа**.

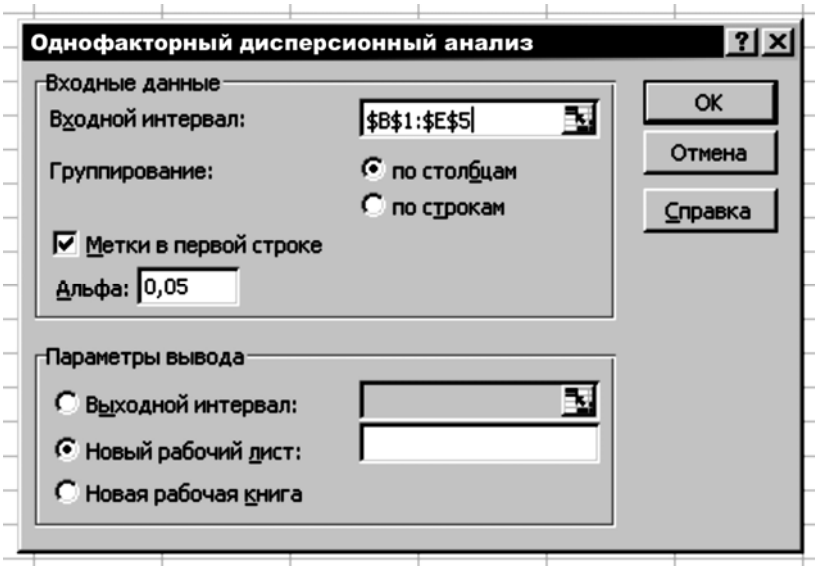


Рис. 2.3.2. Окно ввода параметров дисперсионного анализа в «Excel»

Результаты однофакторного дисперсионного анализа выводятся в окно итогового отчета в виде таблицы (табл. 2.3.2).

Таблица 2.3.2

Итоговый отчет о результатах однофакторного дисперсионного анализа в «Excel»

Однофакторный дисперсионный анализ

ИТОГИ

Группы	Счет	Сумма	Среднее	Дисперсия
Магазин № 1	4	571	142,75	4,9166667
Магазин № 2	4	601	150,25	1,5833333
Магазин № 3	4	590	147,5	1,6666667
Магазин № 4	4	611	152,75	4,9166667

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-значение	F критическое
Между группами	220,1875	3	73,39583	22,43949	3,28029E-05	3,490299605
Внутри групп	39,25	12	3,270833			
Итого	259,4375	15				

В первой части табл. 2.3.2 приведена общая информация о величинах, вычисляемых в процессе выполнения алгоритма. В столбце **Счет** указывается объем соответствующей группы (n_i), а в столбце **Сумма** – сумма элементов группы. Последующие столбцы **Среднее** и **Дисперсия** содержат средние групповые значения ($\bar{x}$) и оценку дисперсии соответственно (D).

Нижняя часть табл. 2.3.2 содержит отчет о проведении дисперсионного анализа. Здесь приводятся такие межгрупповые и внутригрупповые характеристики, как сумма квадратов отклонений от среднего или вариации (**SS**), числа степеней свободы (**df**) и дисперсии (**MS**).

Здесь же приведены наблюдаемое (**F**) и критическое (**F критическое**) значения F-статистики Фишера – Снедекора, а также p – минимальный уровень значимости (**P-значение**) этой статистики.

В рассмотренном примере получилось $F_{\text{набл.}} > F_{\text{кр.}}$, что позволяет сделать вывод о том, что фактор оказывает влияние на изменение случайной величины. Другими словами, влияние вида этикетки на объем продаж значимо.

2.4. Реализация однофакторного дисперсионного анализа в пакете «Stadia»

Реализацию дисперсионного анализа в пакете “Stadia” рассмотрим на модельном примере проверки гипотезы о наличии сезонной волны в объемах реализации некоторого товара по имеющейся статистической информации. Данные об объемах продаж, полученные за год, сведем в таблицу 2.4.1.

Таблица 2.4.1

Помесячные объемы продаж

Месяц	янв.	февр.	март	апр.	май	июнь	июль	авг.	сент.	окт.	нояб.	дек.
Объем продаж	*58*	*23*	*29*	*37*	*30*	*38*	*36*	*55*	*62*	*48*	*84*	*124*

В качестве фактора в рассматриваемой задаче выступает сезонность. Естественно зафиксировать четыре уровня фактора: зима, весна, лето, осень. В таком случае исходные данные можно представить в виде следующей таблицы (табл. 2.4.1).

Таблица 2.4.1

Исходные данные для дисперсионного анализа

Уровень фактора	Объемы продаж	Объем выборки
1 – зима	*58, 23, 124*	*3*
2 – весна	*29, 37, 30*	*3*
3 – лето	*38, 36, 55*	*3*
4 – осень	*62, 48, 84*	*3*

В электронной таблице пакета «Stadia» введем данные, относящиеся к первому сезону, в переменную **x1**, ко второму – в переменную **x2**, и так далее (рис. 2.4.1).

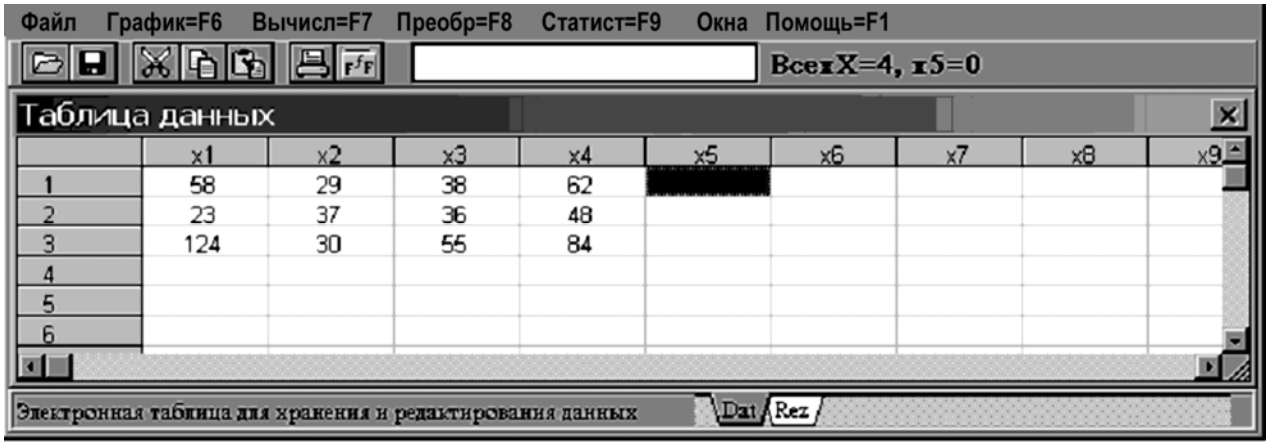


Рис. 2.4.1. Данные для однофакторного дисперсионного анализа в пакете «Stadia»

Процедуры однофакторного анализа пакета «Stadia» требуют, чтобы данные, отвечающие различным способам обработки (уровням фактора), относились к отдельным переменным.

При этом в файле данных не должны содержаться «посторонние» переменные. Отсюда следует, что для проведения анализа только для части способов обработки или объединения не-

скольких способов обработки в один, следует завести новый файл данных и осуществить в нем требуемые преобразования.

Для инициации дисперсионного анализа требуется в меню **Статистические методы** (рис. 2.4.2) выбрать пункт **B=Однофакторный**, в появившемся затем запросе выбрать нужные переменные для анализа, а в следующем – нажать кнопку **1=параметрический**.

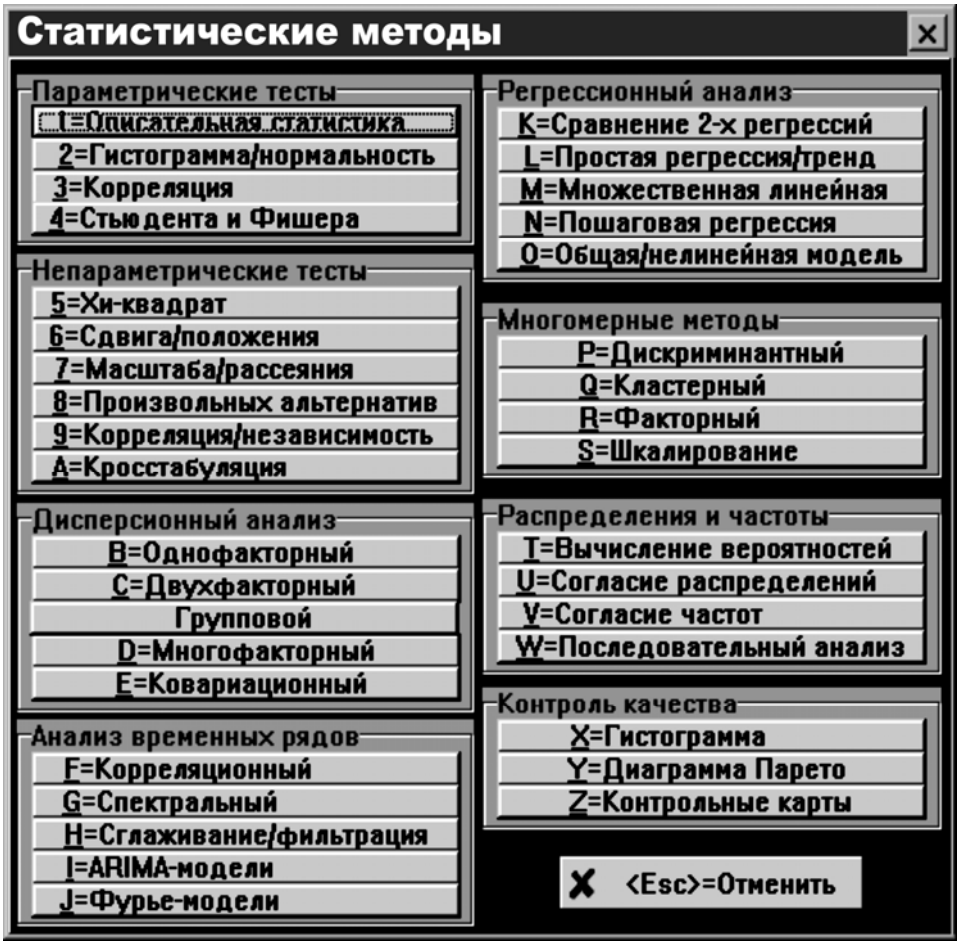


Рис. 2.4.2. Меню статистических методов в пакете «Stadia»

Результаты выводятся в окно в виде базовой таблицы дисперсионного анализа и значения оценок параметров модели (рис. 2.4.3). Назначение базовой таблицы дисперсионного анализа – дать ответ на вопрос о наличии значимого влияния уровней фактора на исследуемый отклик, или, другими словами, о присутствии эффектов обработки (в рассматриваемом случае принимается гипотеза H_0 об отсутствии влияния фактора, о чем сообщается «Нет влияния фактора на отклик»).

```
1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл:
          параметрический

Источник   Сум.квадр  Ст.своб Ср.квадр      F   Значимость Сила влияния
  Факт.1        2725        3    908,2  1,177   0,3782  -0,2247
  Остат.        6175        8    771,9
  Общая         8900       11    809,1

F(фактор1)=1,177, Значимость=0,3782, степ.своб = 3,8
   Гипотеза 0: <Нет влияния фактора на отклик>

          Параметры модели:
Среднее = 52, доверит.инт.=61,85
Эффект1 = 16,33, доверит.инт.=106,4
Эффект2 = -20, доверит.инт.=106,4
Эффект3 = -9, доверит.инт.=106,4
Эффект4 = 12,67, доверит.инт.=106,4
```

*Рис. 2.4.3***.** Итоговый отчет в базовой таблице дисперсионного анализа и оценки параметров модели в пакете «Stadia»

Кроме того, в таблице приведена информация об основных величинах, вычисляемых в процессе выполнения алгоритма. В столбце **Сум. квадр.**, в строке **Общая** указана общая сумма квадратов разностей наблюдений и их среднего значения, то есть полная вариация Q, вычисленная по (2.2.3). В строке **Факт. 1** того же столбца приведен вклад в общую сумму квадратов, обусловленный различиями в уровнях фактора, то есть межгрупповая (систематическая) вариация: $Q_A = \sum_{j=1}^{k} n_j (\overline{x}_j - \overline{x})^2$.

В строке **Остат.** указан вклад в общую сумму квадратов, вызванный случайной изменчивостью данных внутри групп, то есть внутригрупповая (остаточная) вариация, определяемая по (2.2.1). Как было отмечено выше, сумма (2.2.1) первой и второй строк столбца **Сум. квадр.** таблицы совпадает с величиной, указанной в третьей строке этого столбца. Таким образом, смысл анализа вариации сводится к выяснению разложения общей суммы квадратов отклонений на две части. Первая из них интерпретируется как вариация, обусловленная введенной моделью, учитывающей наличие фактора, а вторая – как случайная изменчивость данных внутри самой модели (то есть изменчивость, вызванная любыми другими причинами, но не выделенным фактором).

В третьем столбце таблицы (**Ст. своб.**) приведены числа степеней свободы систематической, остаточной и полной вариаций соответственно.

Наконец, в третьем столбце таблицы (**Ср. квадр.**) находятся частные от деления величин первого столбца на соответствующие величины второго столбца, то есть систематическая, остаточная и полная дисперсии соответственно. Заметим, что, несмотря на название столбца, мы получили именно дисперсии (квадраты средних квадратических отклонений), а не средние квадратические отклонения.

Отношение систематической дисперсии к остаточной (2.2.4) и определяет наблюдаемое значение F-статистики. Как видно из таблицы (рис. 2.4.3), в рассматриваемом примере оно оказалось равным 1,177. Сравним наблюдаемое значение статистики с ее критическим значением. При уровне значимости $\alpha = 0{,}10$ и степенях свободы $k_1 = 3$ и $k_2 = 8$ критическое значение статистики, в соответствии с таблицей (Приложение 2), равно $F_{кр.} = 2{,}92380$. Это позволяет принять гипотезу о незначительности влияния сезонности на объем продаж.

Справа от F-отношения указывается минимальный уровень значимости указанной F-статистики. Если значимость F-статистики близка к нулю, есть основание принять гипотезу H_1. При этом система сравнивает уровень значимости F-статистики с 0,05, на основании чего формирует вывод, который выводит на экран в виде заключения «Нет влияния фактора на отклик» или, наоборот, «Есть влияние фактора на отклик».

В последнем столбце таблицы (рис. 2.4.3) выводится значение величины силы влияния фактора (по Снедекору), которое определяется на основании формулы:

$$h_{\delta}^2 = \frac{D_A - D_R}{D_A + (k-1)D_R},$$

где, как и прежде, D_A – межгрупповая дисперсия, D_R – внутригрупповая дисперсия, а величина k равна числу наблюдений в группе, если в каждой группе одинаковое число наблюдений. Если число наблюдений для каждого уровня фактора различно, в качестве n в этой формуле используют величину

$$k = \frac{1}{l-1}\left(n - \left(\sum_{i=1}^{k} n_i\right) \Big/ n\right),$$

где n – общее число наблюдений, l – число уровней фактора, n_i – число наблюдений на i-м уровне фактора. Величина силы влияния показывает, какую долю в разбросе данных можно объяснить моделью, учитывающей наличие фактора.

Раздел базовой таблицы **Параметры модели** (рис. 2.4.3) включает оценку общего среднего значения – **Среднее** и оценки отклонений от среднего в терминах размаха доверительного интервала. Такая же информация приводится для каждого уровня фактора в строках **Эффект 1**, **Эффект 2** и т.д. (по количеству зафиксированных уровней фактора).

2.5. «ANOVA» – реализация дисперсионного анализа в пакете «Statistica»

Особенности применения однофакторного дисперсионного анализа в пакете «Statistica» рассмотрим на том же, что и выше примере, а именно на задача о наличии сезонной волны в объемах продаж. Для ввода исходных данных требуется сформировать две переменные одна из них отображает помесячные продажи конкретного товара (назовем ее «Sales»), вторая характеризует сезонность («Season») (рис. 2.5.1).

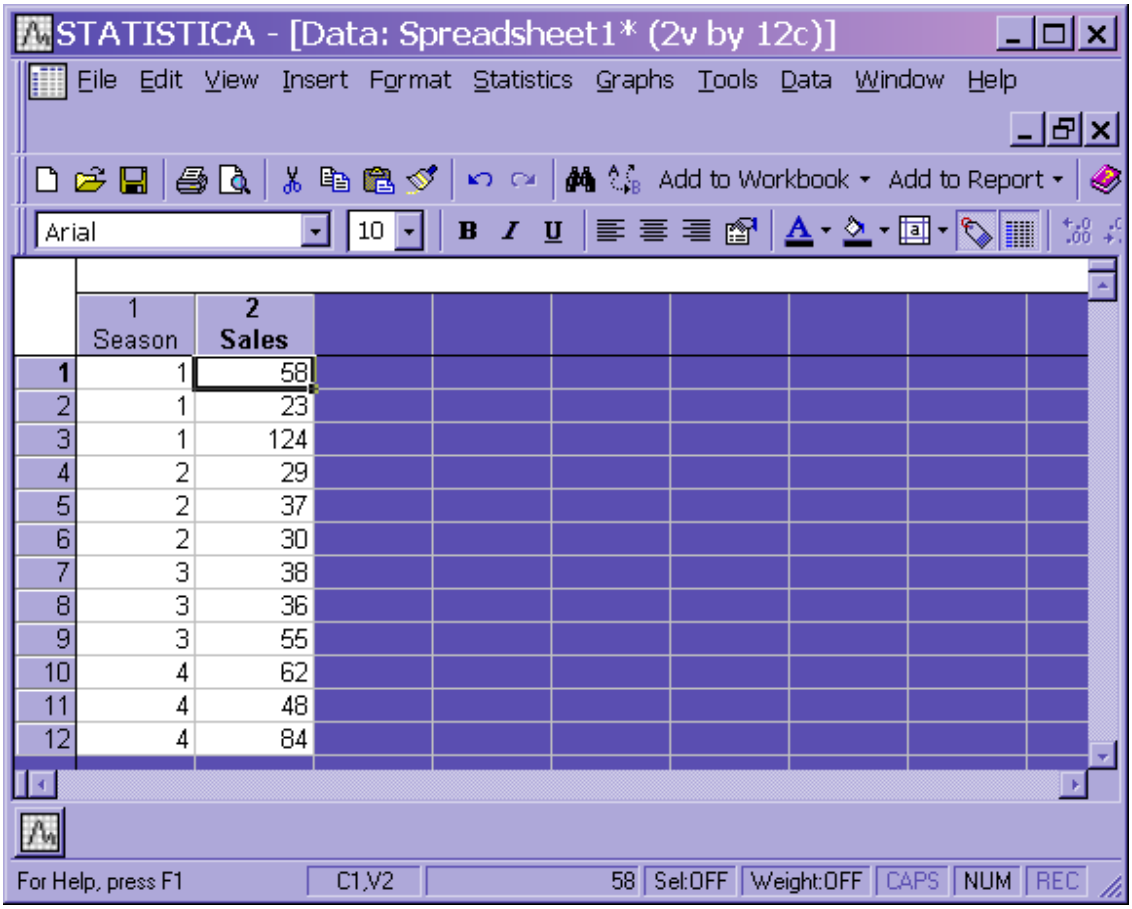


Рис. 2.5.1. Данные для однофакторного анализа в пакете «Statistica»

Следует отметить, что все рассматриваемые здесь пакеты допускают непосредственное экспортирование данных из электронных таблиц «Excel».

Для инициации процедуры однофакторного дисперсионного анализа в меню процедур **Basic Statistics and Tables** нужно выбрать пункт **ANOVA** (от англ. «analysis of variance» – дисперсионный анализ). В появившемся окне (рис. 2.5.2) необходимо задать тип анализа (**Type of analyses**), выбрав **One-way ANOVA** (однофакторный дисперсионный анализ), а в качестве метода спецификации (**Specification method**) выбрать **Quick specs dialog** (быстрый анализ).

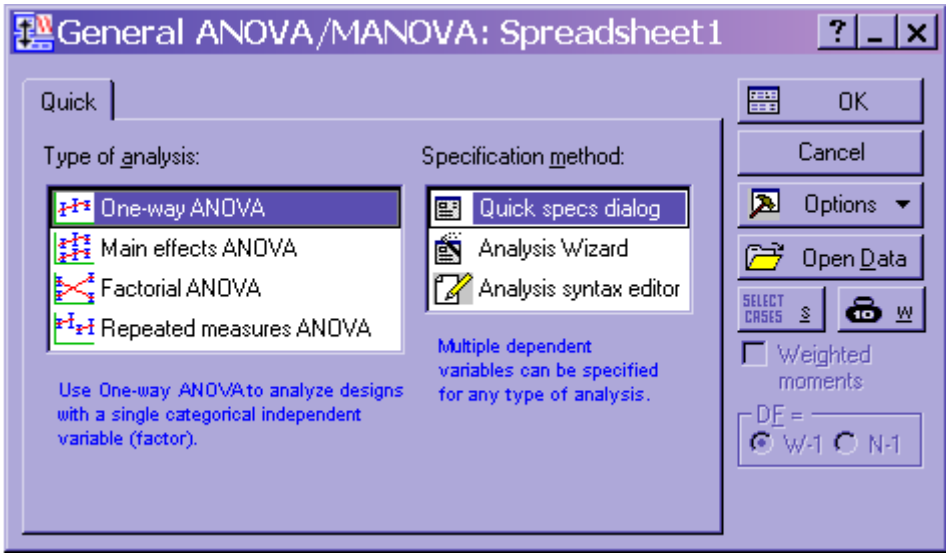


Рис. 2.5.2. Окно параметров дисперсионного анализа в пакете «Statistica»

В следующем окне (рис. 2.5.3), которое появится после нажатия на кнопку **Variances** (переменные) в меню дисперсионного анализа, будет предложено определить значения исходных данных, то есть выбрать зависимую переменную и переменную-фактор. В нашем примере ими будут соответственно «Sales» и «Season».

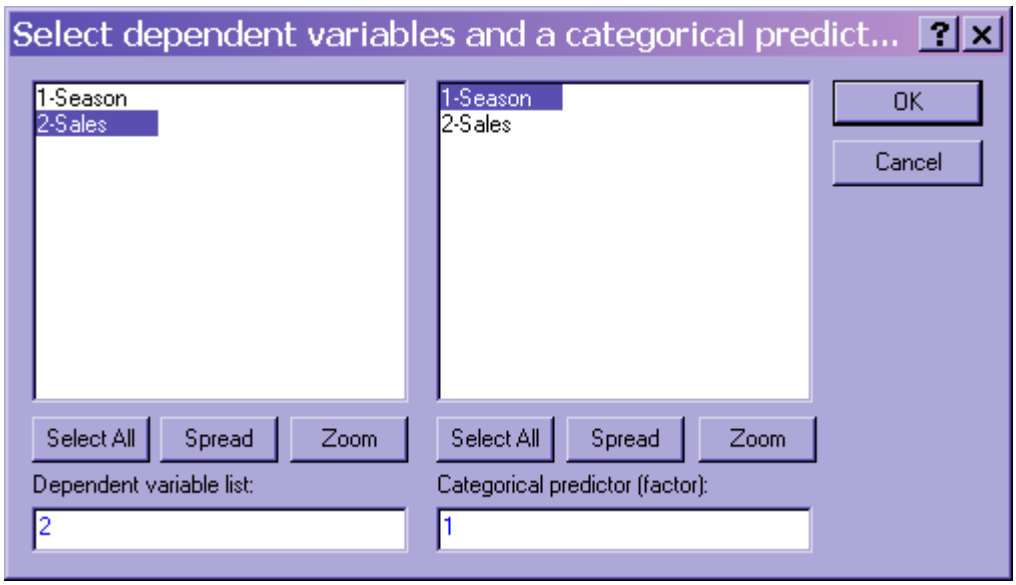
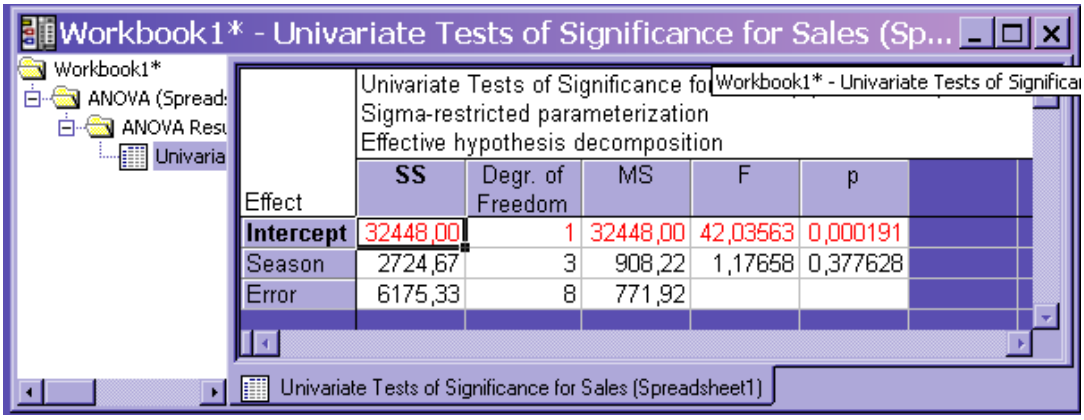


Рис. 2.5.3. Меню выбора переменных в пакете «Statistica»

Для получения результатов остается только нажать на **OK** и в появившемся окне – на кнопку **All effects**. В качестве результата в окно выводится таблица, по своей структуре аналогичная таблице дисперсионного анализа пакета «Stadia» (см. рис. 2.4.3).



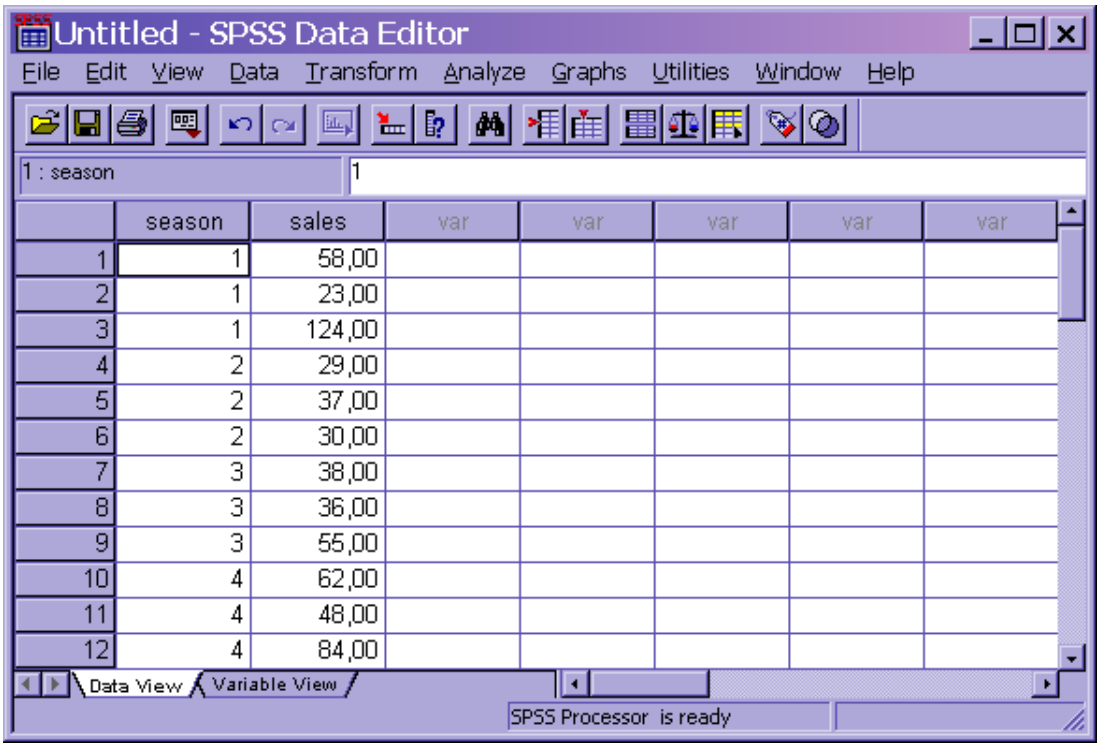
Effect	SS	Degr. of Freedom	MS	F	p
Intercept	32448,00	1	32448,00	42,03563	0,000191
Season	2724,67	3	908,22	1,17658	0,377628
Error	6175,33	8	771,92		

Рис. 2.5.4. Итоговый отчет о результатах дисперсионного анализа в пакете «Statistica»

Результаты, приведенные в строке «Season», относятся к данным, характеризующим влияние фактора (систематические характеристики), а в строке «Error» – к неучтенным воздействиям (остаточные характеристики). В итоговой таблице использованы следующие обозначения: **SS** – сумма квадратов, здесь приведены величины вариаций (SS, *Season* обусловливается межгрупповой изменчивостью, SS, *Error* – внутригрупповой изменчивостью); **Degr. of Freedom** – количество степеней свободы; **MS** – средний квадрат, здесь приведены соответствующие дисперсии; *F* – наблюдаемое в рассматриваемой задаче значение *F*- статистики; *p* – минимальный уровень значимости указанной *F*-статистики.

2.6. Особенности реализации дисперсионного анализа в пакете SPSS

Подготовка данных в пакете SPSS осуществляется аналогично тому, как это делается в пакете «Statistica». Таблица исходных данных имеет следующий вид (рис. 2.6.1).



	season	sales	var	var	var	var	var
1	1	58,00					
2	1	23,00					
3	1	124,00					
4	2	29,00					
5	2	37,00					
6	2	30,00					
7	3	38,00					
8	3	36,00					
9	3	55,00					
10	4	62,00					
11	4	48,00					
12	4	84,00					

Рис. 2.6.1. Данные для однофакторного анализа в пакете SPSS

Для запуска процедуры нужно в верхнем меню выбрать пункт **Analyze**, а в нем – пункт **Compare Means**, и далее – **One-Way ANOVA**, после чего появится подменю разделения переменных (рис. 2.6.2).

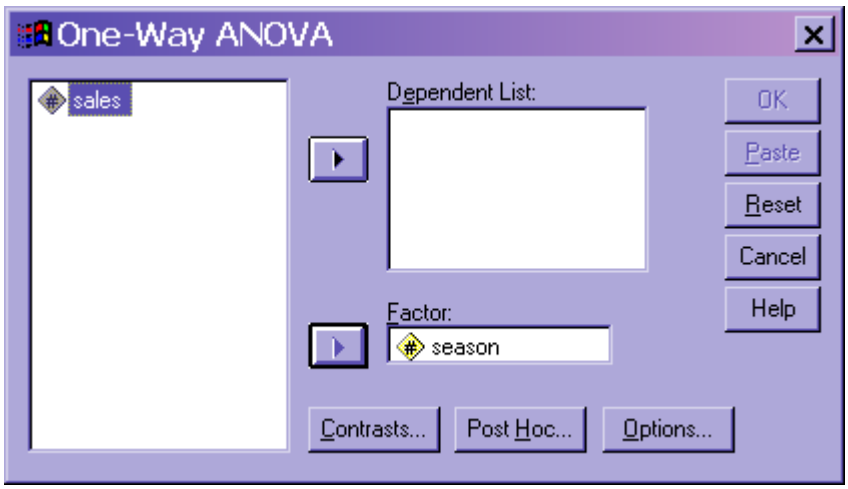


Рис. 2.6.2. Подменю разделения переменных в пакете SPSS

В этом окне следует разделить переменные на зависимые (**Dependent List**) и независимые (**Factor**). В рассматриваемом случае в список зависимых переменных попадет переменная «Sales», а фактором будет являться переменная «Season».

Результаты выводятся в отдельном окне **Output 1** в виде таблицы (рис. 2.6.3).

ANOVA

SALES

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2724,667	3	908,222	1,177	,378
Within Groups	6175,333	8	771,917		
Total	8900,000	11			

Рис. 2.6.3. Итоговый отчет о результатах дисперсионного анализа в пакете SPSS

Здесь **Sum of Squares** – сумма квадратов; **df** – количество степеней свободы; **Mean Square** – оценка дисперсии; **F** – наблюдаемое значение *F*-статистики; **Sig.** – минимальный уровень значимости указанной *F*-статистики.

2.7. Многофакторный дисперсионный анализ

Как отмечалось выше, случайная величина может быть подвержена влиянию не одного, но нескольких факторов. Другими словами, однофакторная модель может оказаться незначимой, если влияние фактора *A,* определяемое *F*-отношением, является несущественным на фоне большого внутригруппового разброса (остаточная вариация). Этот разброс может быть вызван не только случайными причинами, но также действием еще одного «мешающего» фактора *B*.

В этой ситуации фактор *B* дополнительно включается в модель (двухфакторный дисперсионный анализ), чтобы попытаться уменьшить действие неучтенных факторов и повысить влияние на отклик закономерных причин. Аналогично возникает необходимость рассмотрения трех- и многофакторных моделей. Соответственно, в процессе дисперсионного анализа варьируется не один, а несколько факторов. При полном многофакторном дисперсионном анализе отклик наблюдается для каждого сочетания уровней изучаемых факторов.

Рассмотрим матрицу наблюдений двухфакторного анализа (табл. 2.7.1). Главный фактор – фактор *A,* к примеру, влияние настройки станка; дополнительный фактор – фактор *B,* например, влияние качества сырья. Фактор *A* принимает *n,* а фактор *B* – *m* различных значений, т.е. *n* – число станков, *m* – число партий сырья.

Таблица 2.7.1

Исходные данные для двухфакторного дисперсионного анализа

Блоки	Уровни фактора *A* (способы обработки)					
	A_1	A_2	…	A_j	…	A_n
B_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n}
B_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n}
…	…	…	…	…	…	…
B_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}
…	…	…	…	…	…	…
B_m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mn}

Уровни фактора *A* (способы обработки) – отображаются в таблице по столбцам, а уровни фактора *B* (блоки) – по строкам. В блоке отклики могут значимо различаться только за счет различных уровней фактора *A,* т.е. за счет различных обработок. Это простейшая матрица наблюдений двухфакторного анализа, т.к. в каждой ячейке имеется только одно наблюдение x_{ij}. В отличие от матрицы однофакторного анализа наблюдения в любом столбце не являются однородными, т.е. не образовывают выборки, если влияние мешающего фактора значимо.

Вклады факторов *A* и *B* в значения отклика на соответствующих уровнях *i* и *j* обозначим через a_i и b_j. Между факторами нет взаимодействия. Таким образом, каждое наблюдение x_{ij} представляется в виде следующей аддитивной модели:

$$x_{ij} = b_i + a_j + \varepsilon_{ij}, i = \overline{1,m}, j = \overline{1,n}.$$

Как и в случае однофакторного дисперсионного анализа, предполагается, что для случайных величин ε_{ij} справедливо требование наличия нормального закона распределения $N(0,\sigma^2_\varepsilon)$, причем дисперсия σ^2_ε одинакова при всех значениях *j* и *i*.

Величины вкладов a_j и b_i, не могут быть восстановлены однозначно. Так, увеличение всех b_i, и уменьшение всех a_j одновременно на одну и ту же константу не изменит значения x_{ij}. Для однозначного определения вкладов факторов следует использовать отклонения β_i и τ_j отклика от общая средняя значений отклика μ (оценкой которой является величина генеральной средней - $\bar{x}$) в результате действия факторов *B* и *A*:

$$\beta_i + \tau_j = b_i + a_j - \mu,$$

$$\sum_{i=1}^{m} \beta_i = 0, \sum_{j=1}^{n} \tau_j = 0.$$

Величины $\beta_1, ..., \beta_m$, называются эффектами блоков, они характеризуют отклонения от β в результате действия фактора *B.* Эффекты обработки: $\tau_1,..., \tau_n$, характеризуют отклонения отклика из-за действия фактора *A*.

Тогда:

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, i = \overline{1,m}, j = \overline{1,n} \ .$$

Как и в случае однофакторного анализа, нулевая гипотеза H_0 об отсутствии эффектов обработки имеет вид: $\tau_1 = \tau_2 = ... = \tau_n = 0$.

При двухфакторном дисперсионном анализе общая сумма квадратов Q разбивается уже не на две, а на три части: Q_a и Q_b обусловленные влиянием факторов, и остаточную часть Q_R, обусловленную случайной изменчивостью самих наблюдений за счет неучтенных факторов, т.е.

$$\underbrace{\sum_{j=1}^{n}\sum_{i=1}^{m}(x_{ij}-\bar{x})^2}_{Q} = \underbrace{m\sum_{j=1}^{n}(\overline{x_{\bullet j}}-\bar{x})^2}_{Q_a} + \underbrace{n\sum_{i=1}^{m}\overline{x_{i\bullet}}}_{Q_b} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\overline{x_{\bullet j}}-\bar{x}_{i\bullet}+\bar{x})^2}_{Q_R},$$

где $\overline{x_{\bullet j}} = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$ - среднее по j-му столбцу, $(\overline{x_{\bullet j}} - \bar{x})$ - оценка эффекта обработки τ_j ; $\overline{x_{i\bullet}} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$ - среднее по i-му блоку, $(\overline{x_{i\bullet}} - \bar{x})$ - оценка эффекта блока β_i . Базовая таблица двухфакторного дисперсионного анализа принимает вид:

Таблица 2.7.2

Базовая таблица двухфакторного дисперсионного анализа

Источник разброса	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Главные эффекты	$Q_A = Q_a + Q_b$	$n+m-2$	$s_A^2 = \frac{Q_A}{n+m-2}$
Фактор A	Q_a	$n-1$	$s_a^2 = \frac{Q_a}{n-1}$
Фактор B	Q_b	$m-1$	$s_b^2 = \frac{Q_b}{m-1}$
Остаточное рассеяние	Q_R	$(n-1)(m-1)$	$s_R^2 = \frac{Q_R}{(n-1)(m-1)}$
Полная вариация	$Q = Q_A + Q_R$	$nm-1$	$s^2 = \frac{Q}{nm-1}$

При выполнении гипотезы H_0 об отсутствии эффектов обработки сумма статистик s_A^2 и s_R^2 является несмещенной, состоятельной и эффективной оценкой общей дисперсии $D_x = \sigma_x^2$.

Для проверки нулевой гипотезы дисперсия по фактору A сравнивается с остаточной дисперсией. С этой целью вычисляется F-статистика $F_{набл}^a = \frac{s_a^2}{s_R^2}$, имеющее F-распределение с $n-1$, $(n-1)(m-1)$ степенями свободы. Чем больше различие между эффектами обработки $\tau_1, ..., \tau_n$, тем большую тенденцию к возрастанию проявляет F-статистика. На уровне значимости α гипотеза H_0 отвергается, если $F_{набл}^a > F_\alpha(n-1;(n-1)(m-1))$, F_α - критическое значение. В этом случае влияние фактора A на отклик следует признать значимым. Аналогично проверяется гипотеза об отсутствии влияния фактора B. По F-отношению $F_{набл}^A = \frac{s_A^2}{s_R^2}$ проверяется значимость двухфакторной модели с независимым действием факторов.

2.8. Одна задача применения двухфакторного дисперсионного анализа

Особенности применения двухфакторного дисперсионного анализа рассмотрим на примере исследований потребительской оценки одного товара табачной промышленности (при этом автор целиком согласен с тем, что курение вредит здоровью). Некоторая компания выпустила в продажу модификацию одной из своих марок сигарет. Во время проведения маркетинговых исследований в Екатеринбурге было опрошено 100 респондентов, которые оценили качество сигарет и оформление упаковки по 11-и бальной шкале (от -5 до 5). При этом фиксировали такие характеристики, измеряемые в интервальной шкале как, возраст респондента и его официальный доход за последний год. Числовые результаты опроса приведены в Приложение 4.

Исследование проводилось в 90-х годах. В частности, был сделан вывод о том, что и возраст, ни доход потребителей не оказывают существенного влияния на оценки. Это заключение базировалось на результатах однофакторного дисперсионного анализа.

В этом году было проведено повторное исследование, но уже с привлечением двухфакторной модели. Основанием для повторного анализа послужили тот факт, что при построении графиков зависимости качества и оформления от факторов возраст и доход такая зависимость визуально просматривается (см. рис. 2.8.1), что, однако, противоречит выводам первоначальных исследований.

На этом графике по осям абсцисс и ординат откладывался возраст и доход респондентов, а по оси аппликат суммарная оценка качества и оформления сигарет. Поверхность получена сглаживанием дискретных данных с помощью сплайна.

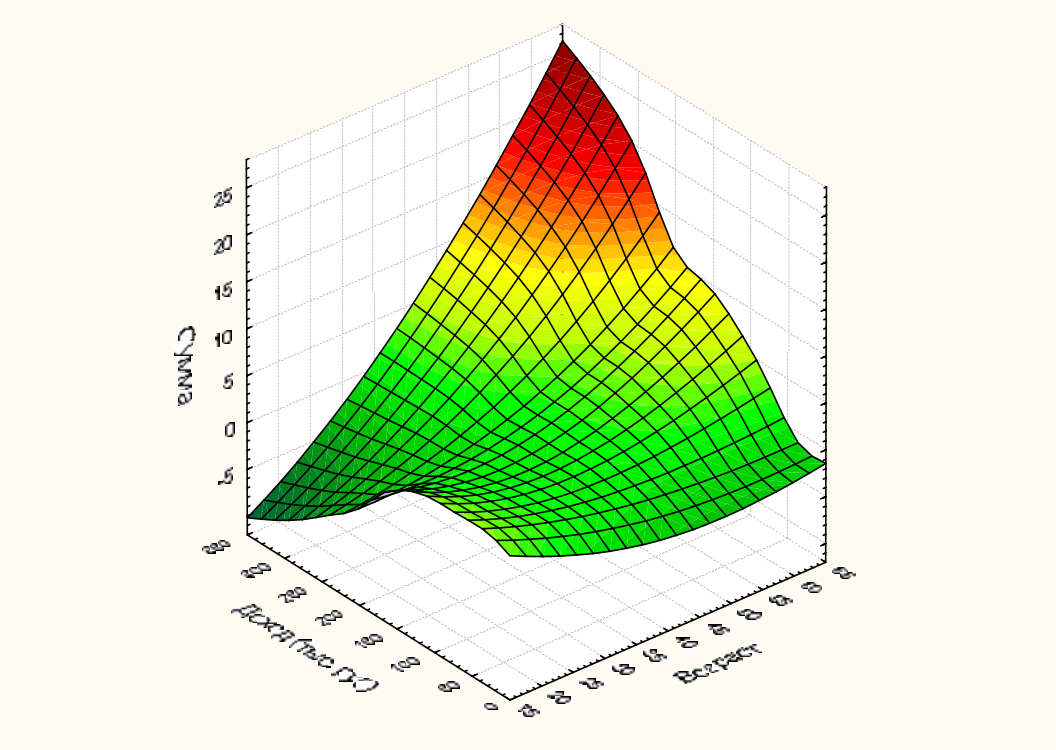


Рис. 2.8.1. Зависимость суммы оценок качества и оформления от возраста и доходов респондентов

Анализ этого графика послужил основанием для выдвижения гипотезы, состоящей в том, что люди старшего возраста и большего достатка отдают предпочтение исследуемой марке сигарет. Для подтверждения этого был проведен повторный однофакторный и после этого двухфакторный дисперсионный анализ данных, который подтвердил справедливость выдвинутой гипотезы.

2.9. Компьютерная реализация двухфакторного дисперсионного анализа

В рассматриваемой задаче влияющими факторами являются возраст и доход респондентов. Сразу же отметим, что коэффициент корреляции (который легко вычисляется с помощью встроенных функций) между ними равен *0,01*, что свидетельствует об отсутствии линейной корреляционной зависимости между факторами.

Для проведения дисперсионного анализа, разделим респондентов на 3 группы по доходу и на 3 группы по возрасту следующим образом. Группы по возрасту определим в следующих пределах: 1 группа - от 19 до 33 лет, 2 группа – 34–44 года, 3 группа – 45-59 лет. Группы по доходам: 1 группа – до 59 тыс. руб., 2 группа – 60-119 тыс. руб., 3 группа – 120-300 тыс. руб. При проведении анализа эти группы буду соответствовать уровням факторов.

Результаты однофакторного дисперсионного анализа при такой постановке задачи приведены в таблицах 2.9.1 и 2.9.2.

Таблица 2.9.1

Результат ОДА зависимости качества от дохода в Ms Excel

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Столбец 1	34	15	0,441176	5,70855615
Столбец 2	39	37	0,948718	5,997300945
Столбец 3	27	8	0,296296	7,37037037

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	8,090582	2	4,045291	0,645479725	0,52664838	3,090186675
Внутри групп	607,9094	97	6,267107			
Итого	616	99				

Таблица 2.9.2

Результат ОДА зависимости оформления упаковки от дохода

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	12,45427	2	6,227134238	0,952874	0,3892	3,090186675
Внутри групп	633,9057	97	6,535110634			
Итого	646,36	99				

Аналогичные результаты демонстрирует анализ оценки качества и оформления в зависимости от возраста респондентов. Соответствующие наблюдаемые значения *F*-статистики оказались равными соответственно *1,851863549* и *2,415426753* при том же критическом значении. Таким образом, подтверждаются выводы предыдущих исследований об отсутствии зависимости оценок от факторов возраста и дохода, взятых в отдельности.

Теперь проведем анализ зависимости оценки качества продукции и оформления упаковки исследуемой марки сигарет от двух факторов одновременно: уровня дохода и возрастной группы респондентов.

Ниже в таблицах представлены результаты двухфакторного дисперсионного анализа (ДфДА) зависимости оценки качества сигарет от возрастной группы и уровня дохода респондентов в Microsoft Excel, Statistica и SPSS.

Таблица 2.9.3

Результат ДфДА в Microsoft Excel

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Выборка	26,72222222	2	13,36111111	2,652573529	0,088765165	3,354131
Столбцы	13,38888889	2	6,694444444	1,329044118	0,281499904	3,354131
Взаимодействие	100,7777778	4	25,19444444	5,001838235	0,003786274	2,727765
Внутри	136	27	5,037037037			
Итого	276,8888889	35				

Таблица 2.9.4

Результат ДфДА в Statistica 6.0

	SS	Degr. of	MS	F	P
Intercept	37,6549	1	37,65492	6,483975	0,012566
"Var1"	18,6215	2	9,31074	1,603260	0,206871
"Var2"	15,2246	2	7,61232	1,310800	0,274646
"Var1"*"Var2"	58,7554	4	14,68886	2,529342	0,045846
Error	528,4717	91	5,80738		

Таблица 2.9.5

Результат ДфДА в SPSS 12.0

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	87,5282726	8	10,94103408	1,883987448	0,072003727
Intercept	37,65491566	1	37,65491566	6,483974728	0,012565543
VAR00001	18,62148399	2	9,310741997	1,603259887	0,206871024
VAR00002	15,22463573	2	7,612317863	1,310800351	0,27464552
VAR00001 * VAR00002	58,75542694	4	14,68885673	2,529342429	0,045846095
Error	528,4717274	91	5,80738162		
Total	652	100			
Corrected Total	616	99			
a	R Squared = ,142 (Adjusted R Squared = ,067)				

В таблице результатов двухфакторного дисперсионного анализа, как и в однофакторном случае, приводятся значения групповых и межгрупповых вариаций (суммы квадратов), числа степеней свободы, оценки дисперсий, наблюдаемое значение статистики - $F_{набл.}$, P-значение и в пакете Microsoft Excel, в отличие от SPSS 12.0 и Statistica 6.0, приводится F-критическое значение статистики. В пакетах SPSS 12.0 и Statistica 6.0 значение F-критическое не приводится.

Для того чтобы сделать вывод о совместном влиянии факторов на изменение случайной величины, необходимо сравнивать $F_{набл}$ и F-критическое, которое в этом случае можно найти в статистических таблицах (см., например [16]).

Следует также обратить внимание на величину *P*-значения (в пакете SPSS 12.0 он обозначается “Sig.”, а в пакете SPSS 12.0 он обозначается “VAR00001 * VAR00002”) в сравнении с уровнем значимости α (α = 0.05).

Как видно из таблиц, *P*-значение меньше уровня значимости α = 0.05, откуда следует, что оценки качества сигарет зависит от возрастной группы и уровня дохода респондентов при рассмотрении влияния этих факторов одновременно.

Далее представлены результаты двухфакторного дисперсионного анализа зависимости оценок оформления упаковки сигарет от возрастной группы и уровня дохода респондентов в Excel, SPSS и Statistica.

Таблица 2.9.6

Результат ДфДА в Microsoft Excel

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Выборка	4,222222222	2	2,111111111	0,298429319	0,744398613	3,354131
Столбцы	11,05555556	2	5,527777778	0,781413613	0,467837158	3,354131
Взаимодействие	84,27777778	4	21,06944444	2,978403141	0,036933412	2,727765
Внутри	191	27	7,074074074			
Итого	290,5555556	35				

Таблица 2.9.7

Результат ДфДА в Statistica 6.0

	SS	Degr. of	MS	F	P
Intercept	16,1297	1	16,12967	2,681874	0,104950
"Var1"	22,5058	2	11,25290	1,871015	0,159842
"Var2"	12,9908	2	6,49540	1,079987	0,343912
"Var1"*"Var2"	60,5913	4	15,14783	2,518623	0,046593
Error	547,3039	91	6,01433		

Таблица 2.9.8

Результат ДфДА в SPSS 12.0

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	99	8,00000	12,3820131	2,058752371	0,048060411
Intercept	16,12967256	1	16,12967256	2,681874213	0,104949752
VAR00004	22,50579546	2	11,25289773	1,87101481	0,159842116
VAR00005	12,99079699	2	6,495398495	1,079987313	0,343912055
VAR00004 * VAR00005	60,59130293	4	15,14782573	2,518622933	0,046593203
Error	547,3038952	91	6,014328519		
Total	664	100			
Corrected Total	646,36	99			
a	R Squared = ,153 (Adjusted R Squared = ,079)				

Из таблиц результатов видно, что оценки оформления упаковки сигарет зависит от возрастной группы и уровня дохода респондентов при рассмотрении влияния этих факторов одновременно.

Отличие результатов анализа Microsoft Excel от результатов пакетов Statistica 6.0 и SPSS 12.0 обусловлены ограничениями работы Microsoft Excel. В Excel встроены процедуры двухфакторного дисперсионного анализа с повторениями и без повторений.

В рассматриваемой задаче мы использовали первую из этих процедур. Эта процедура требует, что бы для анализа была сформирована таблица с выборками, в каждой из которых одинаковое количество значений. Из-за этого ограничения для анализа пакетом Microsoft Excel формируется таблица, в которую попадает лишь часть исходных значений (36 из 100).

Таблица формируется следующим образом: находим выборку, в которой наименьшее количество значений (такой выборкой является 3 группа по возрасту – 3 группа по доходу, в ней 4 значения), из каждой выборки произвольным образом выбираем одинаковое количество значений (равное 4-ем).

В качестве примера ниже приведены таблицы исходных данных и таблица подготовленная для анализа зависимости оценки качества от возрастной группы и уровня дохода респондентов в Excel.

Таблица 2.9.9

Подготовка таблицы для проведения анализа в Excel

Исходная таблица

Возраст \ доход	1	2	3
1	5	-1	-2
	3	-3	-2
	3	-1	-2
	-1	-2	-3
	1	-5	-1
	-1	3	3
	0	3	2
	1	0	4
	2	-2	0
	0	2	-2
	0	2	3
	0	1	
	0		
	3		
	2		
	-2		
2	0	-1	-4
	2	-1	-5
	0	2	0
	3	4	3
	-5	1	-4
	-1	-2	1
	-3	3	2
	3	3	-1
	2	0	1
	3	1	4
	0	4	3
	-3	2	-2
	2	2	
		2	
		3	
		2	
		1	
		-5	
		-3	

Таблица для анализа в Microsoft Excel

Возраст \ доход	1	2	3
1	5	-1	-2
	3	-3	-2
	3	-1	-2
	-1	-2	-3
2	0	-1	-4
	2	-1	-5
	0	2	0
	3	4	3
3	-3	2	3
	1	3	1
	-5	3	3
	3	5	3

3	-3	2	3
	1	3	1
	-5	3	3
	3	5	3
	0	2	
		3	
		3	
		1	

Несмотря на эти расхождения, результаты анализа показывают, что *P*-значение меньше 0.05, т.е. приходим к тем же выводам о зависимостях.

Таким образом, однофакторный анализ не позволяет утверждать, что показатели зависят от возрастной группы и уровня дохода респондентов принятых к рассмотрению по отдельности, однако, результаты двухфакторного анализа указывают на обратное при рассмотрении влияния этих двух некоррелированных факторов одновременно.

При маркетинговых и других подобных исследованиях надо учитывать и анализировать воздействие не только каждого фактора по отдельности, но и влияние взаимодействия этих факторов на показатели.

Глава 3. КЛАСТЕРНЫЙ АНАЛИЗ

Цель кластерного анализа – систематизация множества исследуемых объектов (признаков) с выделением однородных групп, или кластеров (от англ. «cluster» – гроздь, пучок, скопление). Этот метод дает инструмент для классификации данных и выявления в них соответствующей структуры. Кластерный анализ можно применять в самых различных случаях, даже когда речь идет о простой группировке, в которой все сводится к образованию групп по признаку количественного сходства.

Кластерный анализ (термин впервые появился в научной литературе в середине XX века) включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры. В отличие от многих других статистических процедур, методы кластерного анализа используются в большинстве случаев тогда, когда не имеется каких-либо априорных гипотез относительно классификации.

Кластерный анализ – это описательная (не параметрическая) процедура, он не позволяет сделать никаких статистических выводов, но дает возможность провести своеобразную «разведку» – изучить структуру совокупности. Кластерный анализ можно проводить циклически – до тех пор, пока не будут получены необходимые результаты. При этом каждый цикл может давать информацию, которая способна изменить направление исследования и подходы к дальнейшему применению кластерного анализа. Этот процесс можно представить системой с обратной связью.

Кластерный анализ позволяет рассматривать большой (хотя и имеющий ограничения) объем информации и резко сокращать массивы информации, делать их компактными и наглядными.

Существенную роль этот метод играет при анализе совокупностей временных рядов, характеризующих экономическое развитие (например, общехозяйственной и товарной конъюнктуры). Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа. В задачах социально-экономического прогнозирования весьма перспективно сочетание кластерного анализа с другими количественными методами (например, с факторным, дисперсионным и регрессионным анализом).

Большое достоинство кластерного анализа состоит в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ, в отличие от большинства математико-статистических методов, не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество практически произвольно взятых исходных данных.

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения. В частности, состав и количество кластеров зависит от выбираемых критериев вариативности (мер разброса). При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов за счет замены их характеристиками обобщенных значений – параметров кластера. При классификации объектов очень часто игнорируется возможность отсутствия в рассматриваемой совокупности каких-либо характеристик кластеров.

В кластерном анализе по умолчанию принимается, что:

а) выбранные характеристики в принципе допускают желательное разбиение на кластеры (так называемая проблема выбора свойств или характеристик объектов). Вообще, предполагается, что эта проблема решена до начала процесса кластеризации;

б) единицы измерения (масштаб) выбраны правильно (обычно данные нормализуют вычитанием среднего и делением на стандартное отклонение – то есть производят обычную процедуру статистической стандартизации, – так что математическое ожидание оказывается равным нулю, а дисперсия – единице).

3.1. Постановка задачи

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов $X_i \in X$ на m (m – целое) кластеров (подмножеств) $Q_1, Q_2, \ldots, Q_m$ так, чтобы каждый объект X_j принадлежал одному и только одному подмножеству разбиения (кластеру Q_l) и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, а объекты, принадлежащие разным кластерам, – разнородными.

Пусть, например, X – это перечень из n стран, выбранных для сравнительного анализа. Каждая из этих стран может быть охарактеризована набором экономических показателей: ВНП на душу населения, число автомашин на тысячу человек, душевое потребление электроэнергии, душевое потребление стали и т.д. Тогда X_1 (вектор измерений) представляет собой набор указанных характеристик для первой страны: $X_1 = (x_{1,1}, x_{1,2}, x_{1,3}, x_{1,4})$, X_2 – для второй, X_3 – для третьей, и т.д. Задача заключается в том, чтобы сгруппировать страны по уровню развития с учетом этих и только этих показателей.

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией.

Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

$$W = \sum_{j=1}^{n} (x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2,$$

где x_j – измерения j-го объекта.

Основными понятиями кластерного анализа являются понятия сходства и разнородности. Именно с их помощью и производится группировка объектов.

Понятно, что объекты i-й и j-й попали бы в один кластер, если бы расстояние (разнородность, различие) между точками X_i и X_j было достаточно маленьким, и в разные кластеры, если бы это расстояние было достаточно большим.

Таким образом, попадание в один или разные кластеры объектов определяется понятием расстояния между X_i и X_j, введенном в p-мерном пространстве признаков, по которым сравниваются элементы. Неотрицательная функция $d(X_i, X_j)$ называется функцией расстояния (метрикой), если:

а) $d(X_i, X_j) \geq 0$ для всех X_i и X_j,

б) $d(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$,

в) $d(X_i, X_j) = d(X_j, X_i)$,

г) $d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$, где X_j; X_i *и* X_k – любые три вектора.

При проведении кластерного анализа чаще других употребляются следующие функции расстояний:

евклидово расстояние:

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{p}\left(x_{ki} - x_{kj}\right)^2\right]^{\frac{1}{2}}; \qquad (3.1.1)$$

хэммингово расстояние:

$$d_1(X_i, X_j) = \left[\sum_{k=1}^{p}\left|x_{ki} - x_{kj}\right|\right];$$

чебышевское расстояние:

$$d_\infty(X_i, X_j) = \max_k \left\{\left|x_{ki} - x_{kj}\right|\right\};$$

l_p-расстояние:

$$d_p(X_i, X_j) = \left[\sum_{k=1}^{p} \left| x_{ki} - x_{kj} \right|^p \right]^{1/p}.$$

Наиболее популярна евклидова метрика. Метрика Хэмминга наиболее проста для вычислений. Чебышевское расстояние легко считается и включает в себя процедуру упорядочения, а l_p-расстояние охватывает первые две функции расстояний. Если признакам приписывается разный вес, то эти веса можно учесть при вычислении расстояния.

Исходные данные измерений состоят из n векторов $X_1, X_2, \ldots, X_n$ размерности p. Их удобно представить в виде $p \times n$ матрицы:

$$\chi = \begin{pmatrix} x_{11} & x_{12} & \ldots x_{1n} \\ x_{21} & x_{22} & \ldots x_{2n} \\ x_{p1} & x_{p2} & \ldots x_{pn} \end{pmatrix} = \left(X_1, X_2, \ldots, X_n \right). \qquad (3.1.2)$$

Расстояния между парами векторов $d(X_i, X_j)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & d_{12} & \ldots d_{1n} \\ d_{21} & 0 & \ldots d_{2n} \\ d_{n1} & d_{n2} & \ldots\ 0 \end{pmatrix}.$$

Двойственным к понятию расстояния является понятие сходства [10] между объектами. Неотрицательная вещественная функция $S(X_i; X_j) = S_{ij}$ называется мерой сходства, если:

а) $0 \le S(X_i, X_j) < 1$ *для* $X_i \ne X_j$,

б) $S(X_i, X_i) = 1$,

в) $S(X_i, X_j) = S(X_j, X_i)$.

Пары значений мер сходства векторов $X_1, X_2, \ldots, X_n$ объединяются в симметричную (свойство «в») матрицу сходства:

$$S = \begin{pmatrix} 1 & s_{12} & \ldots & s_{1n} \\ s_{21} & 1 & \ldots & s_{2n} \\ s_{n1} & s_{n2} & \ldots & 1 \end{pmatrix}.$$

Величину $s_{ij} = S(X_i, X_j)$ называют коэффициентом сходства между векторами X_i и X_j. Чем ближе к единице значение сходства между векторами X_i и X_j, тем более «схожи» эти векторы. Наоборот, чем коэффициент сходства ближе к нулю, тем разнороднее эти векторы.

Как в случае определения разнородности используются различные определения расстояния, так для определения сходства могут быть взяты различные определения [10] коэффициентов сходства (лишь бы они удовлетворяли перечисленным выше аксиомам). Однако чаще всего применяются коэффициенты сходства, определенные на основании коэффициентов корреляции

$$r_{ij} = \frac{\sum_{h=1}^{n} \left(x_{hi} - \bar{x}_i \right) \left(x_{hj} - \bar{x}_j \right)}{\sigma_i \sigma_j},$$

где m_i, m_j, σ_i, σ_j соответственно – средние и среднеквадратичные отклонения для характеристик X_j и X_i (если только обе величины измерены в интервальных шкалах).

Таковыми являются модуль коэффициента корреляции $|r_{ij}|$ или квадрат коэффициента корреляции r_{ij}^2 – коэффициент детерминации [6].

3.2. Подготовка данных для кластеризации

Кластерный анализ можно применять к интервальным данным, частотам, бинарными данным (то есть к данным, измеренным в любой шкале [23]). Важно, чтобы переменные изменялись в сравнимых шкалах.

Неоднородность единиц измерения и вытекающая отсюда невозможность обоснованного выражения значений различных показателей в одном масштабе приводит к тому, что величина расстояний между точками, отражающими положение объектов в пространстве их свойств, оказывается зависящей от произвольно избираемого масштаба. Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, то есть выражаются через отношение этих значений к некоторой величине, отражающей определенные свойства данного показателя.

Нормирование исходных данных для кластерного анализа иногда проводится посредством деления исходных величин на среднеквадратичное отклонение соответствующих показателей. Другой способ сводится к вычислению так называемого стандартизованного вклада. Его еще называют *Z*-вкладом.

Z-вклад показывает, сколько стандартных отклонений отделяет данное наблюдение от среднего значения:

$$Z_i = \frac{x_i - \bar{x}}{\sigma}, \qquad (3.2.1)$$

где x_i – значение данного наблюдения, $\bar{x}$ – среднее, σ – стандартное отклонение. Среднее для *Z*-вкладов равно нулю, а стандартное отклонение равно 1.

Стандартизация позволяет сравнивать наблюдения над случайными величинами, которые распределены по разным законам. Если распределение переменной является нормальным (или близким к нормальному), а средняя и дисперсия известны или оцениваются по большим выборкам, то *Z*-вклад обеспечивает более специфическую информацию о расположении объекта.

Применение методов нормирования означает признание всех характеристик равнозначными с точки зрения выяснения сходства рассматриваемых объектов. Такое признание равноценности различных показателей кажется оправданным отнюдь не всегда. В ряде задач оказывается желательным, наряду с нормированием, придать каждому из показателей вес, отражающий его значимость в ходе установления сходств и различий объектов.

Определить вес отдельных показателей можно с помощью опроса экспертов. Экспертные оценки дают известное основание для определения важности индикаторов, входящих в ту или иную группу показателей. Довольно часто при решении подобных задач используют не один, а два расчета: первый – при котором все признаки считаются равнозначными; второй – когда им придаются различные веса в соответствии со средними значениями экспертных оценок.

Кроме перечисленных выше, используются и другие методы стандартизации:

а) Разброс от –1 до 1. Линейным преобразованием переменных добиваются разброса значений от –1 до 1.

б) Разброс от 0 до 1. Линейным преобразованием переменных добиваются разброса значений от 0 до 1.

в) Максимум 1. Значения переменных делятся на их максимум.

г) Среднее 1. Значения переменных делятся на их среднее.

Кроме того, возможны преобразования самих расстояний. В частности, можно все расстояния преобразовать так, чтобы они изменялись от 0 до 1, как это делается при введении сходства через коэффициенты корреляции.

3.3. Иерархический кластерный анализ. Алгоритм последовательной кластеризации

Алгоритмы кластерного анализа разнообразны. Все их можно подразделить на иерархические и неиерархические. Иерархические алгоритмы связаны с построением дендрограмм и делятся на агломеративные, характеризуемые последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров, и дивизимные (делимые), в которых число кластеров возрастает, в результате чего образуется последовательность расщепляющихся групп.

В настоящей работе рассматриваются иерархические агломеративные алгоритмы, или, иначе, алгоритмы агрегирования. Их смысл заключается в следующем: перед началом кластеризации все объекты считаются отдельными кластерами (по одному элементу в каждом кластере), которые в ходе реализации алгоритма объединяются. Вначале выбирается пара ближайших многомерных элементов, которые объединяются в кластер; в результате количество кластеров становится равным (n – 1). Процедура повторяется: либо вновь объединяются два элемента, либо элемент добавляется в уже существующий ближайший кластер.

Так продолжается до тех пор, пока не объединятся все кластеры, то есть пока не получится единственный кластер, содержащий все элементы. На любом этапе объединение можно прервать, получив нужное число кластеров. Результат работы алгоритма агрегирования зависит от способов вычисления расстояния между объектами и определения близости между кластерами.

Процедура иерархического кластерного анализа предусматривает группировку как объектов (строк матрицы данных), так и переменных (столбцов). Можно считать, что в последнем случае роль объектов играют переменные, а роль переменных – столбцы.

Агломеративный кластерный анализ допускает применение различных методов кластеризации, некоторые из них будут рассмотрены ниже. Различие между этими методами состоит в способе вычисления степени близости элемента к кластеру. Несмотря на это, все они базируются на едином алгоритме последовательной кластеризации.

На первом этапе последовательной кластеризации все n многомерных объектов ($X_1, X_2, \ldots, X_n$), подлежащих кластеризации, разбиваются на n различных кластеров (по одному элементу в каждом кластере): $\{X_1\}$, $\{X_2\}$, …, $\{X_n\}$. Среди всех этих объектов-кластеров выбираются два наиболее близкие друг к другу в одном из указанных выше смыслах (п. 3.1) и объединяются в один кластер. Пусть объединенные объекты – это X_i и X_j. Новое множество кластеров состоит уже из (n – 1) кластера: $\{X_1\}$, $\{X_2\}$, …, $\{X_i, X_j\}$, …, $\{X_n\}$. Повторяя процесс, получим последовательные множества, состоящие из n – 2, n – 3 и т.д. кластеров. Естественное окончание процедуры наступает тогда, когда все объекты будут сведены в единственный кластер: $\{X_1, X_2, \ldots, X_n\}$.

В качестве исходного числового материала для последовательной кластеризации берется матрица расстояний или сходств, в зависимости от существа рассматриваемой проблемы. При этом сами матрицы принято преобразовывать в верхние (или нижние) треугольные таблицы. В силу симметричности матрицы элементы, находящиеся ниже (выше) диагонали, для реализации алгоритма не представляются необходимыми и не указываются. Так, если кластеризация осуществляется по расстояниям, то исходным числовым материалом будет предварительно вычисленная таблица следующего вида (табл. 3.3.1).

Таблица 3.3.1

Таблица расстояний

	X_1	X_2	X_3	...	X_n
X_1	0	d_{12}	d_{13}	...	d_{1n}
X_2		0	d_{23}	...	d_{2n}
X_3			0	...	d_{3n}
...				...	...
X_n					0

Исторически первой схемой последовательной кластеризации является схема объединения без пересчета. При реализации этой схемы степень близости объектов и кластера определяется как близость каждого объекта к ближайшему из объектов, уже принадлежащих кластеру. Рассмотрим реализацию этого алгоритма на простом примере.

Пример. В качестве объектов исследования рассмотрим четыре различных тарифных плана компании МТС. Каждый план охарактеризуем несколькими признаками (абонентская плата, количество оплаченных минут, стоимость минуты разговора). Таблица сходств (под сходством понимаем модуль коэффициента корреляции, который легко подсчитать с использованием «Excel») будет иметь следующий вид (табл. 3.3.2).

Таблица 3.3.2

Таблица сходств четырех тарифных планов МТС

		«Оптима+50»	«Оптима-Вечер»	«Оптима+100»	«Оптима-Унив.»
1	«Оптима+50»	1			
2	«Оптима-Вечер»	0,159647519	1		
3	«Оптима+100»	0,99973666	0,182283333	1	
4	«Оптима-Унив.»	0,156782137	0,999963128	0,179424634	1

На первом этапе объединим элементы, имеющие максимальное сходство, то есть 4 и 2 («Оптима-Универсал» и «Оптима-Вечер»): {1}, {2, 4}, {3}. Среди оставшихся наибольшее сходство имеют элементы 3 и 1. Объединив их, получаем: {1, 3}, {2, 4}. Следующий этап завершает кластеризацию, ибо объединению подлежат элементы 3 и 2, входящие в разные кластеры: {1, 2, 3, 4}. Следует заметить, что разумно было бы закончить кластеризацию на втором шаге, поскольку полученные классы имеют высокий уровень схожести (0,999963128 и 0,99973666), тогда как на третьем шаге – всего 0,182283333.

В этом случае был бы получен результат, легко интерпретируемый с практической точки зрения. Кроме того, полезно сравнить результаты кластеризации с группировкой, проведенной на основании применения метода многомерной средней (см. п. 1.3).

Схема кластеризации без пересчета существенно упрощает процедуру вычисления, ибо весь расчет производится на основании одной исходной таблицы мер близости. Однако эта схема не корректна, ибо у кластера (уже сформированного объекта) отсутствует единая мера сравнения с остальными объектами. Это приводит к необходимости определять некоторую центральную характеристику положения кластера и пересчитывать ее для каждого вновь образованного кластера.

Пусть исходная таблица расстояний для набора из n векторов задается таблицей 3.3.1. Допустим, число d_{ij} оказалось наименьшим, тогда векторы X_i и X_j объединяются в кластер $\{X_i, X_j\}$. Построим новую $((n-1) \times (n-1))$ таблицу расстояний (табл. 3.3.3).

Таблица 3.3.3

Модифицированная таблица расстояний

	$\{X_i, X_j\}$	X_1	X_2	X_3	…	X_n
$\{X_i, X_j\}$	0	$\hat{d}_{11}$	$\hat{d}_{12}$	$\hat{d}_{13}$	…	$\hat{d}_{1n}$
X_1		0	d_{12}	d_{13}	…	d_{1n}
X_2			0	d_{23}	…	d_{2n}
X_3				0	…	d_{3n}
…						…
X_n						0

В табл. 3.3.3 нижние (n – 2) строки взяты без изменения из предыдущей таблицы (таблица 3.3.1), а первая строка вычислена заново. Вычисление расстояния между кластерами, произведенное разными способами, приводит к алгоритмам кластеризации с различными свойствами. Например, можно положить расстояние между $\{X_i, X_j\}$ и X_k равным среднему арифметическому из расстояний между элементами X_i, X_j и X_k: $d_{i+j,k} = 1/2\ (d_{ik} + d_{jk})$, или определить $d_{i+j,k}$ как минимальное из этих двух расстояний: $d_{i+j,k} = \min(d_{ik,}d_{jk})$.

Довольно широкий класс алгоритмов может быть получен, если для перерасчета расстояний использовать следующую общую формулу [10]:

$$d_{i+j,k} = A(w)\ \min(d_{ik}\, d_{jk}) + B(w)\ \max(d_{ik}\, d_{jk}),$$

$$\text{где } A(w) = \begin{cases} \dfrac{wn_i}{wn_i + n_j}, & d_{ik} \le d_{jk} \\ \dfrac{wn_j}{wn_i + n_j}, & d_{ik} > d_{jk} \end{cases}, \qquad B(w) = \begin{cases} \dfrac{n_i}{wn_i + n_j}, & d_{ik} \le d_{jk} \\ \dfrac{n_j}{wn_j + n_i}, & d_{ik} > d_{jk} \end{cases},$$

здесь n_i и n_j – число элементов в кластерах i и j, а w – свободный параметр, выбор которого определяет конкретный алгоритм. Например, при $w = 1$ мы получаем так называемый, алгоритм «средней связи», для которого формула перерасчета расстояний принимает вид:

$$d_{i+j,k} = \frac{n_i}{n_i + n_j} d_{ik} + \frac{n_j}{n_i + n_j} d_{jk}.$$

В данном случае расстояние между двумя кластерами на каждом шаге работы алгоритма оказывается равным среднему арифметическому из расстояний между всеми такими парами элементов, что один элемент пары принадлежит к одному кластеру, другой – к другому.

Наглядный смысл параметра w становится понятным, если принять $w \to \infty$. Формула пересчета расстояний приобретает вид: $d_{i+j,k} = \min(d_{i,k}\ d_{jk})$. Такой метод кластеризации носит название алгоритма «ближайшего соседа». Он позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов.

В данном случае расстояние между двумя кластерами на каждом шаге работы алгоритма оказывается равным расстоянию между двумя самыми близкими элементами, принадлежащими к этим двум кластерам. Таким образом, мы вернулись к кластеризации без пересчета, как в методе, рассмотренном выше.

3.4. Графическое представление результатов кластеризации

Любой формальный алгоритм становится более понятным и удобным для использования, если его аналитическую часть можно дополнить графической интерпретацией. Кластерный анализ дает такую возможность.

Очевидно, что совокупности многомерных объектов можно представить в виде точек в пространстве соответствующей размерности. Сам процесс объединения может быть изображен как соединение этих точек отрезками. Однако создать такой наглядный образ можно лишь в случае, если объединяемые векторы двух- или трехмерны.

Другая идея графической интерпретации связана с изображением самого процесса кластеризации, а не исходных образов. На плоскости будем изображать последовательные объединения, определенные на основании анализа, состояния матрицы расстояний или сходства. Дендрограмму (граф последовательного объединения; от греч. «dendron» – дерево) можно опреде-

лить как графическое изображение результатов процесса последовательной кластеризации, которая осуществляется в терминах матрицы расстояний или сходств [20].

Для построения дендрограммы (рис. 3.4.1) следует расположить все объекты (*A, B, C, D, F*) вдоль одной из осей, а вдоль другой – разместить линейку расстояний или сходств. По мере объединения элементов в кластеры объекты соединяются на уровне того расстояния, на котором произведено объединение.

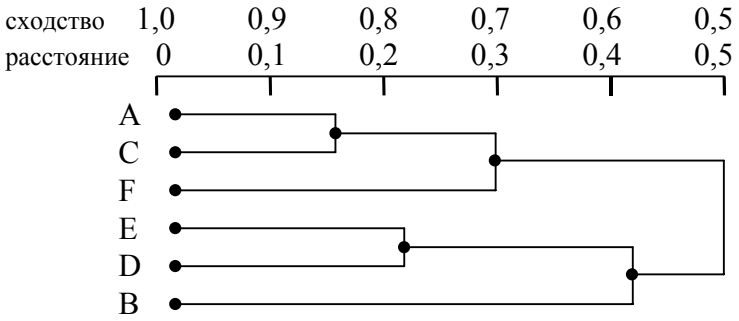


Рис. 3.4.1. Дендрограмма

Объекты *A* и *C* наиболее близки и поэтому объединяются в один кластер на уровне сходства, равном 0,85. Объекты *D* и *E* объединяются на уровне 0,8. Ограничиваясь объединением до уровня сходства 0,8, получим четыре кластера: {*A*, *C*}, {*F*}, {*D*, *E*}, {*B*}. Далее образуются кластеры {*A*, *C*, *F*} и {*E*, *D*, *B*}, соответствующие уровню сходства, равному 0,7 и 0,6. В один кластер все объекты группируются на уровне 0,5.

Вид дендограммы зависит от выбора меры сходства или расстояния между объектом и кластером и от метода кластеризации. Но наиболее важным моментом является выбор меры сходства или меры различия между объектом и кластером.

3.5. Определение числа кластеров

При разделении многомерных объектов на кластеры возможны два подхода.

Первый подход связан с тем, что изначально задается требуемое число классов. В этом случае процесс кластеризации продолжается до тех пор, пока не будет сформировано заранее заданное количество групп.

Другой подход связан с введением некоторого порогового значения: объединение производится до тех пор, пока уровень сходства (различия) не достигнет некоторого критического значения. Процесс объединения завершается, когда это критическое значение достигнуто. Иными словами, объекты объединяются в кластеры, пока расстояние между ними меньше, чем заданное. Преимущество такого метода заключается в том, что результаты легко визуализировать в виде дендрограммы.

Один из методов определения оптимального числа кластеров с помощью заданных критических значений базируется на применении так называемой мерой принадлежности $S(\alpha, \beta)$ Хользенгера – Хармана [10]. Принадлежность $S(\alpha, \beta)$ – функция двух аргументов, где α – вероятность того, что найдено наилучшее разбиение, β – доля наилучших разбиений в общем числе возможных. Критические значения задаются по таблице (см. Приложение 1).

Другой метод определения оптимального числа кластеров связан с анализом изменения некоторой тестовой функции.

В качестве такой функции можно принять, например, сумму квадратов отклонений элементов E_j внутри (j-го)кластера $r_{ij}^2 = (x_{ij} - \bar{x}_j)^2$:

$$E_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} r_{ij}\right)^2,$$

играющую роль внутригрупповой дисперсии. Процессу группировки соответствует последовательное возрастание значения величины E_j. Наличие резкого скачка в значении E_j можно интерпретировать как сигнал для прекращения кластеризации. Действительно, такой скачок определяется фазовым переходом от сильносвязанного к слабосвязанному состоянию объектов.

3.6. Методы минимальной дисперсии

Существует много методов кластерного анализа. В данной работе рассматриваются лишь так называемые методы минимальной дисперсии [20], поскольку именно они чаще представлены в компьютерных пакетах статистической обработки данных.

Общим для этих методов является анализ матрицы расстояний (сходств) размерностью $n \times n$ (где n – число объектов) с пошаговым последовательным объединением пар наиболее схожих объектов (отдельных векторов или кластеров). Различие заключается в способе измерения расстояния.

Способ измерения расстояния определяет особенности получаемых кластеров. Например, при использовании метода «ближайшего соседа» объединяются те кластеры, которые имеют наименьшее расстояние между периферийными объектами. Центроидный метод определяет расстояние между кластерами как расстояние между их «центрами тяжести».

Рассмотрим основные методы минимальной дисперсии, сопровождая описание каждого метода построением соответствующей дендрограммы. За основу будет взят один и тот же исходный числовой материал – набор из семи двумерных объектов: X_1= (1, 1), X_2 = (3, 1), X_3 = (1, 2), X_4 = (5, 2), X_5 = (6, 2), X_6 = (2, 5), X_7 = (3, 5).

На рис. 3.6.1 приведено расположение точек на плоскости, соответствующее выбранным объектам.

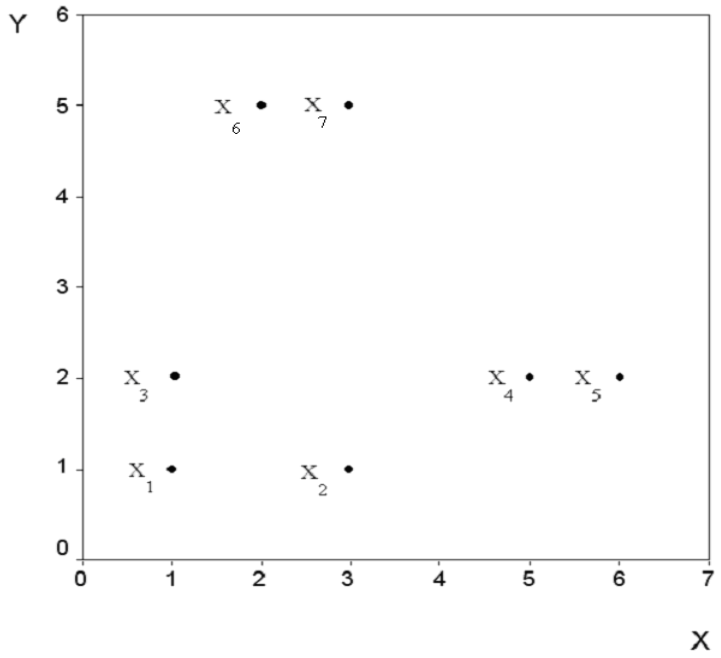


Рис. 3.6.1. Расположение на плоскости семи двумерных объектов

В качестве меры различия выберем евклидово расстояние, определяемое по (3.1.1). Таблица расстояний будет иметь вид:

Таблица 3.6.1

Таблица расстояний для семи двумерных объектов

	1	2	3	4	5	6	7
1	0,00	2,00	1,00	4,12	5,10	4,12	4,47
2		0,00	2,24	2,24	3,16	4,12	4,00
3			0,00	4,00	5,00	3,16	3,61
4				0,00	1,00	4,24	3,61
5					0,00	5,00	4,24
6						0,00	1,00
7							0,00

3.6.1. Метод полных связей, или метод «дальнего соседа» (Complete linkage, Furthest neighbour)

Суть данного метода состоит в том, что два объекта объединяются в группу (кластер), если они имеют коэффициент сходства, меньший некоторого порогового значения. В терминах евклидова расстояния это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения R, определяющего максимально допустимый диаметр подмножества, которое образует кластер.

Расстояние между двумя кластерами определяется по принципу «дальнего соседа», то есть как максимальное расстояние между всеми возможными парами элементов из этих кластеров. При реализации этого метода, как правило, формируются относительно компактные гиперсферические кластеры, состоящие из объектов с большим сходством.

Приняв в качестве порогового уровня объединения сходство с расстоянием между объектами не более $R = 1$, получим для рассматриваемого примера следующие группы: $\{X_1, X_3\}$, $\{X_2\}$, $\{X_4, X_5\}$, $\{X_6, X_7\}$. Если в качестве порогового уровня для объединения в кластер выбрать большее расстояние, например $R = 2{,}5$, то, очевидно, кластеры будут иными: $\{X_1, X_2, X_3\}$, $\{X_4, X_5\}$, $\{X_6, X_7\}$.

Дендрограмма кластеризации, проведенной по методу полных связей, представлена на рис. 3.6.2

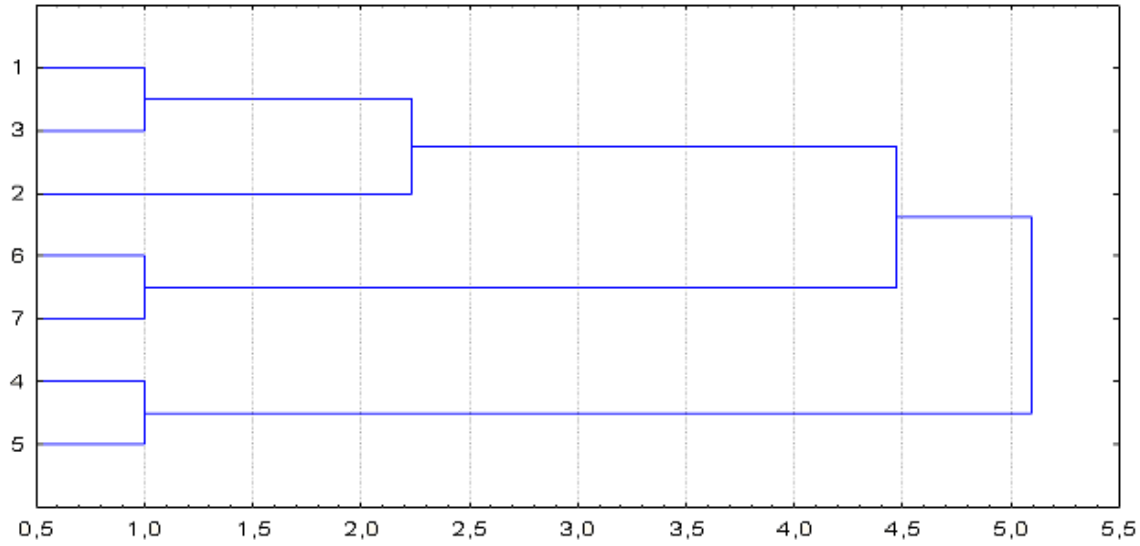


Рис. 3.6.2. Дендрограмма кластеризации по методу полных связей

3.6.2. Метод одиночной связи, или метод «ближайшего соседа» (Single linkage, Nearest neighbour)

Алгоритм этого метода схож с алгоритмом метода полных связей. Отличие заключается лишь в том, что расстояние между кластерами определяется как минимальное расстояние между всеми возможными парами элементов из этих кластеров («ближайшие соседи»). Преимущество метода одиночной связи [10, 20] заключается в его математических свойствах: полученные результаты инвариантны к монотонным преобразованиям матрицы сходства.

Применению метода не мешает наличие «совпадений» в данных. Это означает, что метод одиночной связи является одним из немногих, результаты применения которых не изменяются при любых преобразованиях данных, оставляющих без изменения относительное упорядочение элементов матрицы сходства.

Недостаток метода одиночной связи состоит в том, что он приводит к появлению «цепочек», то есть к образованию больших продолговатых кластеров.

По мере приближения к окончанию процесса кластеризации образуется один большой кластер, а все остающиеся объекты добавляются к нему один за другим.

По данным примера получаем дендрограмму (рис. 3.6.3):

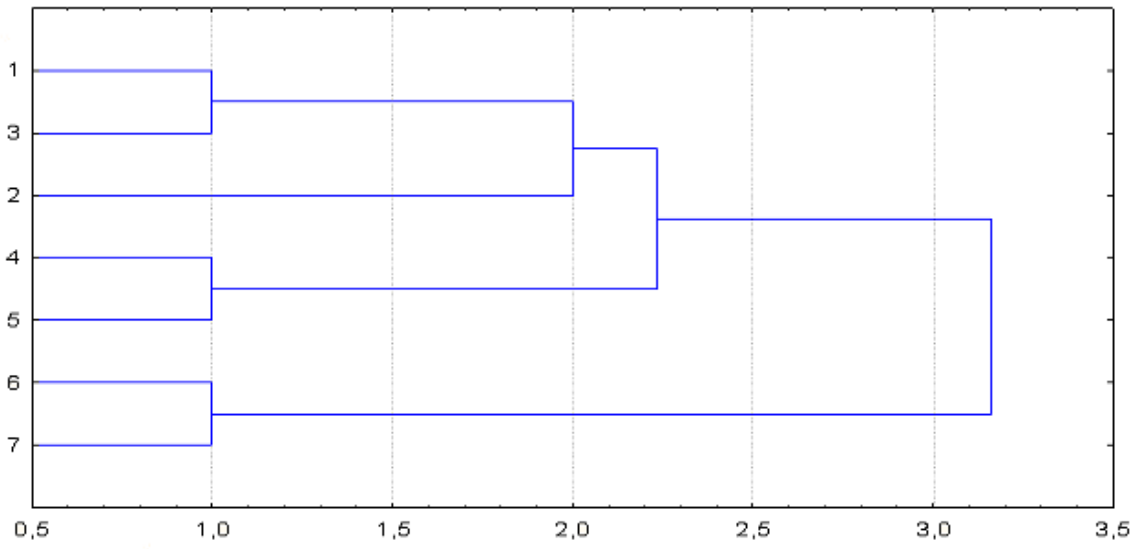


Рис. 3.6.3. Дендрограмма кластеризации по методу одиночной связи

3.6.3. Метод невзвешенного попарного среднего (Unweighted pair-group average, Between-groups linkage)

Расстояние между кластерами определяется по принципу «средней связи», то есть новый элемент, включаемый в кластер, обладает наименьшим средним расстоянием до уже содержащихся в кластере элементов. Метод эффективен, когда объекты в действительности формируют кластеры разных размеров, однако он работает одинаково хорошо и в случаях протяженных («цепочного» типа) кластеров.

Используя данные примера, при методе невзвешенного попарного среднего получим дендрограмму, приведенную на рис. 3.6.4:

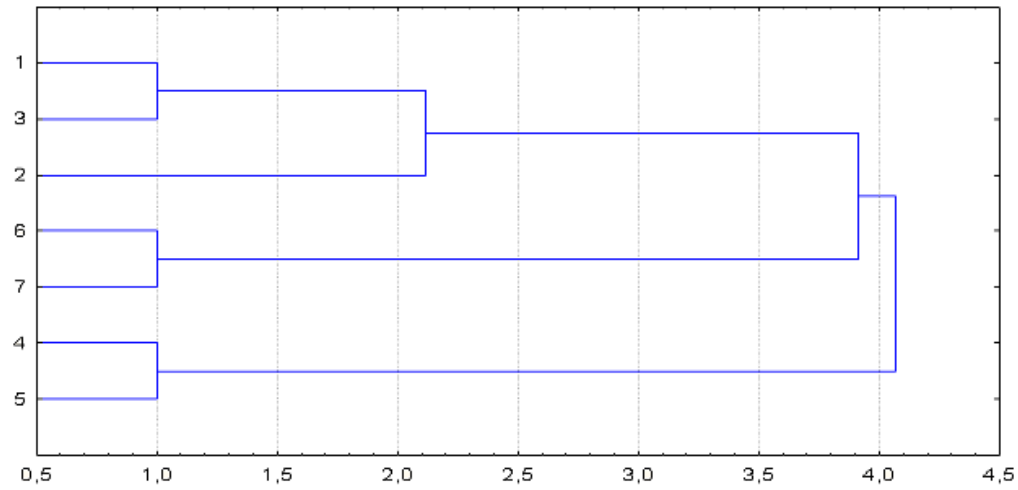


Рис. 3.6.4. Дендрограмма кластеризации по методу невзвешенного попарного среднего

3.6.4. Метод Варда (Ward's method)

Этот метод отличается от остальных тем, что в нем используются методы дисперсионного анализа для оценки расстояний между кластерами. В качестве расстояния $dis(X,Y)$ между кластерами X и Y берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения:

$$dis(X,Y) = \frac{n_x n_y}{n_x + n_y}(\overline{X} + \overline{Y})^T(\overline{X} + \overline{Y}),$$

где радиусы-векторы $\overline{X}$, $\overline{Y}$ центров кластеров, n_x, n_y – число элементов в них, верхний индекс T означает транспонирование.

Метод Варда минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и имеет тенденцию к нахождению (или созданию) кластеров приблизительно равных размеров и имеющих гиперсферическую форму. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

Дендрограмма, построенная по методу Варда на данных рассматриваемого примера, приведена на рис. 3.6.5:

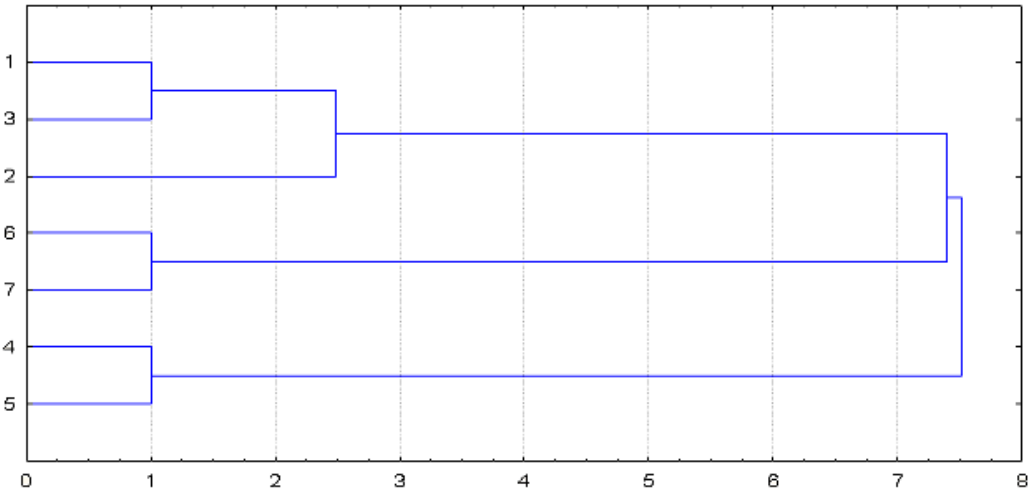


Рис. 3.6.5. Дендрограмма кластеризации по методу Варда

3.6.5. Невзвешенный центроидный метод (Unweighted pair-group centroid, Centroid clustering)

Расстояние между двумя кластерами определяется как евклидово расстояние между центрами (средними, «центрами тяжести») этих кластеров. Кластеризация идет поэтапно, при этом на каждом из шагов объединяют два кластера, имеющие минимальное расстояние между центрами. Следует заметить, что в случае объединения двух кластеров, в одном из которых много элементов, а в другом – мало, характеристики последнего практически игнорируются.

Дендрограмма для рассматриваемого примера приведена на рис. 3.6.6:

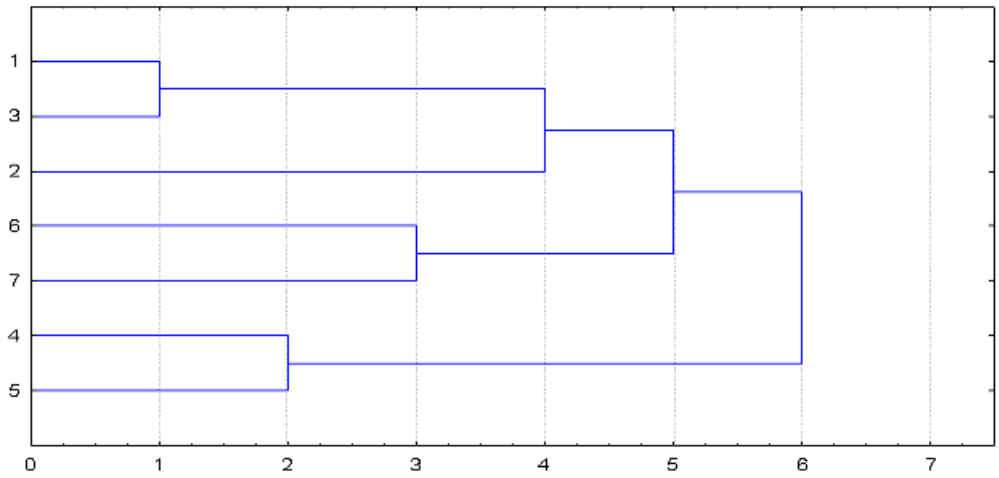


Рис. 3.6.6. Дендрограмма кластеризации по невзвешенному центроидному методу

Кроме рассмотренных выше базовых методов разработаны их модернизированные разновидности, что связано в первую очередь с желанием учесть размеры объединяемых кластеров. С математической точки зрения это приводит к переходу от невзвешенных алгоритмов к взвешенным. Таковыми методами являются метод взвешенного попарного среднего и метод медиан.

3.6.6. Метод взвешенного попарного среднего, или метод минимальной связи (Weighted pair-group average, Within-groups linkage)

Метод идентичен методу средней связи, но при вычислениях размер соответствующих кластеров (то есть число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому метод взвешенного попарного среднего лучше использовать, когда предполагаются неравные размеры кластеров.

Дендрограмма для рассматриваемого примера приведена на рис. 3.6.7:

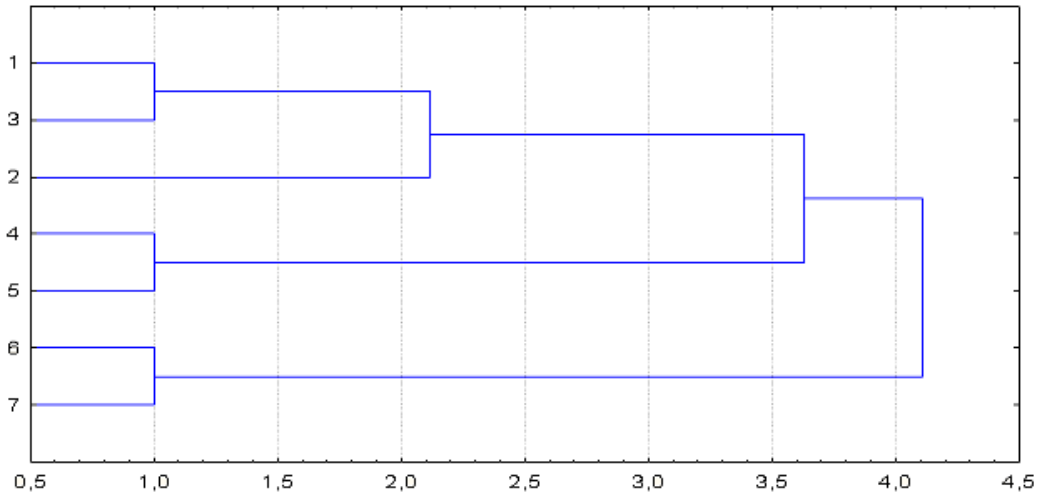


Рис. 3.6.7. Дендрограмма кластеризации по методу взвешенного попарного среднего

3.6.7. Метод медиан, или взвешенный центроидный метод (Weighted pair-group centroid, Median clustering)

Тот же центроидный метод, но при вычислениях используются веса для учета разницы между размерами кластеров (то есть числом объектов в них). Поэтому, если имеются (или предполагаются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее невзвешенного центроидного. Иногда этот метод называют еще методом взвешенных групп.

Дендрограмма для рассматриваемого примера приведена на рис. 3.6.8:

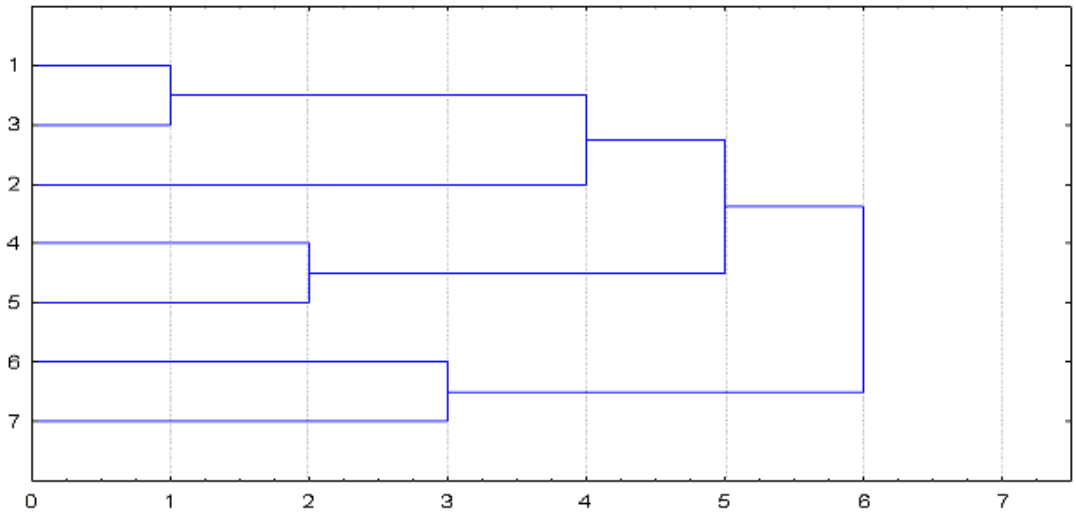


Рис. 3.6.8. Дендрограмма кластеризации по методу медиан

Таким образом, видно, что разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных. Это обычное явление в большинстве прикладных исследований.

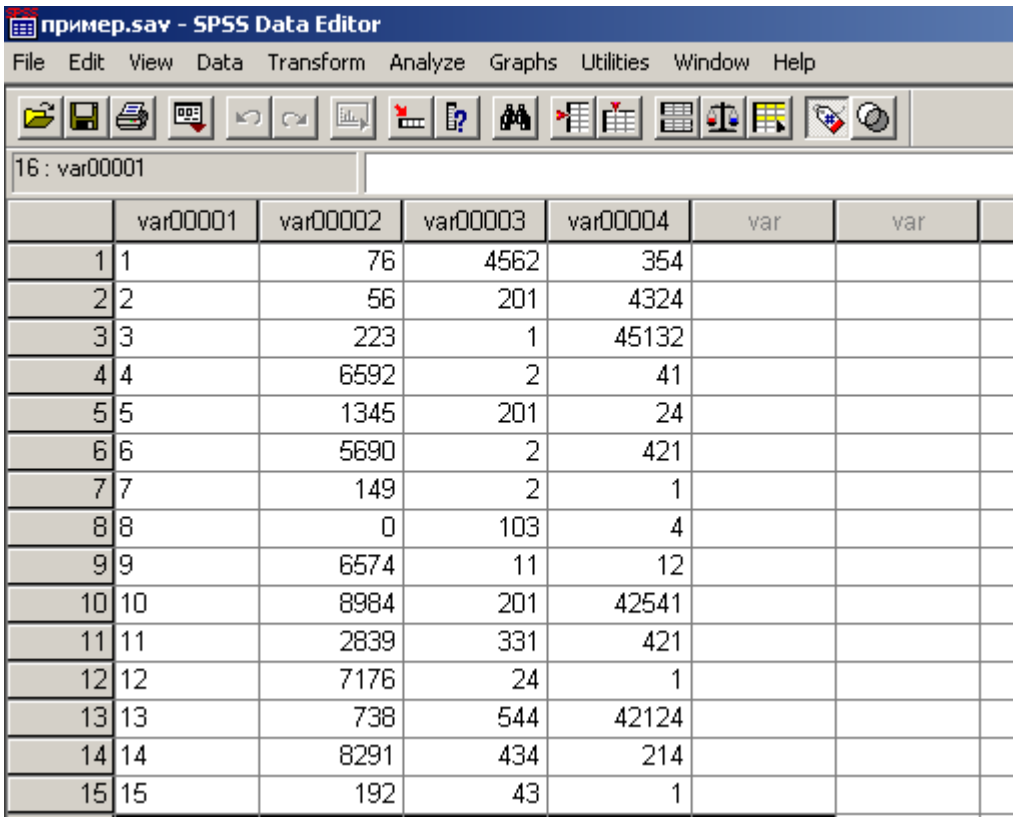
Методы можно условно разделить на три группы по тому, как они преобразуют соотношения между точками в многомерном пространстве. *Сжимающие пространство* методы изменяют эти соотношения, «уменьшая» пространство между любыми группами в данных. Если очередная точка подвергается обработке таким методом то, скорее всего, она будет присоединена к уже существующей группе, а не послужит началом нового кластера.

Расширяющие пространство методы действуют противоположным образом. Здесь кластеры как бы «расступаются» так, что в пространстве образуются мелкие, более «отчетливые» кластеры. Этот способ группировки также зачастую приводит к созданию кластеров гиперсферической формы и приблизительно равных размеров (например, методы Варда и полных связей). И, наконец, есть сохраняющие пространство методы (например, метод средней связи), которые оставляют без изменения свойства исходного пространства.

3.7. Реализация кластерного анализа в пакете SPSS

Одним из наиболее популярных пакетов обработки статистической информации, безусловно, является SPSS. Подробно о достоинствах этой универсальной системы анализа данных можно прочитать, например, в [7, 19].

После ввода данных (рис. 3.7.1) для вызова процедуры иерархического кластерного анализа в пакете SPSS выберите в главном меню последовательность пунктов **Analyze/Classify/Hierarchical Cluster**, после чего на экране появится основное окно процедуры (рис. 3.7.2).



	var00001	var00002	var00003	var00004	var	var
1	1	76	4562	354		
2	2	56	201	4324		
3	3	223	1	45132		
4	4	6592	2	41		
5	5	1345	201	24		
6	6	5690	2	421		
7	7	149	2	1		
8	8	0	103	4		
9	9	6574	11	12		
10	10	8984	201	42541		
11	11	2839	331	421		
12	12	7176	24	1		
13	13	738	544	42124		
14	14	8291	434	214		
15	15	192	43	1		

Рис. 3.7.1. Окно ввода данных в пакете SPSS

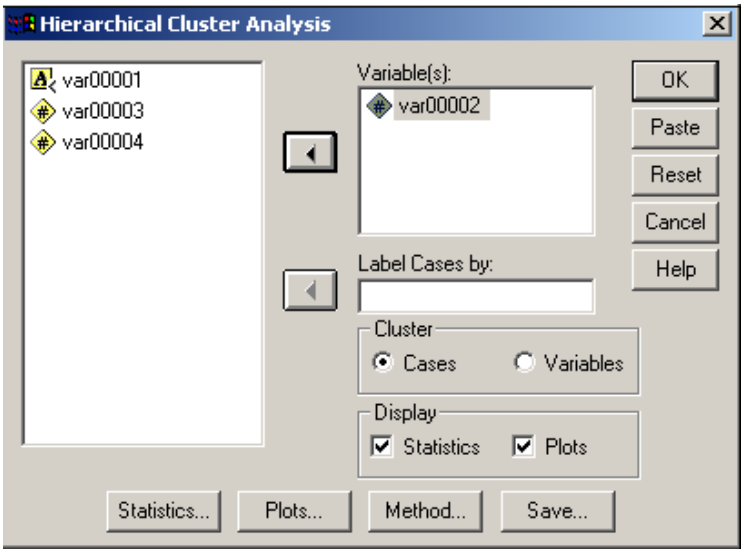


Рис. 3.7.2. Основное окно процедуры кластерного анализа в пакете SPSS

Иерархический кластерный анализ может использоваться как для выявления типологии объектов, так и для классификации переменных. Поэтому первым шагом будет установка значений переключателя **Cluster**. Если кластеризации подвергаются наблюдения (векторы X_i, характеризующие различные объекты по набору признаков), следует оставить его в положении **Cases**. Если же кластеризации подвергаются признаки, характеризующие весь набор объектов, то есть переменные, надо установить этот переключатель в положение **Variables**.

С формальной точки зрения, в первом случае в качестве объектов выступают столбцы матрицы χ, определенной по (3.1.2), а во втором – строки этой матрицы.

В поле **Label Cases by** следует перенести имя строковой переменной (если кластеризации подвергаются наблюдения), с помощью значений которой на графике можно идентифицировать отдельные наблюдения (таким именем, например, может быть фамилия, название или специальный уникальный код). Если классифицируются переменные, данное поле становится недоступным.

При кластеризации наблюдений меры сходства вычисляются на основании тех переменных, которые помещены в список **Variables**. Эти переменные называются критериями кластеризации. Естественно, что именно от их выбора непосредственным образом зависит то, какие объекты окажутся похожими, а какие — нет. Основная проблема состоит в том, чтобы найти ту совокупность переменных, которая наилучшим образом отражает сходство объектов.

Не рекомендуется отбирать для анализа как можно большее количество переменных в надежде на то, что структура «проявится», как только будет собрано достаточное количество данных. Чем больше переменных учитывается при вычислении меры близости, тем сложнее вскрыть реальную структуру данных, поскольку не имеющие отношения к делу признаки вносят статистический «шум» и затрудняют интерпретацию полученных кластеров. Лучше начинать с 2–3 существенных признаков, постепенно добавляя новые переменные при необходимости. Естественным методом для отбора минимального числа независимых переменных может послужить факторный анализ признаков (см. гл. 4).

В режиме экрана **Display** расставляются «флажки»: в окне **Statistics**, если требуется получение отчета о кластеризации, и в окне **Plot**, если требуется получение графиков (дендрограмм).

После того как определен состав наблюдений и критериев кластеризации, необходимо выбрать меру сходства и метод кластеризации, которые должны быть использованы. Для этого, нажав кнопку **Method**, вызываем дополнительное окно (рис. 3.7.3).

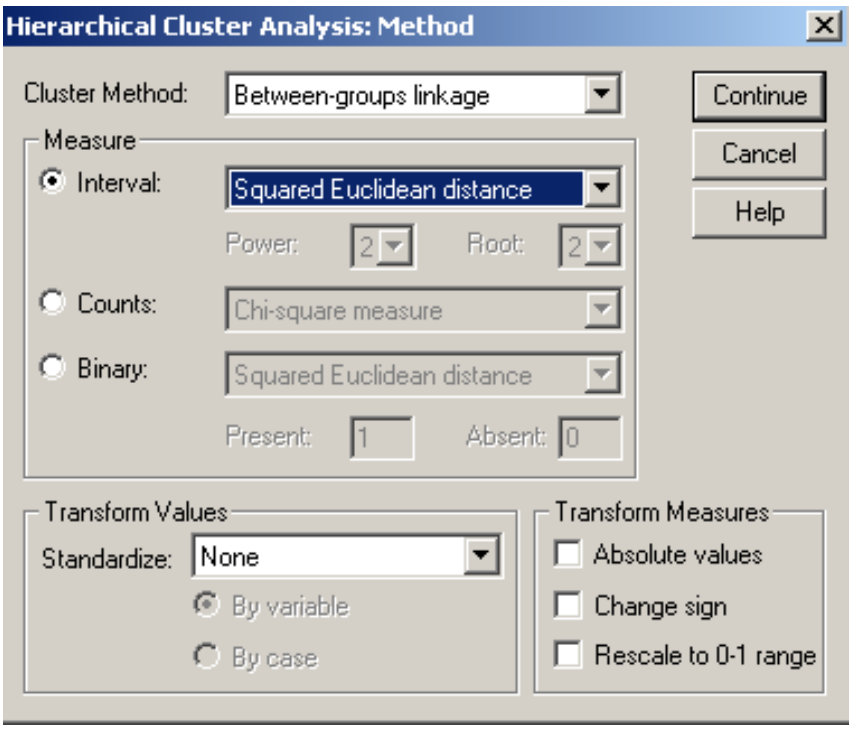


Рис. 3.7.3. Окно выбора меры сходства и метода кластеризации в пакете SPSS

Отметим, что в пакете SPSS не предусмотрена возможность анализа разнотипных переменных. То есть все основания кластеризации должны относиться к одному и тому же типу: интервальные шкалы (**Interval**), частоты (**Counts**) или дихотомические переменные (**Binary**). Для корректного продолжения процедуры требуется указать в окне шкалу (см. Приложение 3), с помощью которой были произведены измерения.

Если переменные вычислены в разных шкалах, то перед вычислением мер сходства следует выполнить их преобразование. Цель таких преобразований состоит в том, чтобы привести переменные к единому масштабу, что позволяет производить осмысленное сравнение объектов по разным признакам.

Для преобразования переменных служат несколько элементов управления. Список **Standardize** в группе **Transform Values** позволяет включить стандартизацию определенного типа (например, *Z*-шкалы, диапазон от 0 до 1, и т.п.) или отключить ее совсем (значение **None**). При включенной процедуре стандартизации можно выбрать, каким образом она будет производиться: по наблюдениям (**By case**) или по переменным (**By variable**).

Группа **Transform Measures** дает возможность дополнительного преобразования знака меры сходства: можно взять абсолютные значения (**Absolute values**), поменять знак (**Change sign**) или перевести значения в диапазон от 0 до 1 (**Rescale to 0–1 range**). Например, использование абсолютного значения при введении коэффициента корреляции может дать хороший результат: ведь сильная отрицательная связь указывает на такую степень несходства, при которой объекты похожи, как говорят, «с точностью до наоборот».

Метод кластеризации (называемый еще «стратегией кластеризации») определяет способ объединения объектов в кластеры. Раскрывающийся список **Cluster Method** позволяет выбрать один из следующих методов: **Between-groups linkage** (метод средней связи), **Within-groups linkage** (метод минимальной связи), **Nearest neighbour** (метод «ближайшего соседа»), **Furthest neighbour** (метод «дальнего соседа»), **Centroid clustering** (центроидный метод), **Median clustering** (медианная кластеризация), **Ward's method** (метод Варда).

Процесс агрегирования данных может быть представлен графически деревом объединения кластеров либо «сосульковой» диаграммой. Подробнее о процессе кластеризации можно узнать по специальной таблице объединения кластеров (**Statistics/Agglomeration Schedule**), вызываемой из основного окна процедуры (рис. 3.7.2). Кроме того, там же можно затребовать матрицу близостей и диапазон решений – минимальное и максимальное количество кластеров.

Кнопка **Plots** главного окна процедуры (рис. 3.7.2) позволяет заказать различные формы представления результатов кластерного анализа. Для вывода дендрограммы требуется установить «флажок» **Dendrogram**. По умолчанию выводится также представление результатов кластеризации при помощи «сосулек» (**Icicles**).

Диалог, вызываемый кнопкой **Save** главного окна, дает возможность сохранить принадлежность объектов к кластерам (рис. 3.7.4).

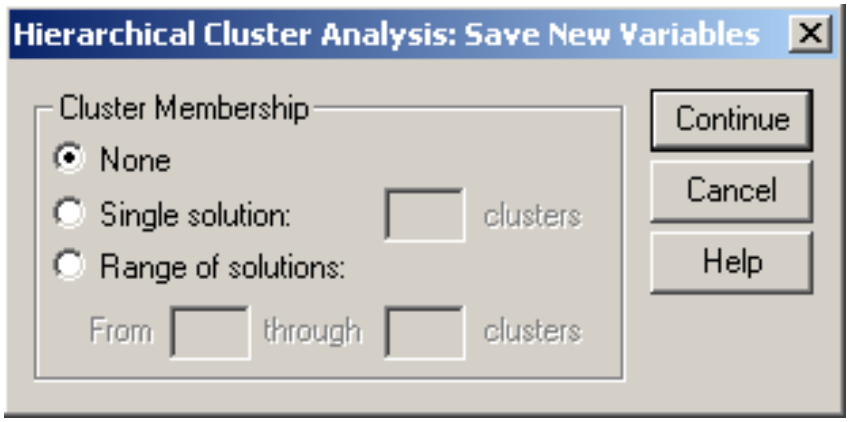


Рис. 3.7.4. Окно сохранения принадлежности объектов кластерам в пакете SPSS

При этом можно указать, сохраняется ли одно решение (**Single solution**) или диапазон решений (**Range of solutions**). Когда будет выбран один из этих пунктов, станут доступными поля ввода, где можно указать число кластеров в решении, которое необходимо сохранять. Создаваемые пакетом новые переменные представляют собой признаки, измеренные в номинальной шкале. Их значения – номера кластеров, к которым принадлежат наблюдения.

Эти номинальные признаки могут в дальнейшем участвовать в других процедурах: дискриминантном анализе (в качестве независимой переменной), в дисперсионном анализе (в качестве уровней фактора), при построении таблиц сопряженности.

Результаты кластеризации выводятся на экран в графической и аналитической формах. Основным видом графического представления является дендрограмма (рис. 3.7.5).

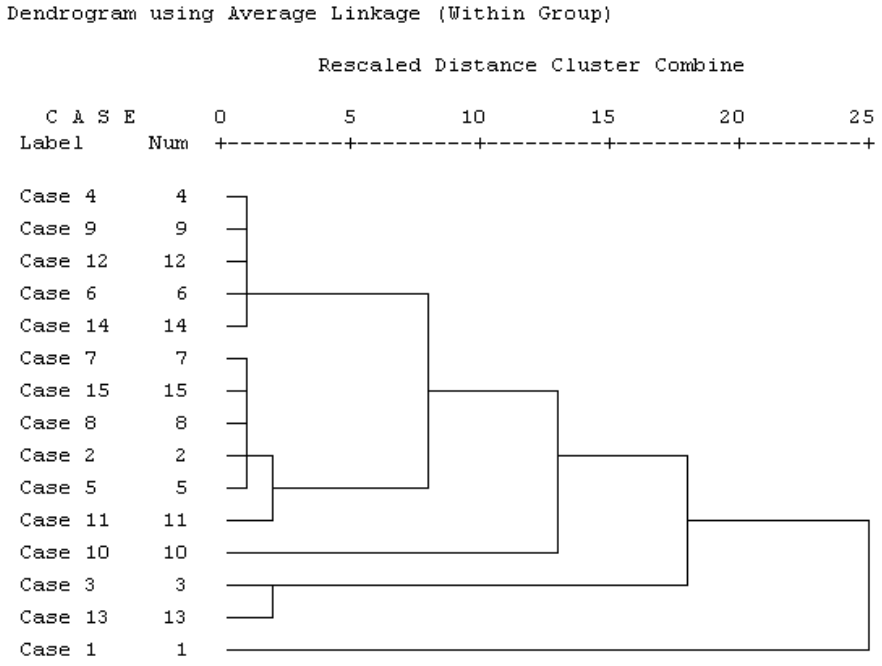


Рис. 3.7.5. Дендрограмма по методу минимальной связи в пакете SPSS

Вверху графика изображена шкала расстояния, значения на которой всегда находятся в диапазоне от 0 (минимальное расстояние между объектами, полное сходство) до 25 (максимальное расстояние). В основной части диаграммы находится «дерево», наглядно показывающее степень сходства объектов. Так, объекты с номерами 4, 9, 12, 6, 14 принадлежат одному кластеру – на графике они входят в одну «гроздь».

Продолжая вертикальную линию, их объединяющую, до шкалы измерения расстояния, можно узнать степень относительной близости объектов в первом кластере; диаметр этого кластера равен 1. Таким образом, для каждого узла (там, где формируется новый кластер) можно определить величину расстояния, при котором соответствующие элементы связываются в новую группу.

Аналитическая форма представления процесса слияния объектов фиксируется в таблице объединения (табл. 3.7.1).

Таблица 3.7.1

Таблица объединения (Agglomeration Schedule) в пакете SPSS

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	9	1246,000	0	0	4
2	7	15	3530,000	0	0	3
3	7	8	25471,333	2	0	5
4	4	12	235693,333	1	0	6
5	5	7	786815,167	0	3	8
6	4	6	833163,667	4	0	7
7	4	14	1972132,60	6	0	12
8	5	11	3019992,40	5	0	9
9	2	5	8632690,07	0	8	11
10	3	13	9608138,00	0	0	13
11	1	2	12954356,9	0	9	12
12	1	4	28598537,9	11	7	14
13	3	10	53801531,3	10	0	14
14	1	3	657675245	12	13	0

Первый столбец в таблице содержит номер этапа кластеризации, два следующих (**Clusters Combined**) – номера объединяемых кластеров. Так, на первом этапе сливаются в один кластер объекты под номерами 4 и 9. Полученному кластеру присваивается номер первого из объединяемых объектов (в данном случае 4).

В столбце **Coefficient** приводится значение расстояния между объединяемыми кластерами. Последние два столбца (**Stage Cluster First Appears**) показывают, на каком этапе впервые появился каждый из кластеров. Нулевые значения приведены для исходных объектов, ненулевые — для кластеров, полученных их слиянием.

3.8. Особенности реализации кластерного анализа в пакете «Statistica»

Система «Statistica» – еще один, наряду с SPSS, популярный пакет обработки статистической информации [2]. Принципиально реализация алгоритмов кластерного анализа в «Statistica» (версия 6.0) не отличается от ее реализации в пакете SPSS. Выявление особенностей работы кластерного анализа в пакете «Statistica» проведем на основе ее сравнения с SPSS.

Стандартизацию данных в системе «Statistica» необходимо проводить заранее, через пункт меню **Data/Standartize**. Так же, как и система SPSS, система «Statistica» на выходе дает аналитический (таблица) и графический (дендрограмма) отчет об объединении объектов.

В качестве примера проведем группировку 14 банков Уральского федерального округа. Каждый банк характеризуется набором показателей: капитал, чистые активы, обязательства, кредиты, ценные бумаги, вклады граждан, расчетные счета средства бюджета и прибыль. Эти данные (по состоянию на конец 2001 года) приведены в табл. 3.8.1.

В качестве исходного материала была взята таблица расстояний, с элементами, подсчитанными на основании (3.1.1) после нормирования с использованием *Z*-вклада.

Таблица 3.8.1

Экономические показатели некоторых банков Уральского региона, по данным на декабрь 2001 года (в тыс. руб.)

Банк	Капитал	Чистые активы	Обяза-тельства	Кредиты	Ценные бумаги	Вклады граждан	Расчет-ные счета	Сред-ства бюджета	При-быль
УБРиР	757457	2514660	1893389	2216767	53776	758137	180741	67197	1691
КУБ	748126	3089747	2566531	2632099	648	631061	820126	1335	148357
ЧИБ	510723	3410124	2579138	2074263	335818	727310	1190751	257348	33233
ХМБ	491173	12344684	12987008	1156238	1662076	315102	1207968	6966037	220355
ЗПБ	452479	3174058	2933912	1256912	1146138	156786	483366	77813	-4623
СГБ	57483	1225817	840246	254539	246495	63479	63928	8768	-1288
АБ	323831	645989	342780	295282	253225	52043	25028	457	40
УТБ	319416	1044876	814121	539291	164380	90373	139401	0	1487
СБ	275589	1185269	991140	894221	40783	66386	409866	173632	15344
УПСБ	270292	2261474	2046850	1382656	90831	561753	542437	89309	8893
Капи-тал	260160	1864879	1730309	722370	648350	178860	439960	493724	27309
МКБ	257092	611858	388318	202279	23898	33649	252165	0	3305
ЧИНБ	247299	1291834	1123550	839344	75350	255224	455420	48706	22834
УТБ	233118	1527899	1377973	773928	122200	378235	287174	60132	13480

Примечание. В таблице приняты следующие сокращения: УБРиР – Уральский банк реконструкции и развития, ЗПБ – Золото-Платина банк, СГБ – Свердловский Губернский банк, УПСБ – Уральский промышленно-строительный банк (все – г. Екатеринбург); ЧИБ – Челябинский индустриальный банк, ЧИНБ – Челябинский инвестиционный банк (оба г. Челябинск); КУБ – Кредит-Урал банк (г. Магнитогорск); ХМБ – Ханты-Мансийский банк (г. Ханты-Мансийск); АБ – Асбест-банк (г. Асбест); УТБ – Уральский трастовый банк (г. Ижевск); СБ – Снежинский банк (г. Снежинск); Капитал (г. Нижневартовск); МКБ – МетКомБанк (г. Каменск-Уральский).

Результат группировки по методу одиночной связи в SPSS приведен в табл. 3.7.1, а дендрограмма – на рис. 3.8.1.

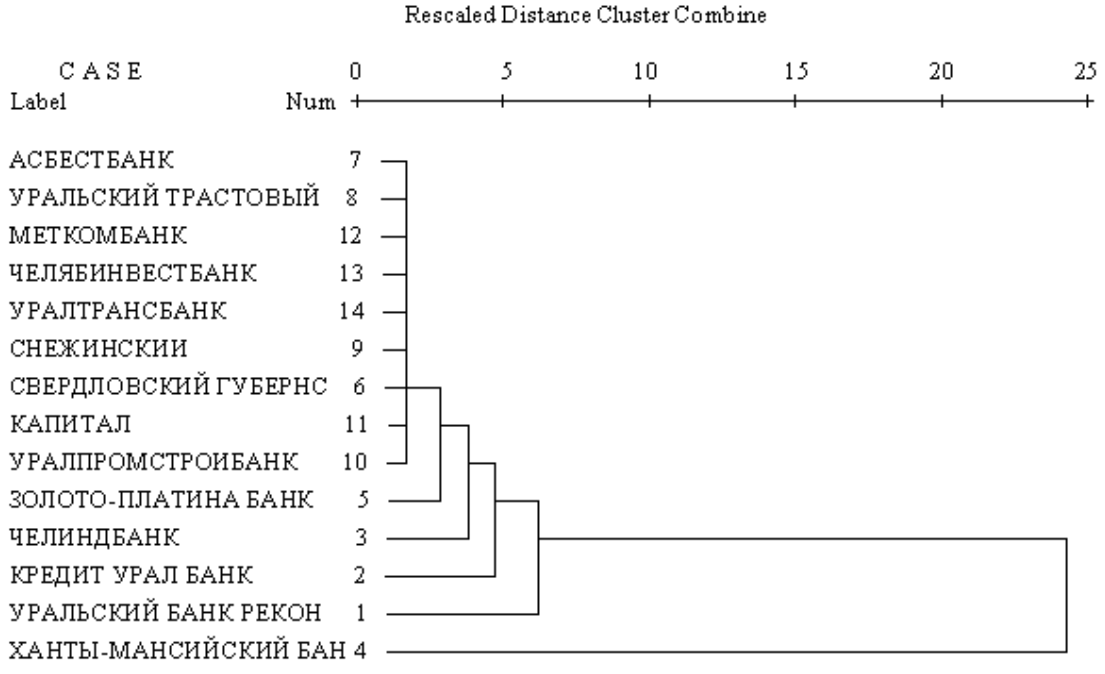


Рис. 3.8.1. Дендрограмма кластеризации банков по методу одиночной связи в пакете SPSS

Для этой же задачи реализуем алгоритм метода одиночной связи в пакете «Statistica 6.0». В качестве отчета получим таблицу последовательной группировки (табл. 3.8.2).

Таблица 3.8.2

Таблица объединения банков по методу одиночной связи в пакете «Statistica»

	Amalgamation Schedule Single Linkage Squared Euclidean distances													
linkage distance	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7	Obj. 8	Obj. 9	Obj. 10	Obj. 11	Obj. 12	Obj. 13	Obj. 14
,2915311	C_7	C_8												
,4738795	C_13	C_14												
,5547080	C_7	C_8	C_12											
,5804698	C_9	C_13	C_14											
,9168113	C_7	C_8	C_12	C_9	C_13	C_14								
1,551917	C_6	C_7	C_8	C_12	C_9	C_13	C_14							
1,625432	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11						
1,741246	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10					
3,114792	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10				
6,243261	C_3	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10			
6,634903	C_2	C_3	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10		
8,497235	C_1	C_2	C_3	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10	
50,58330	C_1	C_2	C_3	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10	C_4

В первом столбце таблицы указывается расстояние между кластерами, в последующих перечисляются сгруппированные на соответствующем шаге объекты. Исходные объекты обозначаются через C_i, где i – номер объекта.

Дендрограмма, построенная в «Statistica», приведена на рис. 3.8.2.

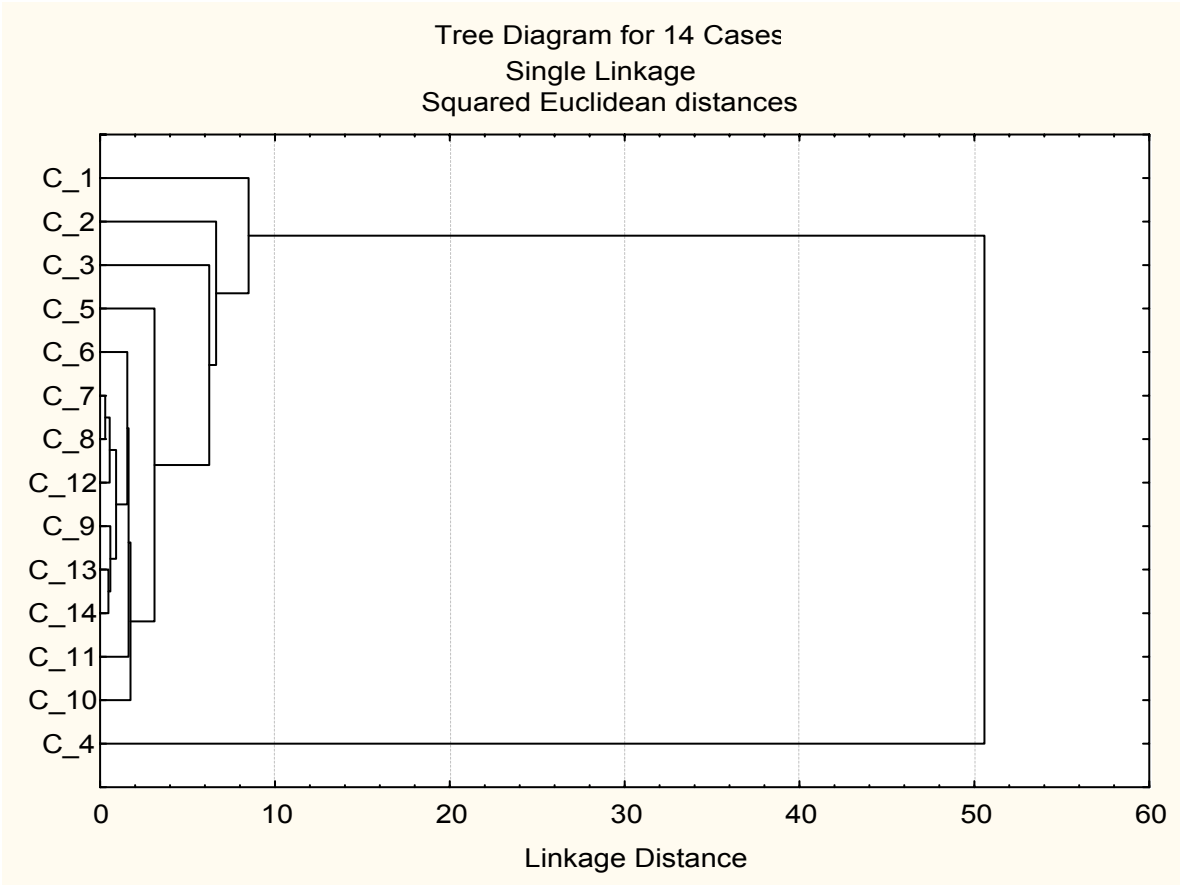


Рис. 3.8.2. Дендрограмма кластеризации банков по методу одиночной связи в пакете «Statistica»

Как видно, при полном совпадении алгоритма группировки и одинаково вычисленных расстояниях между кластерами графическое представление результатов в пакетах «Statistica» и SPSS отличается.

Для более полного представления приведем сравнительные результаты группировки тех же банков другими методами.

Кластеризация центроидным методом дает следующие результаты (табл. 3.8.3, 3.8.4).

Таблица 3.8.3

Таблица объединения банков по центроидному методу в пакете SPSS

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	7	8	,292	0	0	3
2	13	14	,474	0	0	4
3	7	12	,565	1	0	5
4	9	13	,993	0	2	5
5	7	9	1,499	3	4	6
6	7	11	1,763	5	0	7
7	6	7	1,923	0	6	8
8	6	10	4,615	7	0	9
9	5	6	6,610	0	8	12
10	2	3	6,635	0	0	11
11	1	2	7,357	0	10	12
12	1	5	14,788	11	9	13
13	1	4	62,179	12	0	0

Таблица 3.8.4

Таблица объединения банков по центроидному методу в пакете «Statistica»

	Amalgamation Schedule Unweighted pair-group centroid Squared Euclidean distances													
linkage distance	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7	Obj. 8	Obj. 9	Obj. 10	Obj. 11	Obj. 12	Obj. 13	Obj. 14
,2915311	C_7	C_8												
,4738795	C_13	C_14												
,5651486	C_7	C_8	C_12											
,9926276	C_9	C_13	C_14											
1,498850	C_7	C_8	C_12	C_9	C_13	C_14								
1,763476	C_7	C_8	C_12	C_9	C_13	C_14	C_11							
1,923440	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11						
4,614546	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10					
6,610292	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10				
6,634903	C_2	C_3												
7,357102	C_1	C_2	C_3											
14,78765	C_1	C_2	C_3	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10	
62,17867	C_1	C_2	C_3	C_5	C_6	C_7	C_8	C_12	C_9	C_13	C_14	C_11	C_10	C_4

При совпадении аналитических отчетов, приведенных в табл. 3.8.3 и 3.8.4, вновь наблюдается расхождение при построении дендрограмм (рис. 3.8.3, 3.8.4).

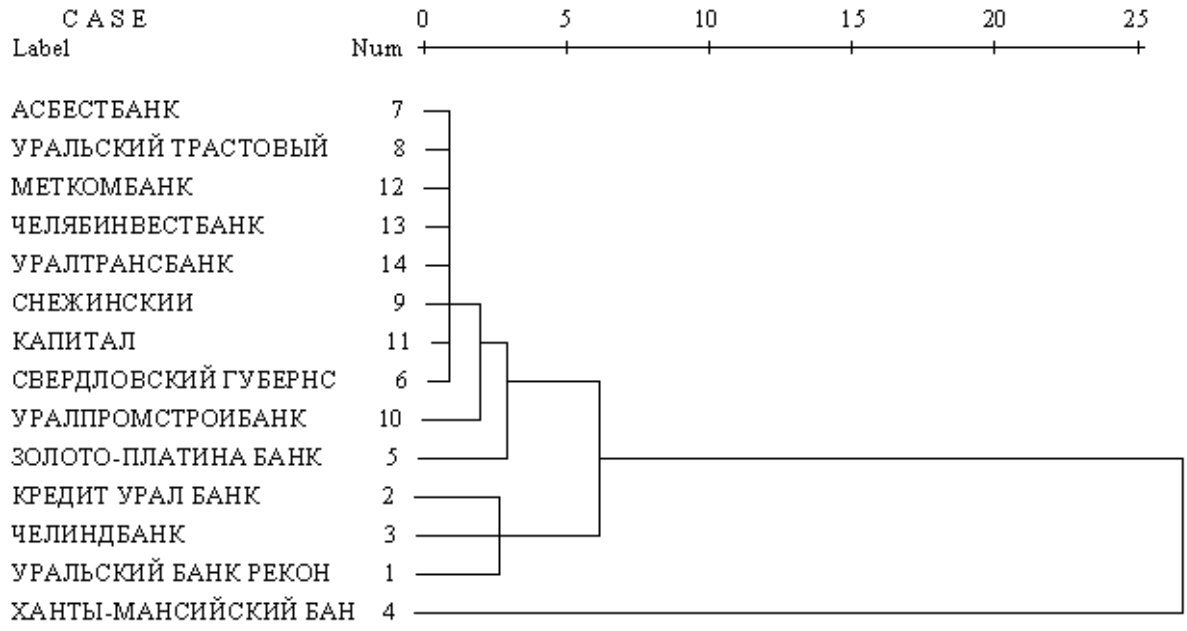


Рис. 3.8.3. Дендрограмма по центроидному методу в пакете SPSS

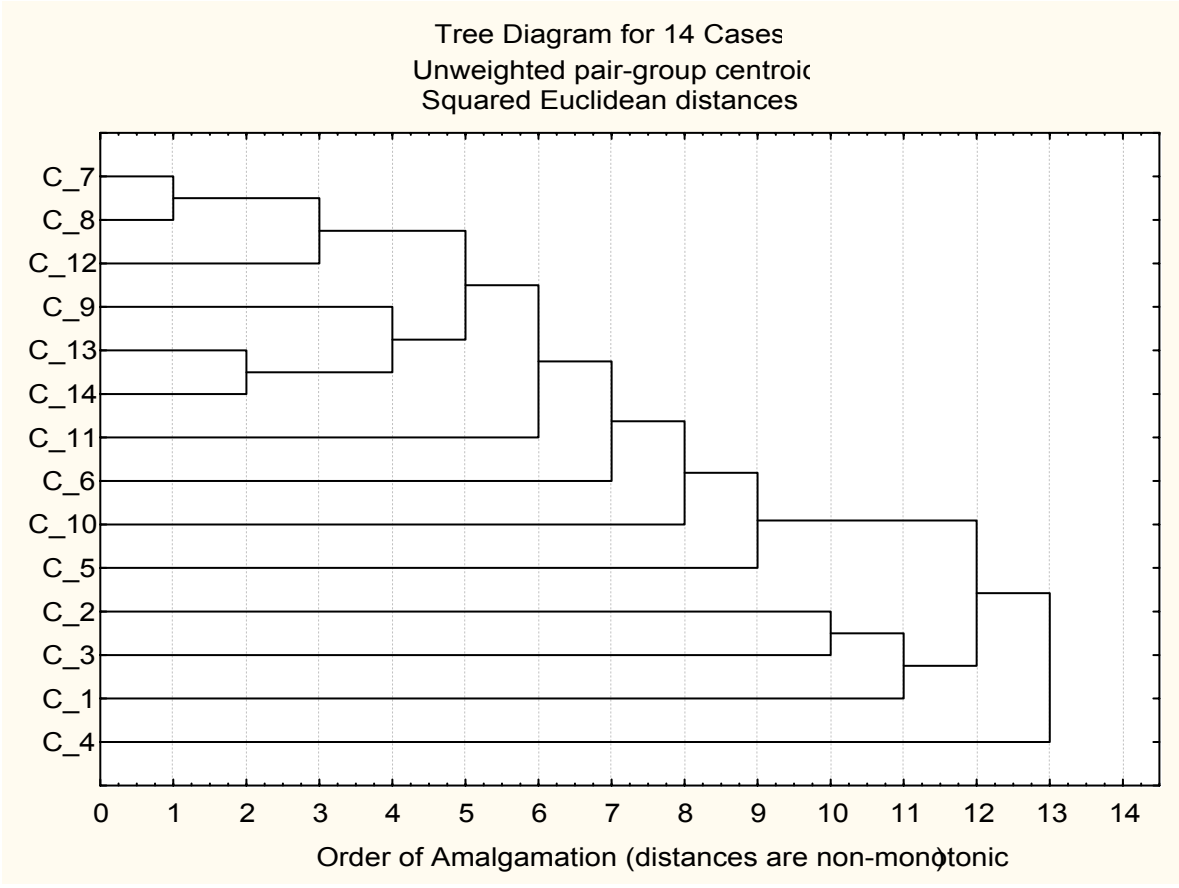


Рис. 3.8.4. Дендрограмма по центроидному методу в пакете «Statistica»

Рассмотрим первые два шага в центроидном методе. В обеих программах они идентичны. Объекты с номерами 7 и 8 объединяются в кластер с расстоянием 0,292, на втором шаге – объекты 13 и 14 с расстоянием 0,474. Следовательно, на дендрограмме кластер {13, 14} должен быть более вытянутым, чем кластер {7, 8}. В SPSS же они изображаются на одном уровне.

Кластеризация по методу Варда, выполненная в этих двух пакетах, демонстрирует различия не только в дендрограммах (рис. 3.8.5, 3.8.6), но и в аналитических отчетах (табл. 3.8.5, 3.8.6).

Таблица 3.8.5

Таблица объединения по методу Варда в пакете SPSS

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	7	8	,146	0	0	3
2	13	14	,383	0	0	4
3	7	12	,759	1	0	6
4	9	13	1,421	0	2	5
5	9	11	2,605	4	0	7
6	6	7	3,820	0	3	11
7	9	10	5,937	5	0	9
8	2	3	9,255	0	0	10
9	5	9	14,104	0	7	11
10	1	2	19,009	0	8	12
11	5	6	25,137	9	6	12
12	1	5	59,263	10	11	13
13	1	4	117,000	12	0	0

Таблица 3.8.6

Таблица объединения по методу Варда в пакете «Statistica»

	Amalgamation Schedule Ward`s method Squared Euclidean distances													
linkage distance	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5	Obj. 6	Obj. 7	Obj. 8	Obj. 9	Obj. 10	Obj. 11	Obj. 12	Obj. 13	Obj. 14
,2915311	C_7	C_8												
,4738795	C_13	C_14												
,7535315	C_7	C_8	C_12											
1,323503	C_9	C_13	C_14											
2,367463	C_9	C_13	C_14	C_11										
2,431087	C_6	C_7	C_8	C_12										
4,233699	C_9	C_13	C_14	C_11	C_10									
6,634903	C_2	C_3												
9,698985	C_5	C_9	C_13	C_14	C_11	C_10								
9,809470	C_1	C_2	C_3											
12,25660	C_5	C_9	C_13	C_14	C_11	C_10	C_6	C_7	C_8	C_12				
68,25067	C_1	C_2	C_3	C_5	C_9	C_13	C_14	C_11	C_10	C_6	C_7	C_8	C_12	
115,4747	C_1	C_2	C_3	C_5	C_9	C_13	C_14	C_11	C_10	C_6	C_7	C_8	C_12	C_4

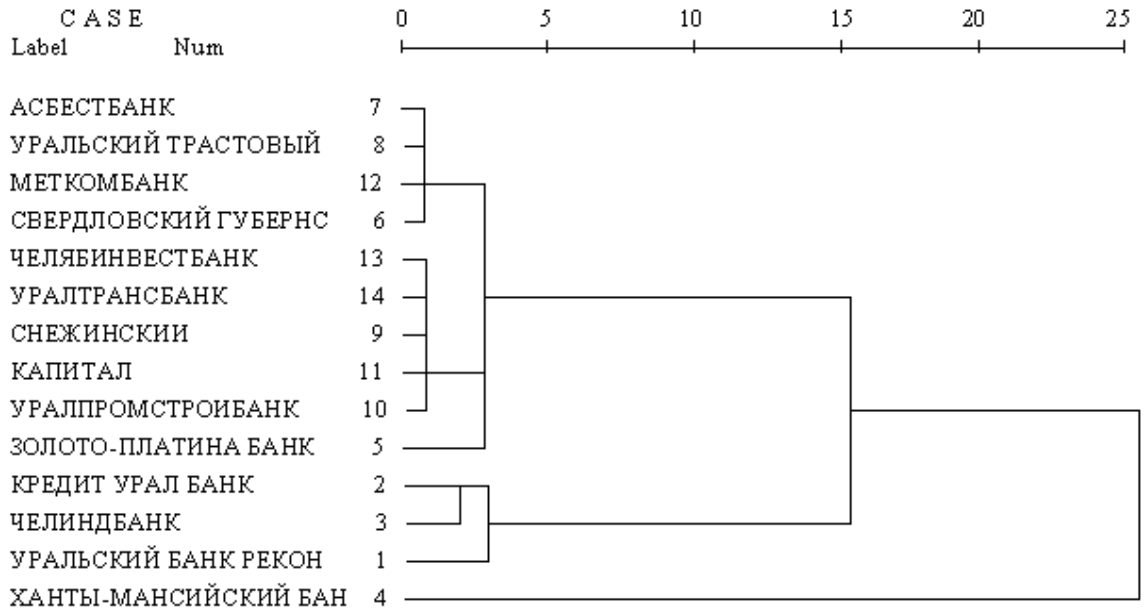


Рис. 3.8.5. Дендрограмма по методу Варда в пакете SPSS

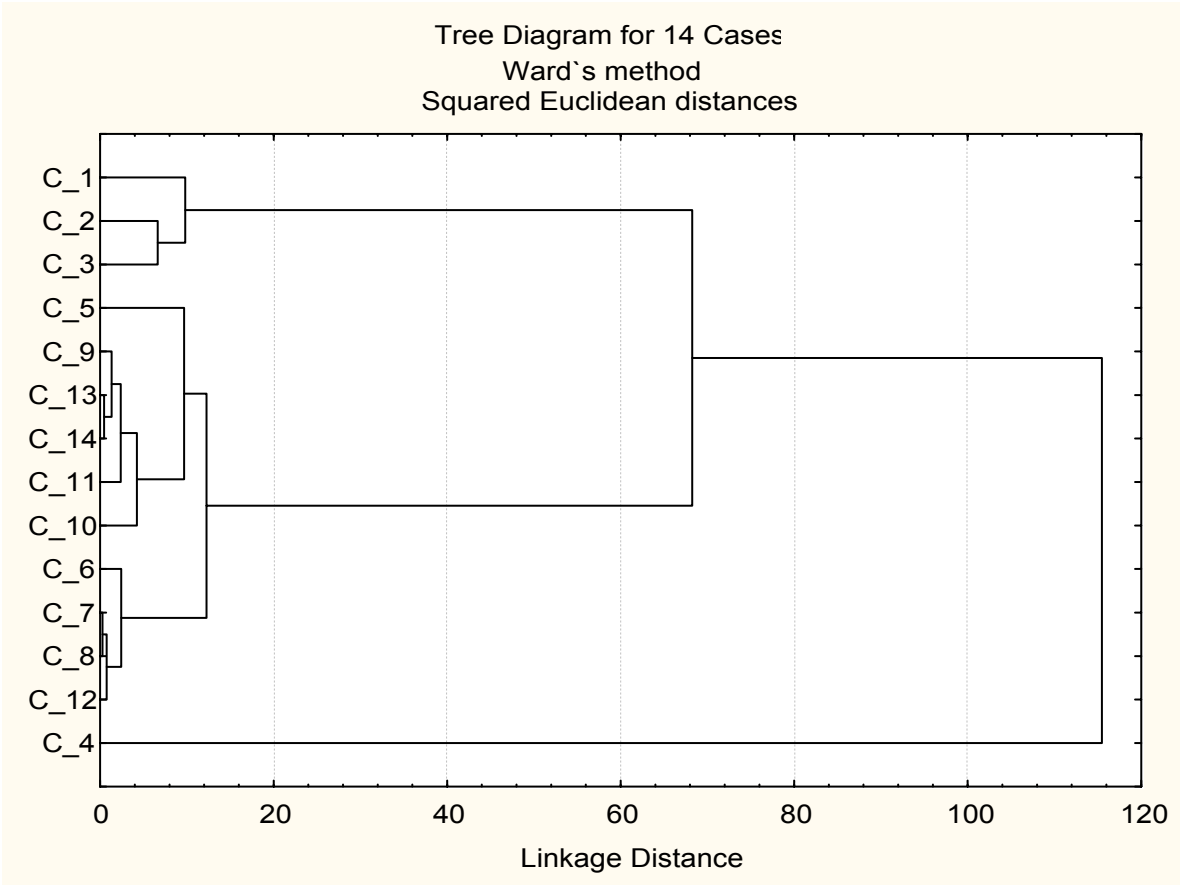


Рис. 3.8.6. Дендрограмма по методу Варда в пакете «Statistica»

Таким образом, кластеризация по методу Варда в двух рассматриваемых пакетах дает наиболее существенные расхождения.

3.9. Пример. Расчет загрузки персонала с учетом специализации

Рассмотрим алгоритм кластерного анализа на примере задачи из области менеджмента. Предположим, необходимо выяснить степень загрузки персонала сети магазинов розничной торговли с учетом профиля специализации (вида товара), что связано, в частности, с планированием отпусков. Для решения задачи нам понадобится провести группировку месяцев года в зависимости от объемов реализации различных товаров в этой сети магазинов. Исходные данные, полученные из отчета о результатах деятельности одной екатеринбургской компании путем умножения на некоторый постоянный коэффициент, приведены в табл. 3.9.1.

Таблица 3.9.1

Объемы реализации товаров по месяцам (в штуках)

Месяц	Характеристики признаков							
	Товар 1	Товар 2	Товар 3	Товар 4	Товар 5	Товар 6	Товар 7	Товар 8
Январь	108	85	131	71	18	85	8	3
Февраль	83	58	58	65	10	55	30	5
Март	112	62	79	67	11	71	95	22
Апрель	136	69	96	85	13	83	63	25
Май	129	49	110	64	26	89	9	4
Июнь	180	76	211	125	40	18	5	3
Июль	139	65	160	108	29	173	10	4
Август	158	64	120	83	14	101	37	21
Сентябрь	125	65	87	71	16	89	54	10
Октябрь	77	61	78	70	8	72	22	15
Ноябрь	138	77	91	60	11	82	3	3
Декабрь	315	119	92	125	25	143	6	3

В качестве метрики использован квадрат евклидова расстояния. Для стандартизации данных применен способ *Z*-шкалы. Расстояние между парой кластеров определялось двумя методами: методом минимальной связи и методом медиан.

В результате обработки данных в пакете SPSS получим следующую таблицу расстояний (табл. 3.9.2).

Таблица 3.9.2

Таблица расстояний

	1	2	3	4	5	6	7	8	9	10	11	12
1	,000	7,440	17,814	12,525	5,219	18,643	10,780	8,116	5,669	7,031	2,128	23,703
2	7,440	,000	9,676	10,009	6,388	34,670	23,608	9,446	3,321	1,978	3,992	41,088
3	17,814	9,676	,000	2,533	17,022	43,248	30,221	6,688	4,580	7,457	16,132	46,962
4	12,525	10,009	2,533	,000	13,493	32,247	20,679	1,784	3,763	5,417	12,271	33,464
5	5,219	6,388	17,022	13,493	,000	20,890	10,406	8,145	5,173	7,328	5,214	33,226
6	18,643	34,670	43,248	32,247	20,890	,000	19,691	25,831	28,171	34,221	27,992	31,135
7	10,780	23,608	30,221	20,679	10,406	19,691	,000	12,712	14,732	20,649	16,232	21,183
8	8,116	9,446	6,688	1,784	8,145	25,831	12,712	,000	3,297	4,748	8,188	27,558
9	5,669	3,321	4,580	3,763	5,173	28,171	14,732	3,297	,000	3,138	4,790	30,044
10	7,031	1,978	7,457	5,417	7,328	34,221	20,649	4,748	3,138	,000	4,606	39,594
11	2,128	3,992	16,132	12,271	5,214	27,992	16,232	8,188	4,790	4,606	,000	25,899
12	23,703	41,088	46,962	33,464	33,226	31,135	21,183	27,558	30,044	39,594	25,899	,000

На основе анализа предложенными методами в пакете SPSS получим графическое представление в виде дендрограмм (рис. 3.9.1, 3.9.2):

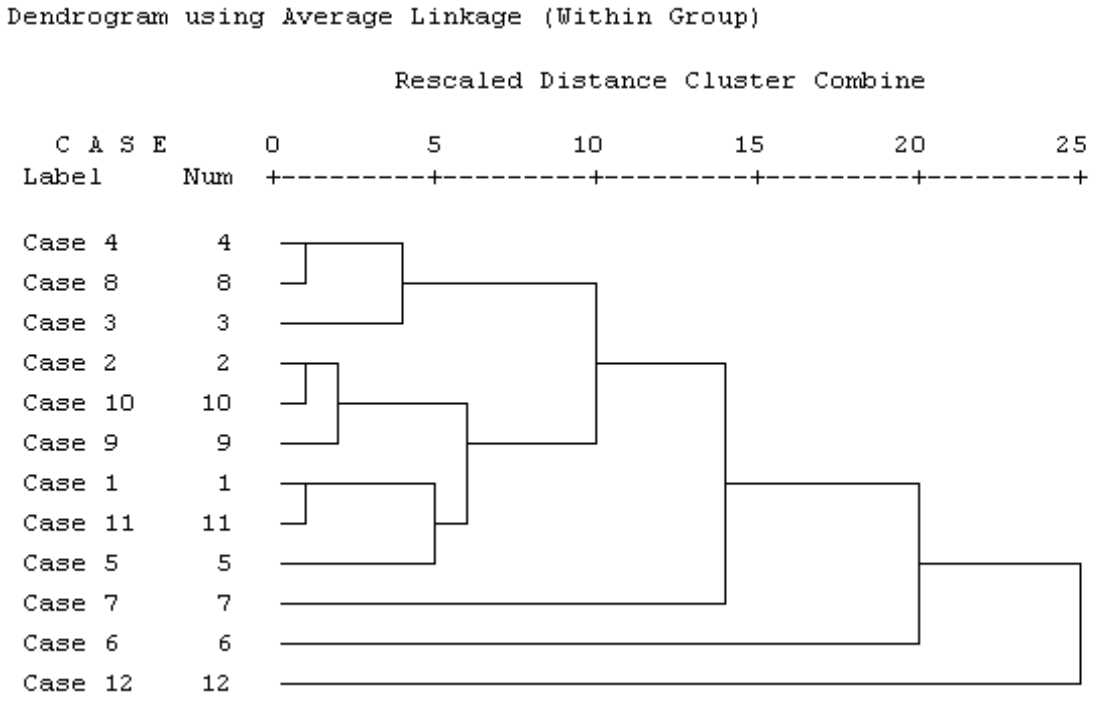


Рис. 3.9.1. Дендрограмма по методу минимальной связи

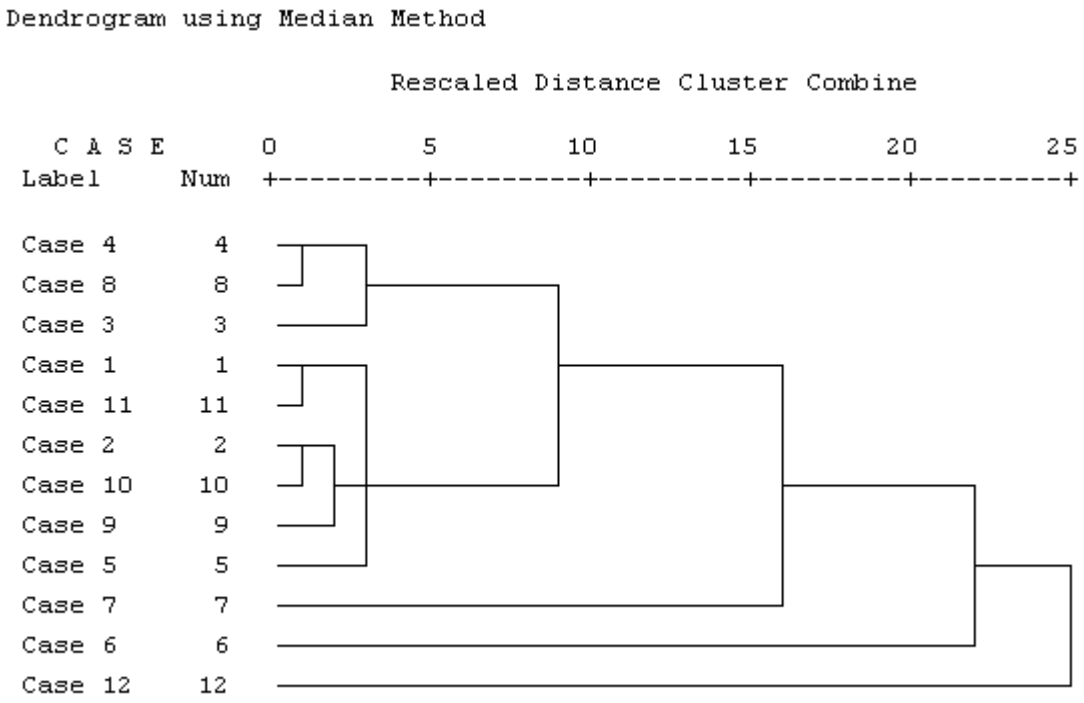


Рис. 3.9.2. Дендрограмма по методу медиан

Проведем деление на кластеры по пятому уровню. В первом случае получаем шесть кластеров: {4, 8, 3}, {2, 10, 9}, {1, 11, 5}, {7}, {6}, {12}; во втором – пять: {4, 8, 3}, {1, 11, 2, 10, 9, 5}, {7}, {6}, {12}. Видно, что в обоих случаях месяцы июнь, июль и декабрь выделяются в отдельные кластеры.

Для большей наглядности воспользуемся вспомогательной диаграммой (рис. 3.9.3), полученной с помощью электронных таблиц Microsoft Excel 2002 [15].

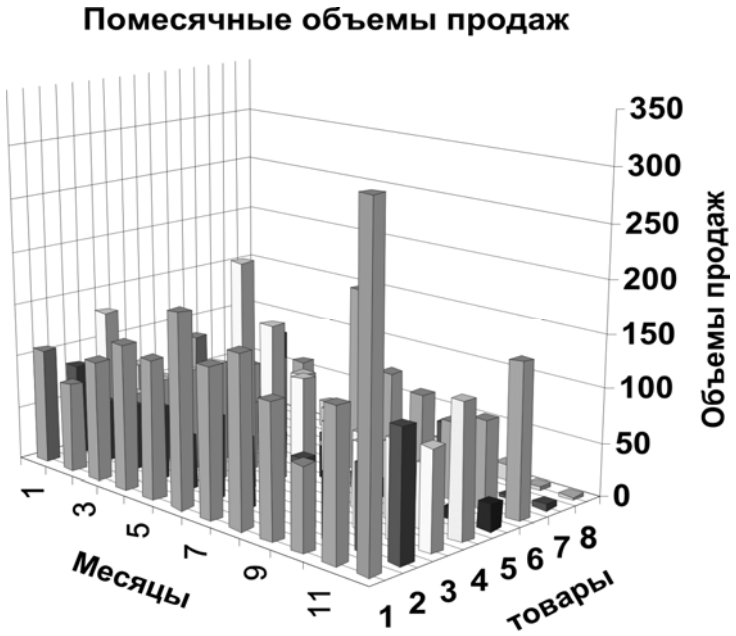


Рис. 3.9.3. Диаграмма объемов продаж, выполненная средствами «Excel»

Сравнительный анализ рис. 3.9.2 и 3.9.3 позволяет сделать следующие выводы: резко выделяются декабрь и его первый товар, июнь и его третий товар, июль и его шестой товар. Особенность декабря в том, что количество продаж первого товара превышает количество продаж любого другого товара в любой другой месяц.

Количество продаж третьего товара в шестом месяце и количество продаж шестого товара в седьмом месяце находятся соответственно на втором и третьем местах. Видно, что декабрь значительно опережает другие месяцы по количеству товаров, а июнь и июль находятся соответственно на третьем и втором местах.

Следовательно, можно сделать вывод, что произошло объединение в кластеры на основе, как минимум, двух признаков: количественного и качественного.

Глава 4. ФАКТОРНЫЙ АНАЛИЗ

Построение математической модели любого явления или объекта связано с восстановлением его состояния по наблюдению за его реакциями на различные входные воздействия [13]. Как правило, непосредственное измерение основных параметров, определяющих свойства системы, невозможно. Иногда неизвестно даже число и содержательный смысл этих параметров, которые в дальнейшем будем называть основными факторами, или просто факторами [26]. При этом для измерения могут быть доступны иные величины (реакции), тем или иным способом зависящие от факторов. Эти величины принято называть косвенными признаками. Если влияние фактора проявляется в нескольких признаках, то эти признаки могут обнаруживать достаточно тесную связь (так называемый эффект мультиколлинеарности [6]).

В статистических исследованиях это проявляется в наличии корреляционной зависимости между признаками. Обычно число существенных (независимых) факторов гораздо меньше числа измеряемых признаков. Суть факторного анализа состоит в переходе от описания системы большим набором косвенных признаков к описанию той же системы меньшим числом максимально информативных факторов.

Важная отличительная черта факторного анализа – возможность одновременного исследования сколь угодно большого числа взаимозависимых переменных, что особенно значимо при описании социально-экономических явлений [12].

Идея метода состоит в сжатии матрицы признаков в матрицу с меньшим числом переменных, сохраняющую почти ту же самую информацию, что и исходная матрица. Обычно применяется линейная модель факторного анализа, в которой полагается представление исходных переменных (признаков) в виде линейной комбинации факторов. С геометрической точки зрения задача факторного анализа аналогична поиску оптимального базиса (базиса наименьшей размерности с заданной точностью представления объектов).

Идея факторного анализа, по-видимому, принадлежит Карлу Пирсону [9]. В 1901 году он предложил метод главных компонент – наиболее используемый метод решения задач факторного анализа. Этот метод был вновь открыт Хоттелингом в 1933 году. Со второй половины прошлого века факторный анализ получил признание как универсальный метод компактного представления больших массивов статистических и экспериментальных данных. Ему посвящено большое количество публикаций, обратившись к которым можно получить сведения различной степени детализации и строгости [14, 17, 21, 26].

В настоящее время факторный анализ широко используется для обработки данных в различных областях естествознания, в первую очередь, в математическом моделировании, экономике, социологии, психологии. Достаточно широкий круг практических задач, решаемых этим методом, описан в работах [12, 25].

Особенно следует подчеркнуть, что факторный анализ представляется чрезвычайно действенным методом борьбы с таким нежелательным явлением множественного регрессионного анализа и эконометрики, как мультиколлинеарность. Факторный анализ – неотъемлемая часть компьютерных систем обработки статистических данных, алгоритм главных компонент реализован практически во всех стандартных пакетах [2, 7, 9, 24].

4.1. Постановка задачи. Линейная модель факторного анализа

Пусть проводится p наблюдений над n признаками $X_1, X_2, ..., X_n$. Под наблюдениями понимаем набор из p однотипных объектов для каждого из которых фиксируются значения заданного набора из n признаков. Таким образом, исходными данными служит набор из n p-мерных векторов:

$$X_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{pmatrix}, \; X_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2p} \end{pmatrix}, \ldots, \; X_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{np} \end{pmatrix},$$

которые, так же как и в кластерном анализе, можно представить в виде матрицы

$$\chi = \begin{pmatrix} x_{11} & x_{12} & \ldots x_{1n} \\ x_{21} & x_{22} & \ldots x_{2n} \\ x_{p1} & x_{p2} & \ldots x_{pn} \end{pmatrix} = \left(X_1, X_2, \ldots, X_n\right).$$

При этом предполагается, что все данные подвергнуты нормированию и центрированию в соответствии с (3.2.1).

Допустим, есть основания полагать, что количество этих признаков избыточно и рассматриваемый феномен в принципе может быть охарактеризован меньшим набором из m ($m < n$) определяющих признаков, которые и будем в дальнейшем называть основными факторами, или просто – факторами:

$$F_1 = \begin{pmatrix} f_{11} \\ f_{12} \\ \vdots \\ f_{1p} \end{pmatrix}, \; F_2 = \begin{pmatrix} f_{21} \\ f_{22} \\ \vdots \\ f_{2p} \end{pmatrix}, \ldots, \; F_m = \begin{pmatrix} f_{m1} \\ f_{m2} \\ \vdots \\ f_{mp} \end{pmatrix}.$$

Основным предположением линейной модели факторного анализа является предположение о том, что признаки выражаются через факторы линейно:

$$X_i = \sum_{k=1}^{m} a_{ik} F_k + a_i U_i \,, \qquad i = 1, 2, \ldots, n, \qquad (4.1.1)$$

где U_i – некоторые «добавки», введение которых обусловлено строгим равенством, с одной стороны, и тем фактом, что m-мерный базис из факторов не обязательно окажется полным для исходного описания явления через n векторов-признаков.

Факторы F_k называются общими факторами, а переменные U_i специфическими факторами («специфический» – это лишь один из переводов применяемого в англоязычной литературе слова «unique», в отечественной литературе в качестве определения U_i используются также «характерный», «уникальный»). Значения a_{ik} называются факторными нагрузками. Каждое из уравнений системы (4.1.1) является векторным и может быть заменено на p скалярных уравнений.

Таким образом, (4.1.1) есть система pn скалярных уравнений. Задача, подлежащая решению, – определение факторных нагрузок a_{ik}, а затем и самих факторов F_k. Очевидно, ее решение будет неоднозначным. По постановке задачи факторы должны быть линейно независимыми. Если набор F_k образует полный базис, то все U_i равны нулю. В противном случае в (4.1.1) будут присутствовать U_i, отличные от нуля.

Такая ситуация возникает в случаях, когда число факторов меньше размерности пространства косвенных признаков, а признаки не могут быть линейно выражены через факторы. Последнее свидетельствует о неточности линейной модели. В любом случае, естественно предположить, что U_i и F_k линейно независимы.

Для дальнейшего анализа важно подробнее рассмотреть величины дисперсий и корреляции в рамках модели линейного факторного анализа.

4.2. Дисперсия, коэффициенты корреляции признаков и их составляющие

Согласно принятой модели (4.1.1), дисперсию признаков можно разложить на две составляющие [26]: общность, обусловленную общими факторами F_k, и характерность, обусловленную характерными факторами U_i. Напомним, что в силу нормировки (3.2.1) оценка дисперсии признаков σ_k^2 равна единице, а средняя $\bar{x}_k = 0$.

$$\sigma_k^2 = \frac{1}{p}\sum_{i=1}^{p}(x_{ki} - \bar{x}_k)^2 = \frac{1}{p}\sum_{i=1}^{p}(x_{ki})^2 = 1.$$

После подстановки в последнее выражение (4.1.1) и элементарных преобразований с учетом некоррелированности факторов, получим:

$$1 = \sigma_k^2 = a_{k1}^2 + a_{k2}^2 + ... + a_{km}^2 + a_k^2 = h_k^2 + a_k^2. \qquad (4.2.1)$$

Такое представление дисперсии признака дает явное представление смысла факторных нагрузок: чем больше величина $-1 \le a_{ki} \le 1$, тем в большей степени k-й признак определяется i-м фактором. Степень неточности модели определяется величиной a_k^2 – влияние неучтенных факторов; общность представляет собой часть дисперсии переменных, объясненную факторами; специфичность – часть, не объясненную факторами. Величину a_k^2 иногда представляют в виде суммы

$$a_k^2 = b_k^2 + e_k^2,$$

где b_k^2 – специфичность; e_k^2 – ненадежность [20]. Величина b_k^2 характеризует влияние того факта, что среди общих факторов не содержатся все необходимые, то есть имеет место объективный недоучет принципиальных причинно-следственных связей.

Величина e_k^2 характеризует субъективную ошибку, связанную с неточностью измерений. В соответствии с постановкой задачи, необходимо искать такие факторы, при которых суммарная общность максимальна, а специфичность – минимальна.

Коэффициенты корреляции признаков X_k и X_l определяются соотношением ($\sigma_k = \sigma_l = 1$):

$$r_{kl} = \frac{1}{p}\sum_{i=1}^{p}(x_{ki} - \bar{x}_k)(x_{li} - \bar{x}_l) = \frac{1}{p}\sum_{i=1}^{p}x_{ki}x_{li}.$$

Аналогично предыдущему нетрудно получить:

$$r_{kl} = a_{k1}a_{l1} + a_{k1}a_{l1} + ... + a_{k1}a_{l1}, \qquad (4.2.2)$$

$$\hat{r}_{kl} = a_{kl}, \quad (4.2.3)$$

где $\hat{r}_{kl}$ – коэффициент корреляции между признаком X_l и основным фактором F_k. Коэффициенты корреляции и дисперсии вычисляются по исходным данным, а это означает, что соотношения (4.2.1) – (4.2.3) могут быть использованы при подсчете величины факторных нагрузок.

Наглядное представление о смысле факторных нагрузок дает геометрическая интерпретация, предложенная в [26] (рис. 4.2.1).

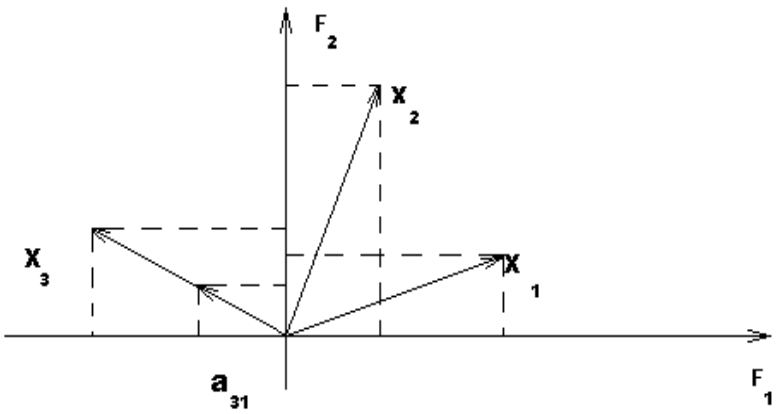


Рис. 4.2.1. Геометрическая интерпретация разложения косвенных признаков по основным факторам

На рис. 4.2.1 представлен случай разложения трех косвенных признаков по двум линейно независимым факторам. Величина, равная проекции единичного направляющего вектора признака на линию фактора, есть факторная нагрузка соответствующего признака на этот фактор. Так, на рисунке величина проекции единичного направляющего вектора X_3 на фактор F_1 обозначена через a_{31} – это и есть факторная нагрузка третьего признака на первый фактор.

4.3. Алгоритмы решения задачи факторного анализа. Метод главных компонент

Существует несколько методов решения задачи факторного анализа. Класс геометрических методов, среди которых наиболее известен центроидный метод, основывается на геометрической интерпретации, приведенной в разделе 4.2. Основным недостатком этого метода является то, что факторы вычисляются неоднозначно. Это связано с тем, что на каждом его этапе осуществляется так называемая процедура обращения знаков, результат которой зависит от субъективного выбора исследователя.

Метод максимального правдоподобия трактует задачу факторного анализа как вариационную. Оптимальная модель определяется через поиск максимума некоторой вспомогательной функции. Достаточное распространение получили также методы минимальных остатков и групповой. Подробное описание этих алгоритмов можно найти, например, в [26]. Однако в большинстве практических исследований применяется метод главных компонент.

Метод главных компонент состоит в последовательном поиске факторов. Первоначально модель содержит такое же число факторов F_k (главных компонент), что и косвенных признаков, что позволяет отказаться от введения специфических факторов U_i. В отличие от (4.1.1) модель главных компонент задается в виде:

$$X_i = \sum_{k=1}^{n} a_{ik} F_k + a_i , \quad i = 1, 2, \ldots, n, \qquad (4.3.1)$$

где a_i – свободный член.

Очевидно, если стандартизация уже произведена (F_k и X_i центрированы), то a_i, имеющая смысл оценки математического ожидания признака, равна нулю. Идея метода главных компонент заключается в поиске ортогональной системы из n векторов со специальными свойствами. Таким образом, в этой новой системе координат ковариационная матрица должна иметь диагональную форму (ограничиваемся здесь случаем, когда все собственные значения матрицы ковариаций простые).

Обозначим ковариационную матрицу (после нормирования $\sigma_k = \sigma_l = 1$ она будет совпадать с корреляционной матрицей) признаков через R:

$$R = \left(\operatorname{cov}(X_k, X_l)\right)_{k,l=1}^{n} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{ni} & r_{n2} & \cdots & 1 \end{pmatrix}.$$

Находим собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_n$ и собственные векторы матрицы R в главных осях (составленных из собственных векторов), матрица примет вид:

$$\widehat{R} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

Понятно, что сумма всех λ_i – след матрицы $tr\widehat{R} = \sum_{i=1}^{n} \lambda_i = n$, являясь инвариантом преобразования, будет равна n. Новые оси, полученные из исходных поворотом (вращением), ортогональны – соответствующие им главные компоненты (новые признаки) не коррелируют.

Главных факторов по-прежнему n, однако не все они одинаково важны с точки зрения рассматриваемой задачи. Собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_n$ являются дисперсиями. Те главные факторы, которым соответствуют большие значения дисперсии, объясняют большую часть разнообразия рассматриваемого набора объектов, они важны для описания системы, а те, дисперсия которых мала, не принципиальны для рассмотрения, ими можно пожертвовать, выигрывая в понижении размерности.

Для выбора факторов, обеспечивающих заданную точность модели α, расположим собственные значения в порядке их убывания: $\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_n$ и будем добавлять к первому, самому значимому, фактору ($\lambda_1 = \max\lambda_i$) последующие, пока $\sum_{i=1}^{n} \lambda_i \le (1-\alpha)n$.

Оставшиеся главные компоненты не интересны с точностью до заданного α. По умолчанию обычно в пакетах предусмотрено продолжать учитывать факторы, пока соответствующие им собственные значения превышают единицу: $\lambda_к > 1$.

Напомним, что переменные стандартизованы, и поэтому нет смысла строить очередной фактор, если он объясняет часть дисперсии меньшую, чем приходящаяся непосредственно на одну переменную. Выделив, таким образом, определяющие факторы, которые совпадают со значимыми главными компонентами, на основании (4.2.3) можно подсчитать факторные нагрузки.

4.4. Пример. Выявление основных факторов, влияющих на объем продаж. Реализация факторного анализа в пакете «Statistica»

Рассмотрим реализацию алгоритма факторного анализа в рамках статистических пакетов «Statistica 6.0» и «SPSS 11.5 for Windows» на конкретном примере. Допустим, необходимо выделить основные факторы, влияющие на объем продаж товаров *a – f*. Как и ранее в качестве исходного материала выбраны данные (табл. 4.4.1), полученные из годового отчета об объемах продаж одной екатеринбургской фирмы.

Таблица 4.4.1

Объемы продаж товаров *a – f*

Месяц	Наименование товара					
	a	*b*	*c*	*d*	*e*	*f*
Янв.	0	58	19	23	23	0
Фев.	0	23	11	8	17	0
Март	0	29	13	8	15	0
Апр.	0	37	12	11	25	0
Май	0	30	18	3	24	0
Июнь	0	38	28	9	15	0
Июль	0	36	24	1	18	0
Авг.	0	55	39	12	38	6
Сент.	2	62	37	46	31	1
Окт.	1	48	51	41	22	1
Нояб.	0	84	77	31	25	0
Дек.	0	124	76	56	25	0

Для проведения процедуры факторного анализа в пакете «Statistica 6.0» в электронной таблице пакета введем данные, относящиеся к товару *a*, в первую колонку, к товару *b* – во вторую, и т. д. (рис. 4.4.1).

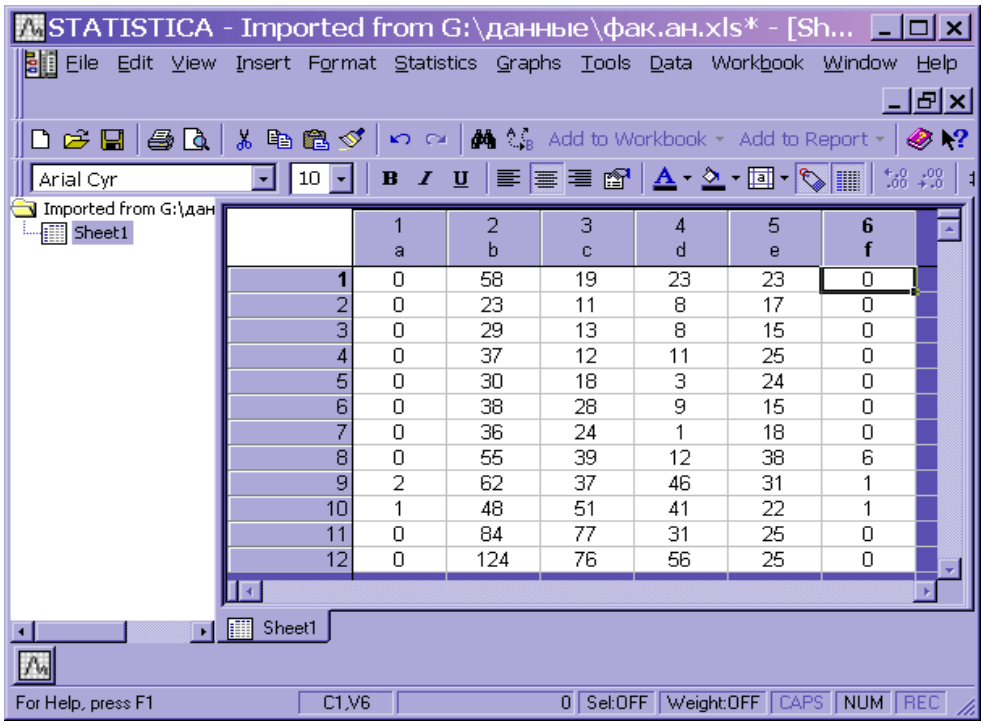


Рис. 4.4.1. Данные для факторного анализа в пакете «Statistica 6.0»

Для перехода в подпрограмму, реализующую факторный анализ, в верхнем меню следует выбрать пункт **Statistics/Multivariate Exploratory Techniques/Factor Analysis**. В появившемся на экране диалоговом окне в подменю **Variables** выбрать все переменные, необходимые для анализа (рис. 4.4.2).

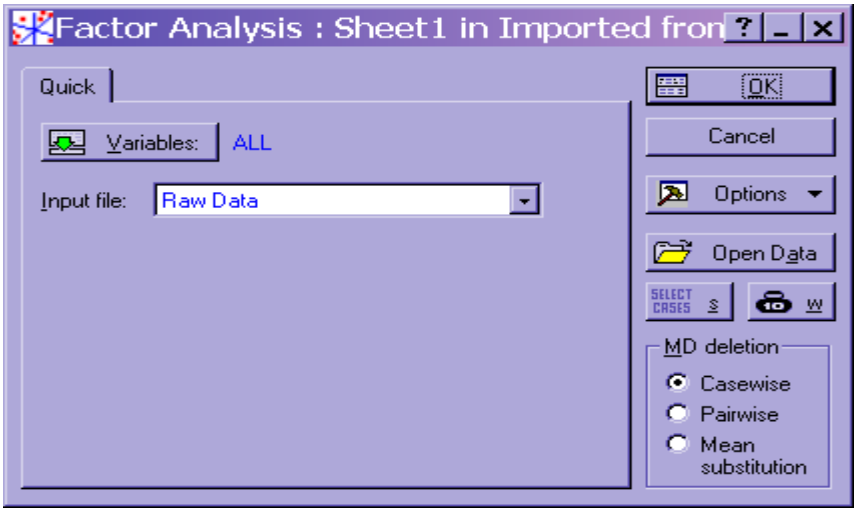


Рис. 4.4.2. Главное меню факторного анализа в пакете «Statistica 6.0»

После инициализации процедуры в появившемся диалоговом окне (рис. 4.4.3) требуется задать максимальное количество общих факторов.

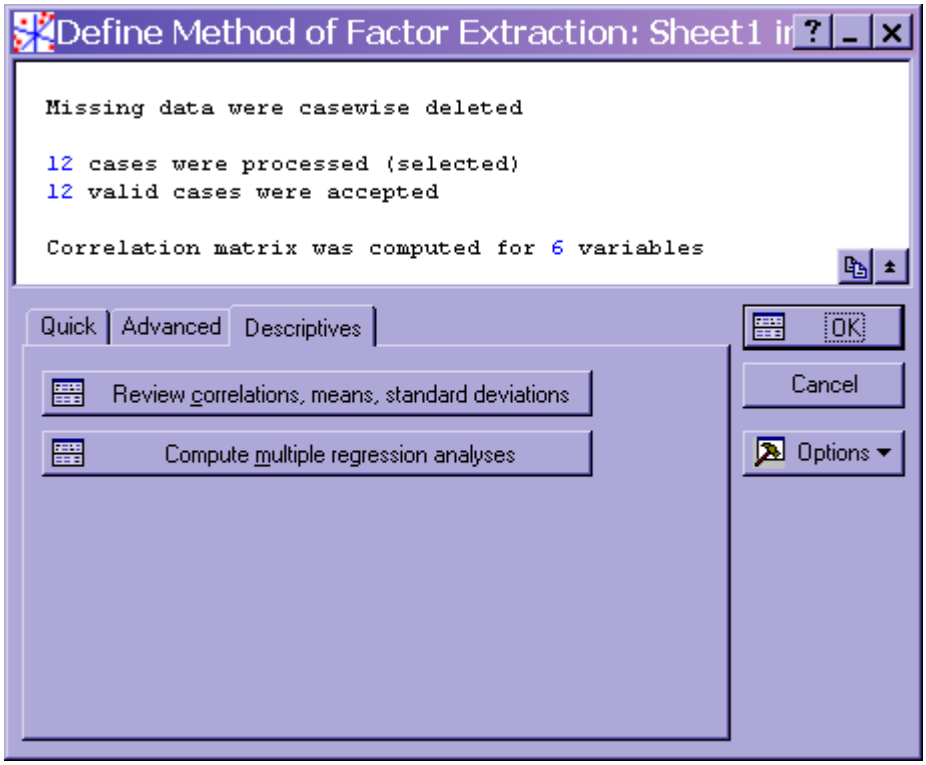


Рис. 4.4.3. Меню выбора метода анализа в пакете «Statistica 6.0»

Для запуска процедуры алгоритма нужно нажать на **Descriptives/Review correlation, means, standard deviations.**

В новом окне (рис. 4.4.4) требуется заказать подсчет попарных коэффициентов корреляции между всеми товарами *a – f.*

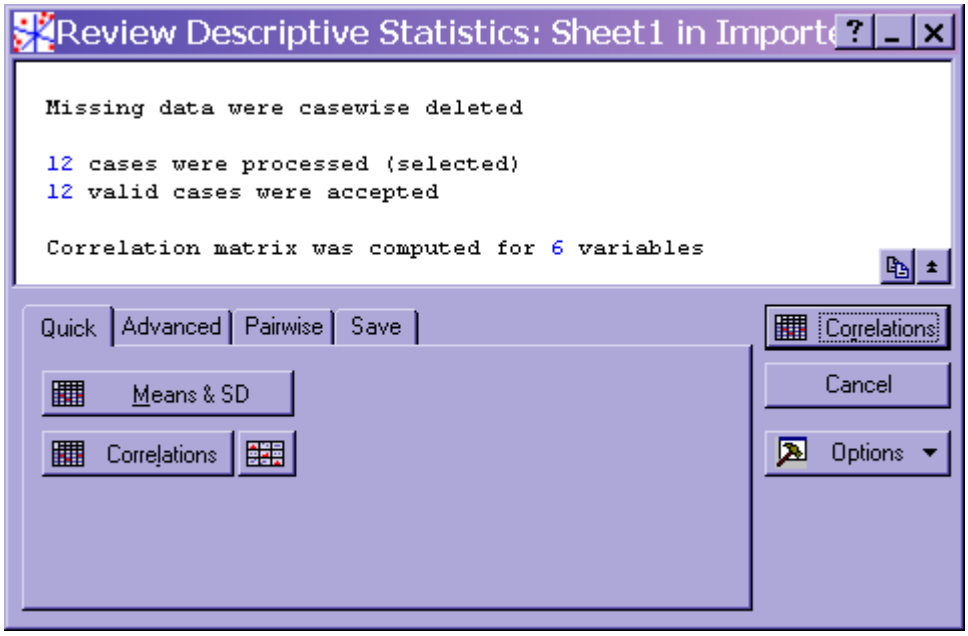


Рис. 4.4.4. Меню подсчета матрицы корреляций в пакете «Statistica 6.0»

Результатом такой операции является корреляционная матрица (рис. 4.4.5).

Correlations (Sheet1 in Imported from G:\дан
Casewise deletion of MD
N=12

Variable	a	b	c	d	e	f
a	1,00	0,08	0,15	0,56	0,32	0,08
b	0,08	1,00	0,87	0,82	0,40	0,04
c	0,15	0,87	1,00	0,76	0,35	0,12
d	0,56	0,82	0,76	1,00	0,35	-0,02
e	0,32	0,40	0,35	0,35	1,00	0,75
f	0,08	0,04	0,12	-0,02	0,75	1,00

Рис. 4.4.5. Матрица корреляций в пакете «Statistica 6.0»

С помощью матрицы корреляций можно установить степень линейной зависимости между наблюдаемыми переменными. Так, из приведенной выше матрицы видно, что степень зависимости между элементами, то есть между объемами продаж товаров *b*, *c*, *d* и *e*, *f* достаточно высока. Последнее явно свидетельствует о том, что имеется возможность перейти к описанию процесса на основании меньшего числа факторов, нежели общее число товаров.

После подсчета матрицы корреляций возвращаемся в меню выбора метода анализа (рис. 4.4.3) для определения факторного метода. Для этого нужно, щелкнув на кнопке **Advanced**, установить режим **Principal components** – метод главных компонент. Как было указано, максимальное количество факторов, которое уже установлено, будет совпадать с числом переменных (в нашем случае их 6).

Минимальную величину собственного значения, **Mini eigenvalue**, до которого проводится процедура, положим равной 1,25 (напомним, эта величина не должна быть менее единицы). После применения таких настроек, экран примет вид, показанный на рис. 4.4.6.

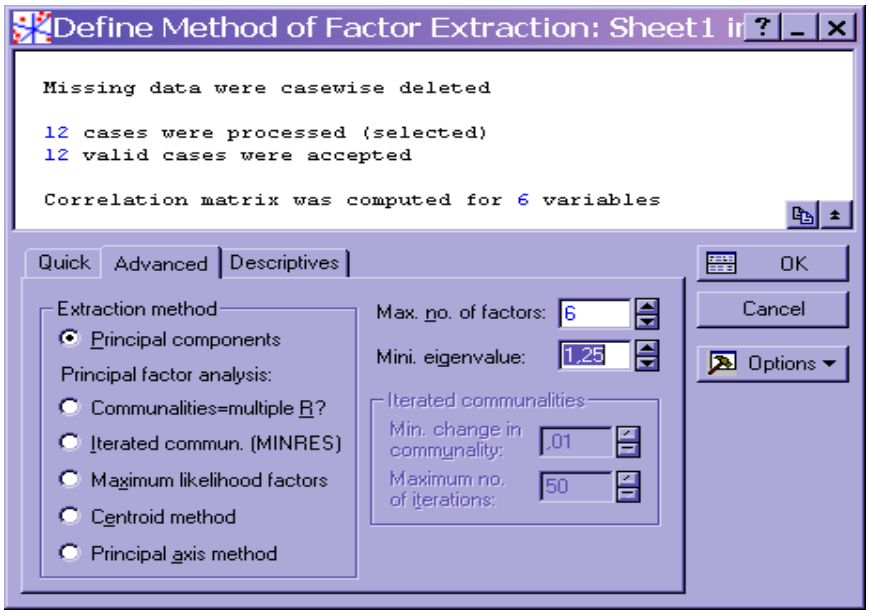


Рис. 4.4.6. Установка режимов проведения факторного анализа в пакете «Statistica 6.0»

Теперь можно осуществить запуск процедуры факторного анализа нажатием кнопки **ОК**. Отчет о результатах работы алгоритма появляется в окне, вид которого показан на рис. 4.4.7.

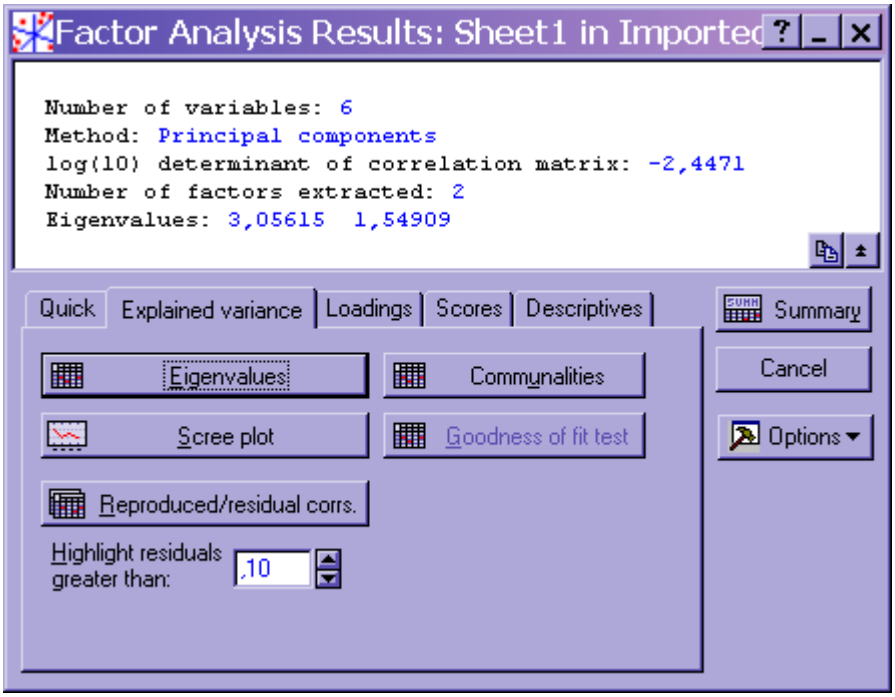


Рис. 4.4.7. Отчет о результатах работы факторного анализа в пакете «Statistica 6.0»

Из отчета видно, что обработке методом главных компонент подвергалось 6 признаков. В результате было выделено два основных фактора, отвечающих наибольшим собственным значениям корреляционной матрицы: $\lambda_1 = 3{,}05615$ и $\lambda_2 = 1{,}54909$, то есть на два этих фактора приходится наибольшая часть (почти 77 %) объясненной дисперсии.

Более подробную информацию об этих значениях можно получить, щелкнув по кнопке **Eigenvalues.** Вид появляющегося при этом окна показан на рис. 4.4.8.

Value	Eigenvalues (Sheet1 in Imported from G:\данные\фак.ан.xls) Extraction: Principal components			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,056147	50,93579	3,056147	50,93579
2	1,549086	25,81809	4,605233	76,75388

Рис. 4.4.8. Информация о наибольших собственных значениях корреляционной матрицы

На долю первого фактора приходится более 50 % объясненной суммарной дисперсии (50,93579), на долю второго – почти 26 % (25,81809). Матрица факторных нагрузок приведена на рис. 4.4.9, она определяет влияние факторов на наблюдаемые переменные.

Variable	Factor Loadings (Unrotated) (Sheet1 Extraction: Principal components (Marked loadings are > ,700000)	
	Factor 1	Factor 2
a	0,456854	0,095331
b	0,873990	-0,305266
c	0,862897	-0,267355
d	0,898081	-0,316561
e	0,656161	0,690423
f	0,319193	0,893553
Expl.Var	3,056147	1,549086
Prp.Totl	0,509358	0,258181

Рис. 4.4.9. Матрица факторных нагрузок

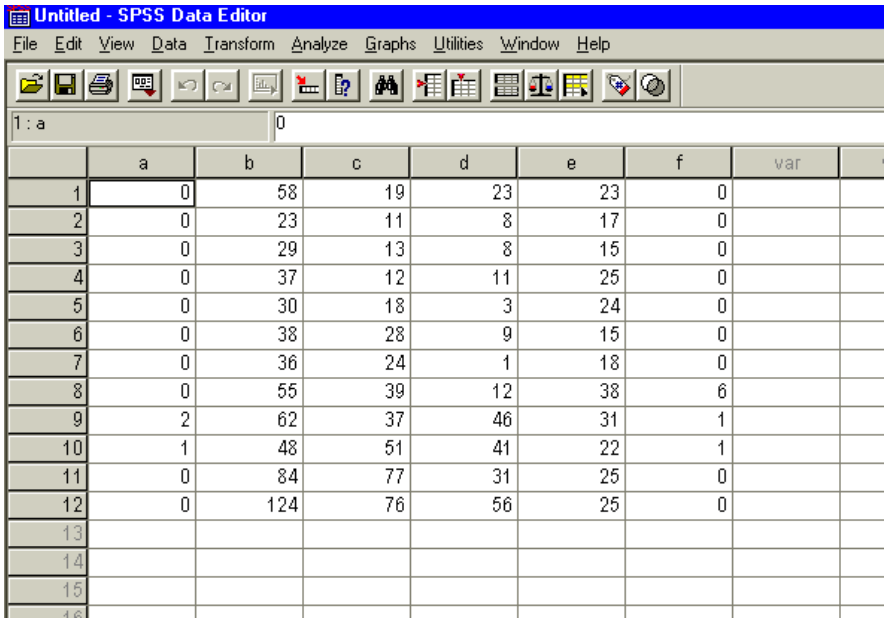
В нашем примере первый фактор оказывает наибольшее влияние на признаки *b*, *c*, *d*, а второй – на признак *f*. На основании матрицы факторных нагрузок можно составить модель вида (4.1.1):

$$
\begin{aligned}
a &= 0{,}456854 f_1 + 0{,}0955331 f_2, \\
b &= 0{,}872990 f_1 + 0{,}305266 f_2, \\
c &= 0{,}862897 f_1 + 0{,}237355 f_2, \\
d &= 0{,}898081 f_1 + 0{,}316561 f_2, \\
c &= 0{,}656161 f_1 + 0{,}690423 f_2, \\
f &= 0{,}456854 f_1 + 0{,}0955331 f_2 .
\end{aligned}
$$

Содержательная трактовка вводимых факторов выходит за рамки математических методов и здесь не рассматривается. Тем не менее, такая задача принципиальна для приложений, ее обсуждение можно найти, например, в [1, 12].

4.5. Реализация факторного анализа в пакете SPSS

Реализация факторного анализа в SPSS аналогична ее реализации в пакете «Statistica». Отметим некоторые особенности SPSS на том же примере. Ввод данных для начала процедуры осуществляется с помощью электронной таблицы (рис. 4.5.1) (Следует отметить, что как «Statistica», так и SPSS допускают экспорт данных из «Excel».)



	a	b	c	d	e	f	var
1	0	58	19	23	23	0	
2	0	23	11	8	17	0	
3	0	29	13	8	15	0	
4	0	37	12	11	25	0	
5	0	30	18	3	24	0	
6	0	38	28	9	15	0	
7	0	36	24	1	18	0	
8	0	55	39	12	38	6	
9	2	62	37	46	31	1	
10	1	48	51	41	22	1	
11	0	84	77	31	25	0	
12	0	124	76	56	25	0	
13							
14							
15							

Рис. 4.5.1. Данные для однофакторного анализа в пакете SPSS

Для запуска процедуры в верхнем меню следует выбрать пункт **Analyze/Data Reduction/Factor…**. В результате появится следующее окно (рис. 4.5.2).

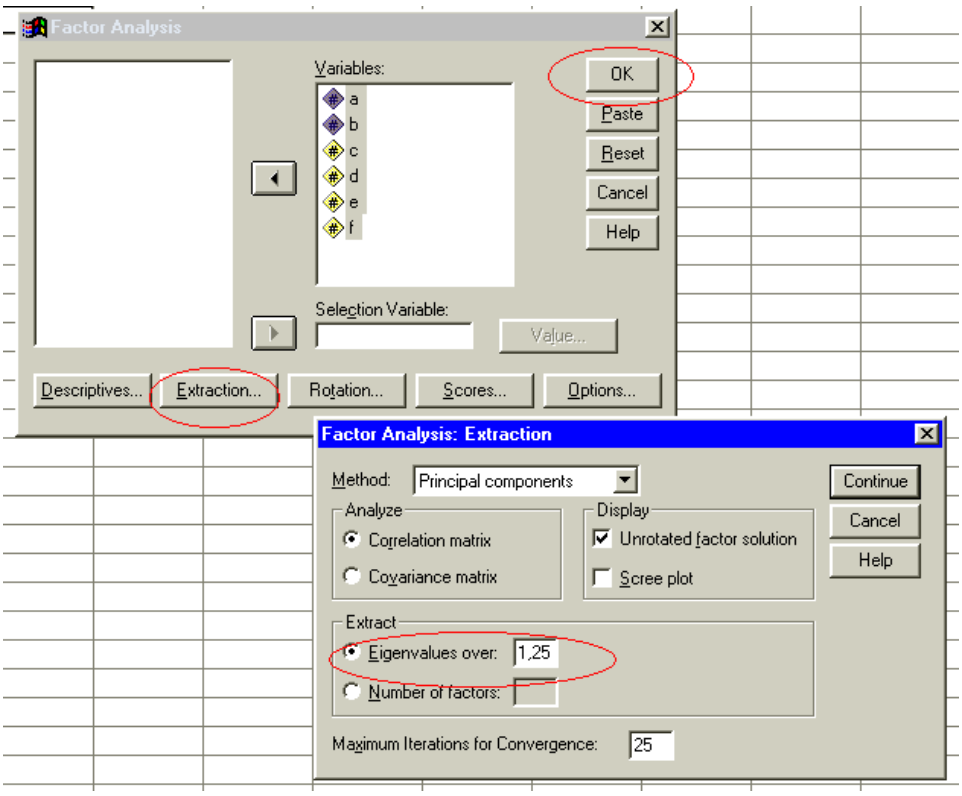


Рис. 4.5.2. Меню разделения переменных

В этом меню следует задать все косвенные признаки, подлежащие факторизации, указать предельное значение для собственных значений (как и ранее, положим $\lambda_{крит.} = 1{,}25$), выбрать алгоритм факторного анализа (**Method/Principal component analysis**) и поставить «флажок», требующий вывода матрицы корреляций (**Correlation matrix**). После этого программа начинает расчет. Результаты выводятся в отдельном окне **Output1** в виде таблицы (табл. 4.5.1).

Таблица 4.5.1

Матрица корреляций (Correlation Matrix) в пакете SPSS

Correlation	*a*	*b*	*c*	*d*	*e*	*f*
a	1,000	,082	,149	,561	,317	,085
b	,082	1,000	,870	,815	,399	,045
c	,149	,870	1,000	,758	,347	,117
d	,561	,815	,758	1,000	,351	-,020
e	,317	,399	,347	,351	1,000	,755
f	,085	,045	,117	-,020	,755	1,000

Значения корреляционной матрицы, полученные в SPSS, совпадают со значениями, полученными в пакете «Statistica» (рис. 4.4.5). Результаты факторного анализа, полученные в «Statistica» (рис. 4.4.8, 4.4.9), в SPSS оформляются в виде аналогичных итоговых таблиц (табл. 4.5.2, 4.5.3).

Таблица 4.5.2

Таблица результатов факторного анализа (Total Variance Explained) в пакете SPSS

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,056	50,936	50,936	3,056	50,936	50,936
2	1,549	25,818	76,754	1,549	25,818	76,754
3	1,036	17,270	94,023			
4	,217	3,613	97,636			
5	,112	1,863	99,499			
6	,030	,501	100,000			

Вторая колонка табл. 4.5.2 содержит величины собственных значений, третья – процент объясненной дисперсии (Variance), четвертая – накопленный процент объяснения дисперсии (Cumulative), подсчитанный нарастающим итогом. Все эти показатели подсчитаны для исходных собственных значений (Initial Eigenvalues), по всем косвенным признакам. Те же показатели приведены в колонках 5–7, но только для выделенных факторов (Extraction Sums of Squared Loadings).

Таблица 4.5.3

Матрица факторных нагрузок (Component Matrix(a)) в пакете SPSS

	Component	
	1	2
a	,457	,095
b	,874	-,305
c	,863	-,267
d	,898	-,317
e	,656	,690
f	,319	,894

Как и в «Statistica», получаем вместо шести косвенных признаков (*A*–*F*) два основных фактора (component).

ЛИТЕРАТУРА

1. *Благуш П. Б.* Факторный анализ с обобщениями. – М.: Финансы и статистика, 1989.
2. *Боровиков В.* STATISTICA. Искусство анализа данных на компьютере. – СПб.: Питер, 2003.
3. *Вуколов Э. А.* Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. – М.: ИНФРА-М, 2004.
4. *Глас Дж., Стэнли Дж.* Статистические методы в педагогике и психологии. – М.: Прогресс, 1976.
5. *Додж М., Стинсон К.* Эффективная работа с Microsoft Excel 2000. – СПб: Питер, 2001.
6. *Доугерти К.* Введение в эконометрику. – М.: ИНФРА-М, 1997.
7. *Дубнов П. Ю.* Обработка статистической информации с помощью SPSS. – М.: Прогресс, 2004.
8. *Дубров А. М.* Многомерные статистические методы. – М.: Финансы и статистика, 2003.
9. *Дюк В.* Обработка данных на ПК в примерах. – СПб.: Питер, 1997.
10. *Дюран Б., Одел П.* Кластерный анализ. – М.: Статистика, 1977.
11. *Елисеева И. И., Юзбашев М. М.* Общая теория статистики. – М.: Финансы и статистика, 1996.
12. *Жуковская В. М., Мучник И. Б.* Факторный анализ в социально-экономических исследованиях. – М.: Статистика, 1976.
13. *Заде Л., ДезоерЧ.* Теория линейных систем. Метод пространства состояний. – М.: Мир, 1970.
14. *Иберла В. Г.* Факторный анализ. – М.: Статистика, 1980.
15. *Карлберг К.* Бизнес-анализ с помощью Microsoft Excel. – М.: ИД «Вильямс», 2002.
16. *Колемаев В. А., Калинина В. Н.* Теория вероятностей и математическая статистика. – М.: ИНФРА-М, 2001.
17. *Лоули Д., Максвелл А.* Факторный анализ как статистический метод. – М.: Мир, 1967.
18. *Орехова Н. А., Левин А. Г., Горбунова Е. А.* Математические методы и модели в экономике. – М.: ЮНИТИ, 2004.
19. *Плис А. И.* Практикум по прикладной статистике в среде SPSS. – М.: Финансы и статистика, 2003.
20. *Плюта В.* Сравнительный многомерный анализ в экономических исследованиях. – М.: Статистика, 1980.
21. *Пугачев В. С.* Теория вероятностей и математическая статистика. – М.: Наука, 1979.
22. *Пугачев В. С., Казаков И. Е., Евланов Л. Г.* Основы статистической теории автоматических систем. – М.: Машиностроение, 1974.
23. Справочник по прикладной статистике. В 2 тт. / Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. – М.: Финансы и статистика, 1989, 1990.
24. *Тюрин Ю. Н., Макаров А. А.* Статистический анализ данных на компьютере. – М.: ИНФРА-М, 1998.
25. *Уотшем Т. Д., Паррамоу К.* Количественные методы в финансах. – М.: ЮНИТИ, 1999.
26. *Харман Г.* Современный факторный анализ. – М.: Статистика, 1972.
27. *Четыркин Е. М., Калихман И. Л.* Вероятность и статистика. – М.: Финансы и статистика, 1982.

ПРИЛОЖЕНИЕ 1

Таблица критических значений числа разбиений $S(\alpha, \beta)$

β / α	0,20	0,10	0,05	0,01	0,001	0,0001
0,20	8	11	14	21	31	42
0,10	16	22	29	44	66	88
0,05	32	45	59	90	135	180
0,01	161	230	299	459	689	918
0,001	1626	2326	3026	4652	6977	9303
0,0001	17475	25000	32526	55000	75000	100000

ПРИЛОЖЕНИЕ 2

Таблица *F*-распределения для уровня значимости α = 0,10

k_1/k_2	1	2	3	4	5	6	7	8	∞
1	39,86346	49,50000	53,59324	55,83296	57,24008	58,20442	58,90595	59,43898	63,32812
2	8,52632	9,00000	9,16179	9,24342	9,29263	9,32553	9,34908	9,36677	9,49122
3	5,53832	5,46238	5,39077	5,34264	5,30916	5,28473	5,26619	5,25167	5,13370
4	4,54477	4,32456	4,19086	4,10725	4,05058	4,00975	3,97897	3,95494	3,76073
5	4,06042	3,77972	3,61948	3,52020	3,45298	3,40451	3,36790	3,33928	3,10500
6	3,77595	3,46330	3,28876	3,18076	3,10751	3,05455	3,01446	2,98304	2,72216
7	3,58943	3,25744	3,07407	2,96053	2,88334	2,82739	2,78493	2,75158	2,47079
8	3,45792	3,11312	2,92380	2,80643	2,72645	2,66833	2,62413	2,58935	2,29257
9	3,36030	3,00645	2,81286	2,69268	2,61061	2,55086	2,50531	2,46941	2,15923
10	3,28502	2,92447	2,72767	2,60534	2,52164	2,46058	2,41397	2,37715	2,05542
11	3,22520	2,85951	2,66023	2,53619	2,45118	2,38907	2,34157	2,30400	1,97211
12	3,17655	2,80680	2,60552	2,48010	2,39402	2,33102	2,28278	2,24457	1,90361
13	3,13621	2,76317	2,56027	2,43371	2,34672	2,28298	2,23410	2,19535	1,84620
14	3,10221	2,72647	2,52222	2,39469	2,30694	2,24256	2,19313	2,15390	1,79728
15	3,07319	2,69517	2,48979	2,36143	2,27302	2,20808	2,15818	2,11853	1,75505
16	3,04811	2,66817	2,46181	2,33274	2,24376	2,17833	2,12800	2,08798	1,71817
17	3,02623	2,64464	2,43743	2,30775	2,21825	2,15239	2,10169	2,06134	1,68564
18	3,00698	2,62395	2,41601	2,28577	2,19583	2,12958	2,07854	2,03789	1,65671
19	2,98990	2,60561	2,39702	2,26630	2,17596	2,10936	2,05802	2,01710	1,63077
20	2,97465	2,58925	2,38009	2,24893	2,15823	2,09132	2,03970	1,99853	1,60738
21	2,96096	2,57457	2,36489	2,23334	2,14231	2,07512	2,02325	1,98186	1,58615
22	2,94858	2,56131	2,35117	2,21927	2,12794	2,06050	2,00840	1,96680	1,56678
23	2,93736	2,54929	2,33873	2,20651	2,11491	2,04723	1,99492	1,95312	1,54903
24	2,92712	2,53833	2,32739	2,19488	2,10303	2,03513	1,98263	1,94066	1,53270
25	2,91774	2,52831	2,31702	2,18424	2,09216	2,02406	1,97138	1,92925	1,51760

ПРИЛОЖЕНИЕ 3

Измерения и шкалы

В широком смысле, измерение есть присваивание числовых значений элементам некоторого множества (объектам, свойствам, признакам, событиям и т.п.) или изменениям, в соответствии с определенными правилами. Другими словами, измерение можно определить как построение шкал посредством изоморфного отображения эмпирической системы с отношениями в численную систему с отношениями. «Измерить все, что измеримо, и сделать измеримым все, что таковым еще не является» – задача для естествознания, поставленная Г. Галилеем [24].

Различают четыре основных типа шкал (соответственно, измерений): номинальную (идентификация), порядковую (задание, в дополнение к предыдущей, на множестве порядка предпочтения), интервальную (задание на множестве, в дополнение к предыдущей, единицы) и шкалу отношений (задание на множестве, в дополнение к предыдущей, нуля). Различия между этими шкалами заключается в том, насколько измеренные в них величины допускают совершать над собой арифметические операции. Рассмотрим эти шкалы в порядке нарастания таковых возможностей.

Номинальная шкала (шкала наименований). Задание такой шкалы сводится к присваиванию каждому объекту (свойству, признаку) определенного обозначения или символа (численного, буквенного и т.д.). Например, классификация вкусовых качеств (1 – сладкое, 2 – горькое, 3 – кислое и т.д.), цвета видимого спектра (К – красный, З – зеленый, С – синий и пр.). Номинальная шкала определяет то, что разные свойства или признаки качественно отличаются друг от друга, но не подразумевает каких-либо количественных операций с ними (сложение, умножение и т.п.). Номинальная шкала допускает любые замены и перестановки численных (буквенных) обозначений.

Важным частным случаем номинальной шкалы является ***дихотомическая шкала.*** Дихотомическая шкала – это такая шкала наименований, в которой измерение объекта может принимать только одно из двух значений: 0 или 1. Например, мужчина – 1, женщина – 0 (да – нет, верующий – атеист, успевающий – неуспевающий).

Ранговая (ординарная, порядковая) шкала. Введение ранговой шкалы предполагает установление на множестве измерений порядка (ранжирование определенного признака или свойства) так, что на любых двух различных элементах A и B можно определить порядок (строгий порядок) предпочтения $A < B$ или $A > B$. Порядковое измерение возможно тогда, когда в объектах можно обнаружить различия в степени выраженности признака или свойства. Шкала порядка не предусматривает меры (степени) различий между элементами ряда, не допускает никаких перестановок внутри ранжированного ряда, невозможны арифметические операции (сложение, умножение и т.п.). Допустимая операция – реверсия шкалы.

Примерами рангового измерения могут служить: места, занятые спортсменами в соревновании; рейтинг студента по среднему баллу успеваемости; оценка в психодиагностическом тесте («Я спокоен, собран»: 1 – никогда, 2 – иногда, 3 – часто, 4 – всегда). Эта шкала более развита по сравнению с номинальной. Однако здесь очевидно присутствуют ограничения на использование арифметических действий. Действительно, быть «мисс» конкурса красоты лучше, чем «вице-мисс», но никому не придет в голову сказать, что «вице-мисс» в два раза менее красива, чем «мисс».

Интервальная шкала. Введение шкалы интервалов предполагает определение расстояния на множестве измерений (с единицей, масштабом). На интервальной шкале нет естественной точки отсчета (нуль условен, он не указывает на отсутствие измеряемого свойства). Интервальная шкала допускает операции нахождения разности, суммы и среднего значения, но не допускает нахождение произведений или отношений величин (умножение, деление, возведение в степень). Примеры интервальной шкалы: температурная шкала Цельсия; шкала уровня субъективного контроля (–3 – «абсолютно не согласен», 0 – «не знаю», +3 – «абсолютно согласен»).

Шкала отношений. Шкала отношений предполагает наличие естественного нуля, который означает полное отсутствие какого-либо свойства или признака. Это наиболее информативная шкала, она допускает любые математические операции и различные статистические приемы. Шкалами отношений являются большинство измерительных шкал физических характеристик (пространство, время, масса, объем, скорость и пр.).

Оперирование различными математическими методами предполагает изначальное определение типа шкалы. Если тип шкалы определен неверно, то исследователь может выбрать неадекватный метод статистической обработки и прийти в результате к неверным выводам. При подготовке к проведению статистических исследований следует обратить внимание на то, чтобы все наблюдаемые признаки допускали возможность измерения в одной из шкал.

ПРИЛОЖЕНИЕ 4

Оценка качества и оформления упаковки новой марки сигарет. Числовые результаты опроса

№	Возраст	Доход (тыс. руб.)	Качество	Оформление
1	35	37	0	5
2	45	72	2	3
3	37	120	-4	3
4	26	60	-1	-3
5	30	140	-2	-2
6	31	40	5	3
7	23	50	3	2
8	27	100	-3	2
9	40	98	-1	3
10	38	65	-1	2
11	38	70	2	-2
12	34	125	-5	-5
13	43	95	4	3
14	33	40	3	4
15	19	30	-1	2
16	36	42	2	-3
17	31	48	1	3
18	56	70	3	-4
19	23	12	-1	2
20	34	96	1	2
21	26	64	-1	-1
22	37	36	0	-1
23	38	60	-2	-1
24	46	144	3	2
25	42	84	3	-2
26	29	180	-2	-2
27	41	120	0	-1
28	44	54	3	3
29	34	96	3	1
30	59	96	3	1
31	44	18	-5	0
32	37	13	-1	-1
33	27	200	-2	-2
34	23	96	-2	2
35	29	180	-3	-5
36	36	70	0	-2
37	31	130	-1	4
38	42	95	1	-4
39	31	80	-5	-3
40	26	36	0	3
41	57	39	-3	-2
42	43	48	-3	-4
43	27	30	1	1
44	45	32	1	-2
45	45	15	-5	0

46	47	100	5	-5
47	51	15	3	3
48	48	90	2	2
49	50	65	3	0
50	21	56	2	2
51	22	90	3	3
52	30	180	3	3
53	21	30	0	3
54	48	42	0	0
55	24	96	3	3
56	27	60	0	0
57	23	120	2	3
58	25	55	0	2
59	28	175	4	4
60	27	42	0	1
61	38	36	3	-2
62	33	96	-2	-2
63	34	240	0	1
64	38	120	3	3
65	41	200	-4	-2
66	44	30	2	-2
67	34	120	1	3
68	34	300	2	-2
69	33	120	-2	-2
70	56	120	1	2
71	25	18	0	2
72	37	144	-1	4
73	33	102	2	1
74	44	72	4	5
75	23	54	3	1
76	53	60	3	-3
77	34	85	2	0
78	35	36	3	1
79	42	140	1	0
80	45	120	3	3
81	34	96	2	0
82	21	45	2	2
83	25	65	2	2
84	36	41	0	0
85	25	125	3	-1
86	55	62	1	0
87	35	100	2	1
88	39	85	3	1
89	37	180	4	4
90	42	160	3	-4
91	38	72	2	2
92	31	96	1	2
93	34	24	-3	-3
94	36	124	-2	-2
95	39	60	1	0
96	42	60	-5	-5
97	44	48	2	1
98	23	24	-2	3
99	36	72	-3	-3
100	45	239	3	3