Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Кафедра математики, информатики
и математических методов в экономике

А.Ю. Трусова

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Часть 1

Утверждено Редакционно-издательским советом университета в качестве учебного пособия

Самара
Издательство «Самарский университет»
2008

ББК 65.051я73
УДК 311(076.1)
Т 78

Рецензент доц. Л.К. Ширяева

Т 78 **Трусова А.Ю.**
Многомерные статистические методы: учебное пособие для студентов факультета экономики и управления: в 2 ч. Ч.1. / А.Ю. Трусова; Федер. агентство по образованию. – Самара: Изд-во «Самарский университет», 2008. – 67 с.

Данная работа содержит конспект лекций отдельных глав курса «Многомерные статистические методы»: многомерный корреляционный анализ, проверка гипотез в многомерном статистическом анализе, дискриминантный анализ, кластерный анализ.

Предназначено для студентов 3 курса специальности «Математические методы в экономике», может быть использовано студентами вузов, обучающихся по экономическим, управленческим, социологическим и психологическим специальностям и направлениям.

ББК 65.051я73
УДК 311(076.1)

ВВЕДЕНИЕ

Социально - экономические процессы и явления зависят от большого числа параметров, их характеризующих, что обусловливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. Методы многомерного статистического анализа используются при изучении стохастической информации, т.е. в ситуации, когда решение принимается на основе неполной информации.

Многомерный статистический анализ позволяет выбрать из множества вероятностно-статистических моделей ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надёжность и точность выводов, сделанных на основании ограниченного статистического материала.

Проведение системного анализа до изучения взаимосвязей в многомерной совокупности требует иметь представление о связях между отдельной зависимой переменной и группой влияющих на неё показателей. Это может быть осуществлено при помощи множественного корреляционного анализа.

Методы многомерной классификации, которые предназначены для разделения совокупности объектов на однородные группы, используют большое количество разных стохастически связанных признаков. В этих случаях используются методы кластерного и дискриминантного анализов.

Многомерный статистический анализ представляет собой неотъемлемую часть фундаментальных курсов университетского образования и активно используется в аналитической практике. В теоретическом плане многомерный статистический анализ представляет собой дальнейшее развитие традиционной одномерной статистики, его отличают трудоёмкие алгоритмы реализации вычислительных процедур, практически всегда рассчитанные на привлечение технических средств, и сложная интерпретируемость аналитических результатов. Это требует от пользователя достаточно серьёзной подготовки как в области математической статистики, так и в области, в которой проводятся конкретные исследования.

ГЛАВА 1
МНОГОМЕРНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

В многомерном корреляционном анализе изучается связь между группой признаков X_1, X_2, X_3, ..., X_m. Изучая связь между парами признаков X_i и X_j, находится коэффициент парной корреляции r_{ij}. Если найти все возможные коэффициенты корреляции r_{ij}, то в результате получается набор данных, которыми являются коэффициенты корреляции r_{ij}. Упорядоченное значение всех коэффициентов корреляции представляется в виде матрицы корреляции (R). На главной диагонали матрицы корреляции располагаются единицы. Матрица корреляции R симметрична относительно главной диагонали, так как $r_{12} = r_{21}$. Матрица корреляций имеет вид:

$$\underset{m \times m}{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1m} \\ r_{21} & 1 & r_{23} & \ldots & r_{2m} \\ r_{31} & r_{32} & 1 & \ldots & r_{3m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{m1} & r_{m2} & r_{m3} & \ldots & 1 \end{pmatrix}$$

В многомерном корреляционном анализе рассматриваются две типовые задачи:

1. Определение тесноты связи одной из переменных с совокупностью остальных ($m - 1$) переменных, включенных в анализ.
2. Определение тесноты связи между переменными при фиксировании или исключении влияния других k переменных, где $k < m - 2$.

Эти задачи решаются с помощью множественных и частных коэффициентов корреляции.

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Теснота линейной взаимосвязи одной переменной X_i с совокупностью других (m-1) переменных X_j, рассматриваемой в целом, измеряется с помощью *множественного* (или *совокупного*) коэффициента корреляции R_{j0}, который является обобщением парного коэффициента корреляции r_{ij}. Выборочный множественный, или совокупный, коэффициент корреляции вычисляется по формуле:

$$R_{j0} = \sqrt{1 - \frac{|R|}{R_j}},$$

где $|R|$ – определитель матрицы корреляции R; R_j – алгебраическое дополнение элемента r_{jj} матрицы корреляции (равного 1).

Множественный коэффициент корреляции изменяется от 0 до 1, он не меньше, чем абсолютная величина любого парного или частного коэффициента корреляции с таким же первичным индексом. Если R стремится

к 1, то делается вывод о тесной линейной взаимосвязи между признаком X_j и всеми остальными признаками, но направление этой связи нельзя определить с помощью множественного коэффициента корреляции.

Величина R_{jo}^2 называется выборочным множественным коэффициентом детерминации и показывает, какая часть вариации исследуемой переменной объясняется вариацией остальных переменных.

Множественный коэффициент корреляции значимо отличается от нуля, если наблюдаемое значение статистики $F_{набл} = \frac{R_{jo}^2 (n-m)}{(1-R_{jo}^2)(m-1)}$ больше критического значения статистики $F_{кр}$ (α, k_1, k_2), $k_1 = m - 1$, $k_2 = n - m$. Значение критической статистики $F_{кр}$ определяется по таблице распределения Фишера-Снедекора.

ЧАСТНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Если переменные коррелируют друг с другом, то на величине парного коэффициента корреляции частично сказывается величина других переменных. В связи с этим возникает необходимость исследовать частную корреляцию между переменными при элиминировании влияния одной или нескольких других переменных. *Выборочным частным коэффициентом корреляции* между переменными X_i и X_j при фиксированных значениях остальных (m-2) переменных называется выражение:

$$r_{ij \bullet k_s} = \frac{-A_{ij}}{\sqrt{A_{ii} A_{jj}}},$$

где A_{ij} – алгебраическое дополнение элемента r_{ij} матрицы корреляций R.

Например: $r_{13\bullet} = \frac{-A_{13}}{\sqrt{A_{11} A_{33}}}$. Знак коэффициенту частной корреляции присваивается согласно знаку соответствующего коэффициента регрессии в линейной модели.

Для определения частного коэффициента корреляции любого порядка l (от 0 до $m - 2$) следует рассмотреть подматрицу ($l + 2$) – порядка матрицы R, составленную из строк и столбцов, отвечающих индексам вычисляемого коэффициента, а далее к подматрице применяется формула:

$$r_{ij \bullet k_s} = \frac{-A_{ij}}{\sqrt{A_{ii} A_{jj}}}.$$

Рассмотрим пример вычисления частного коэффициента корреляции $r_{34/26}$. Составим подматрицу размерности 4 ×4, содержащую коэффициенты пар-

ной корреляции между признаками X_2, X_3, X_4 и X_6: $\begin{vmatrix} 1 & r_{23} & r_{24} & r_{26} \\ r_{23} & 1 & r_{34} & r_{36} \\ r_{24} & r_{34} & 1 & r_{46} \\ r_{26} & r_{36} & r_{46} & 1 \end{vmatrix}$, тогда частный коэффициент корреляции $r_{34/26} = \frac{-A_{34}}{\sqrt{A_{33}A_{44}}}$.

Проверка значимости частного коэффициента корреляции: H_o: $r_{ij} = 0$, H_1: $r_{ij} \neq 0$. Наблюдаемое значение статистики критерия вычисляется по формуле: $t_{набл} = \frac{|r_{ij}|\sqrt{n-m+2}}{\sqrt{1-r_{ij}^2}}$, $t_{кр}(\alpha,k)$ с числом степеней свободы $k=n-m+2$ определяется по таблице распределения Стьюдента.

Вывод: частная корреляция между признаками считается незначимой, если $t_{набл}<t_{кр}$, в противном случае - значимо отличной от нуля ($t_{набл} > t_{кр}$).

ПОНЯТИЕ О РАНГАХ И ИХ ПОСТРОЕНИЕ

На практике довольно часто встречаются числовые данные в выборке, которые носят в определенном смысле условный характер. Это могут быть экспертные оценки, тестовые баллы, данные о каких-либо предпочтениях исследуемой группы людей (например, политических) и т.д. При анализе таких данных часто невозможно соблюсти все предпосылки применения классических статистических методов, которые подразумевают принадлежность выборки одному из известных законов распределения. Часто очень трудно доказать закон распределения (скажем, в силу недостаточного количества наблюдений). В таких случаях делать научно обоснованные выводы, применяя методы прикладной статистики, можно, лишь используя не сами значения (например, баллы), а их порядок, основанный на соотношении «меньше – больше». Порядок значений называют рангами. *Рангом наблюдения называют номер, который получит это наблюдение в упорядоченной совокупности всех данных после упорядочения их согласно определенному правилу (например, от меньшего значения к большему). Ранжирование* – это процедура перехода от совокупности наблюдений к последовательности их рангов. Результат ранжирования называют *ранжировкой*. Рассмотрим процесс ранжирования на примере. Допустим, у нас есть выборка, состоящая из пяти чисел: 8, 25,42, 3, 1. Этим значениям будут присвоены соответствующие ранги: 3, 4, 5, 2, 1. При ранжировании возникают случаи, когда невозможно найти существенные различия между объектами по величине проявления рассматриваемого признака. Говорят, что объекты оказываются связанными. Связанным объектам приписывают одинаковые средние ранги такие, чтобы сумма всех рангов осталась такой же, как и при отсутствии связанных рангов. Совокупность элементов вы-

борки, имеющих одинаковое значение, называют *связкой*, а количество одинаковых значений в связке – её *размером*. Средним рангом является среднее арифметическое рангов элементов связки, которые бы они имели, если бы одинаковые элементы связки оказались различны. Например, пусть дана выборка чисел: 15, 17, 12, 15, 7, 8, 5, 1, 8. Этим значениям будут соответствовать ранги: 7,5; 9;7,5; 6; 3; 4,5; 2; 1; 4,5.

Статистические методы, которые используют ранги для получения научно обоснованных выводов из анализируемых данных, называются ранговыми. Эти методы широко применяют там, где очень сложно (или невозможно) выяснить, какому закону распределения соответствуют анализируемые данные.

РАНГОВАЯ КОРРЕЛЯЦИЯ

На практике существует необходимость изучения связи между ординальными (порядковыми) переменными, измеренными в так называемых порядковых шкалах. В этой шкале можно установить лишь порядок, в котором объекты выстраиваются по степени проявления признака. В таких случаях проблема оценки тесноты связи разрешима, если упорядочить, или ранжировать объекты по степени выраженности измеряемых признаков. На ранговых данных выясняется теснота связи – ранговая корреляция.

КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

Коэффициент ранговой корреляции Спирмена определяется по формуле:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left(rang(X_i) - rang(X_j)\right)^2}{n^2(n-1)},$$

где r_s – коэффициент ранговой корреляции Спирмена, *rang* (X_i), *rang* (X_j) – ранги, полученные для признаков X_i и X_j соответственно, n – объём выборки (количество измерений). При наличии связанных рангов коэффициент ранговой корреляции Спирмена определяется по формуле:

$$r_S = 1 - \frac{\sum_{i=1}^{n}\left(rang(X_i) - rang(X_j)\right)^2}{\frac{1}{6}\left(n^3 - n\right) - \left(T_{X_i} + T_{X_j}\right)},$$

где $T_{X_i} = \frac{1}{12}\sum_{i=1}^{m_1}\left(t_{X_i}{}^3 - t_{X_i}\right)$, $T_{X_j} = \frac{1}{12}\sum_{i=1}^{m_2}\left(t_{X_j}{}^3 - t_{X_j}\right)$, t_{Xi} – количество рангов, входящих в группу неразличимых рангов по переменной X_i, t_{Xi} – количест-

во рангов, входящих в группу неразличимых рангов по переменной X_j, m_1 и m_2 – количество групп неразличимых рангов у переменных X_i и X_j.
Проверка на значимость коэффициента ранговой корреляции Спирмена.
H_0: $r_S = 0$, H_1: $r_S \neq 0$,

$t_{набл} = \dfrac{|r_S|\sqrt{n-2}}{\sqrt{1-r_S^2}}$, $t_{кр}$ определяется по таблице распределения Стьюдента на уровне значимости α с числом степеней свободы k, где $k = n - 2$, $t_{кр}(\alpha; k)$.

Вывод: если $t_{набл} < t_{кр}$ – коэффициент ранговой корреляции Спирмена не значим на уровне α, если $t_{набл} > t_{кр}$ – коэффициент ранговой корреляции Спирмена значим на уровне α.

Рассмотрим *пример*. По результатам тестирования 10 студентов по двум дисциплинам А и В на основе набранных баллов получили следующие ранги:

rang X_i	2	4	5	1	7,5	7,5	7,5	7,5	3	10
rang X_j	2,5	6	4	1	2,5	7	8	9,5	5	9,5

По дисциплине А имеем $m_1 = 1$ – одну группу неразличимых рангов с $t_{Xi}=4$; по дисциплине В - $m_2=2$ –две группы неразличимых рангов с $t_{Xi}=2$. Поэтому

$$T_{X_1} = \frac{1}{12}\left(4^3 - 4\right) = 5, \quad T_{X_2} = \frac{1}{12}\left[\left(2^3 - 2\right) + \left(2^3 - 2\right)\right] = 1,$$

$$r_s = 1 - \frac{39}{\frac{1}{6}\left(10^3 - 10\right) - (5+1)} = 0{,}755.$$

Проверка на значимость. $t_{набл} = \dfrac{0{,}775\sqrt{8}}{\sqrt{1-0{,}775^2}} = 3{,}26$, $t_{кр}(0{,}05; 8) = 2{,}31$. Вывод: так как $t_{набл} > t_{кр}$ коэффициент ранговой корреляции Спирмена значим на 5% уровне.

КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ КЕНДАЛЛА (τ)

Для вычисления коэффициента ранговой корреляции Кендалла используется формула:

$$\tau = 1 - \frac{4K}{n(n-1)},$$

где K – статистика Кендалла (число инверсий). Инверсии – это нарушение порядка. Порядок означает, что большее число стоит справа от меньшего.

Нарушение прядка (инверсия) – это такое распределение чисел, когда справа располагается меньшее число. Для определения числа инверсий K объекты по одному из признаков ранжируются по возрастанию рангов. По другому признаку вычисляется количество инверсий с учетом полученной ранжировки. При полном совпадении двух ранжировок $K = 0$, $\tau = 1$. При полной противоположности двух ранжировок $\tau = -1$, во всех остальных случаях $-1 \le \tau \le 1$.

При проверке значимости τ исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными (при n > 10) τ имеет приближенно нормальный закон распределения с математическим ожиданием, равным нулю, и средним квадратическим отклонением $S_\tau = \sqrt{\frac{9n(n-1)}{2(2n+5)}}$. Поэтому τ значим на уровне α, если значение статистики $t_{набл} = |\tau|\sqrt{\frac{9n(n-1)}{2(2n+5)}}$ больше критического. Значение критической статистики $t_{кр}$ определяется из условия $\Phi(t_{кр}) = \frac{1-\alpha}{2}$.

Рассмотрим *пример*. Два эксперта проранжировали 10 предложенных им проектов реорганизации НПО с точки зрения их эффективности при заданных ресурсных ограничениях.

Эксперт 1:	1	2	3	4	5	6	7	8	9	10
Эксперт 2:	2	3	1	4	6	5	9	7	8	10
Число инверсий	1	1	0	0	1	0	2	0	0	0

$K = 1 + 1 + 1 + 2 = 5$,

$\tau = 1 - \frac{4 \cdot 5}{10(10-1)} = 1 - \frac{2}{9} = \frac{7}{9} \approx 0{,}77$. Проверка на значимость: $t_{набл} = 0{,}77\sqrt{\frac{90 \cdot 9}{2 \cdot 25}}$,

$t_{кр} = 1{,}96$, при $\alpha = 0{,}05$. Вывод: коэффициент ранговой корреляции Кендалла значимо отличен от нуля на 5% уровне.

Если ранги связаны, формула имеет вид:

$$\tau_{св} = \frac{1 - \frac{2(T_{X_1} - T_{X_2})}{n(n-1)}}{\sqrt{\left(1 - \frac{2T_{X_1}}{n(n-1)}\right)\left(1 - \frac{2T_{X_2}}{n(n-1)}\right)}},$$

где $T_{X_i} = \frac{1}{2}\sum_{i=1}^{n_i}\left(t_{X_i}^{\,2} - t_{X_i}\right)$.

Пример.

Десять однородных предприятий подотрасли были проранжированы по степени прогрессивности их организационных структур (признак X_1), по

эффективности их функционирования в отчетном году (признак X_2). Получены следующие ранжировки.

1	2	2	4	4	6	6	8	9	9
1	2	4	4	4	4	8	8	8	10

Выявить коэффициент связанных рангов.

$$T_{X_1} = \frac{1}{2}\left[\left(2^2-2\right)+\left(2^2-2\right)+\left(2^2-2\right)+\left(2^2-2\right)\right] = 4,$$

$$T_{X_1} = \frac{1}{2}\left[\left(4^2-4\right)+\left(3^2-3\right)\right] = 9,$$

$$\tau_{cв} = \frac{1-\frac{2(4+9)}{10(10-1)}}{\sqrt{\left(1-\frac{2\cdot 4}{10\cdot 9}\right)\left(1-\frac{2\cdot 9}{10\cdot 9}\right)}} = \frac{1-\frac{13}{45}}{\sqrt{\frac{41}{45}\cdot\frac{4}{5}}} = \frac{\frac{32}{45}}{\frac{2\sqrt{41}}{3\cdot 5}} = \frac{16}{3\sqrt{41}} = 0{,}83.$$

Коэффициенты Спирмена и Кендалла связаны соотношением $r_S = \frac{3}{2}\tau$ при $n > 10$.

КОЭФФИЦИЕНТ КОНКОРДАЦИИ (СОГЛАСОВАНИЯ) РАНГОВ КЕНДАЛЛА (*W*)

В случаях, когда совокупность характеризуется не двумя, а несколькими последовательностями рангов (ранжировками) и необходимо установить статистическую связь между несколькими переменными (например, в экспертных оценках), используется коэффициент конкордации (согласования) рангов Кендалла:

$$W = \frac{12\sum_{i=1}^{n} D^2}{m^2\left(n^3-n\right)},$$

где $D = \left(\sum_{i=1}^{m} r_{ij}\right) - \frac{m(n-1)}{2}$, n – число объектов; m – число анализируемых порядковых переменных. Коэффициент конкордации (согласования) рангов Кендалла $0 \le W \le 1$, причем $W=1$ при совпадении всех ранжировок.

Проверка значимости коэффициента конкордации W основана на том, что в случае справедливости нулевой гипотезы H_o: $W = 0$ (при конкурирующей гипотезе H_1: $W \neq 0$) об отсутствии корреляционной связи при n > 7 статистика $m(n-1)W$ имеет приближенно χ^2 - распределение. Таким образом, $\chi^2_{набл} = m(n-1)W$, $\chi^2_{кр} = (\alpha; k)$, $k = n-1$.

Вывод: $\chi^2_{набл} > \chi^2_{кр}$ – W значимо отличается от 0, т.е. присутствует согласование по рангам.

Пример. Группа из 5 экспертов оценивает качество изделий, изготовленных на 7 предприятиях. Их предпочтения представлены в таблице. Вычислить коэффициент конкордации (согласования) рангов Кендалла и оценить его значимость на уровне $\alpha = 0{,}05$.

Эксперт (m)	Предприятие i (n)							
	1	2	3	4	5	6	7	Итого
1	1	3	4	2	6	7	5	
2	1	2	5	3	6	4	7	
3	2	1	7	5	6	4	3	
4	1	2	4	6	3	5	7	
5	3	1	5	4	2	6	7	
Сумма	8	9	25	20	23	26	29	140
								Ранг $= \frac{140}{7} = 20$
D	-12	-4	5	0	3	6	9	
D^2	144	121	25	0	9	36	81	416

$$W = \frac{12 \cdot 416}{5^2\left(7^3 - 7\right)} = 0{,}594.$$

Проверка значимости W: $\chi^2_{набл} = 5 \cdot 6 \cdot 0{,}594 = 17{,}83$, $\chi^2_{кр}(0{,}05;6) = 12{,}59$, $\chi^2_{набл} > \chi^2_{кр}$ – коэффициент конкордации значим, т.е. существует тесная согласованность мнений экспертов.

КОРРЕЛЯЦИЯ КАТЕГОРИЗИРОВАННЫХ ПЕРЕМЕННЫХ

Признак называют *категоризованным*, если его «возможные» значения описываются конечным числом состояний (*категорий*, *градаций*). Статистический анализ парных связей между категоризованными переменными X_i и X_j производится на базе исходных данных, представленных в виде так называемых двухвходовых таблиц сопряженности следующего типа:

Градации признака X_i	Градации признака X_j						
	1	2	…	j	…	k	Сумма в строке
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1k}	n_1
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2k}	n_2
…	…	…	…	…	…	…	…
i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ik}	n_i
…	…	…	…	…	…	…	…
m	n_{m1}	n_{m2}	…	n_{mj}	…	n_{mk}	n_m
Сумма в столбце	m_1	m_2	…	m_j	…	m_k	n

В таблице n_{ij} означает число объектов (из общего числа n обследованных), у которых «значение» признака X_i зафиксировано на уровне i-й градации, а значение признака X_j – на уровне j-й градации.

КРИТЕРИЙ χ^2 О НЕЗАВИСИМОСТИ КЛАССИФИКАЦИИ В ТАБЛИЦЕ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

Наблюдаемое значение статистики критерия ХИ-квадрат определяется по формуле:

$$\chi^2_{набл} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}},$$

где $\tilde{n}_{ij}$ – ожидаемая (теоретическая) частота. Критическое значение определяется на уровне значимости α с числом степеней свободы ν по таблице распределения χ^2. $\chi^2_{кр}(\alpha;\nu)$, $\nu = (m-1)(k-1)$, k– количество столбцов, m – количество строк.

Пример.
Среди 190 человек исследовалось мнение относительно какого-то определенного вопроса А. Выделим в выборке 3 независимых категории по возрасту. Рассмотрим следующие гипотезы:
H_0: не существует различие мнений относительно вопроса A среди разных возрастных групп.
H_1: существует различие мнений относительно вопроса A среди разных возрастных групп.

Ответы респондентов	Возраст респондентов, лет			
	Старше 40	25 – 40	Младше 25	Сумма
"Категорически не согласен"	*(а)* 18	*(б)* 13	*(в)* 10	41
"Не согласен"	*(г)* 23	*(д)* 13	*(ж)* 12	48
"Согласен"	*(з)* 11	*(и)* 14	*(к)* 23	48
"Совершенно согласен"	*(л)* 8	*(м)* 16	*(н)* 29	53
Сумма	60	56	74	190

Вспомогательная таблица:

Ячейка	n_i	$\tilde{n}_i$	$\frac{(n_{ij}-\tilde{n}_{ij})^2}{\tilde{n}_{ij}}$
а	18	12,9	2,02
б	13	12,1	0,07
в	10	16	2,25
г	23	15,2	4
д	13	14,1	0,08
ж	12	18,7	2,4
з	11	15,2	1,16
и	14	14,1	0
к	23	18,7	0,99
л	8	16,7	4,53
м	16	15,6	0,01
н	29	20,6	3,42
$\chi^2_{набл}$			20,94

$\chi^2_{кр}$ (0,05; 6) = 16,812. Вывод: $\chi^2_{набл} > \chi^2_{кр}$ – можно говорить о том, что существует различие мнений относительно вопроса *А*.

ГЛАВА 2
ПРОВЕРКА ГИПОТЕЗ В МНОГОМЕРНОМ СТАТИСТИЧЕСКОМ АНАЛИЗЕ

В многомерном статистическом анализе рассматриваются следующие гипотезы:

Многомерная случайная величина	Нулевые гипотезы	Конкурирующие гипотезы
$\overline{X}$ – вектор средних значений; μ – вектор постоянных значений	H_0: $\overline{X}_1 = \overline{X}_2$ H_0: $\overline{X} = \mu$	H_1: $\overline{X}_1 \neq \overline{X}_2$ H_1: $\overline{X} \neq \mu$
Σ – матрица ковариаций	H_0: $\Sigma_1 = \Sigma_2$	H_1: $\Sigma_1 \neq \Sigma_2$

Критериальная проверка многомерных гипотез основывается на теоретических подходах, принятых для одномерного случая.

ПРОВЕРКА ГИПОТЕЗ О РАВЕНСТВЕ ВЕКТОРА СРЕДНИХ ЗНАЧЕНИЙ ПОСТОЯННОМУ ВЕКТОРУ μ

Пусть исходная матрица данных имеет вид:

Многомерная случайная величина X	X_1	X_2	...	X_m
1	x_{11}	x_{12}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2m}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nm}

Вектор средних значений $\overline{X} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \dots \\ \overline{x_m} \end{pmatrix}$ сравнивается с постоянным вектором $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_m \end{pmatrix}$. Выдвигаемые гипотезы: $\begin{matrix} H_0 : \overline{X} = \mu \\ H_1 : \overline{X} \neq \mu \end{matrix}$.

Наблюдаемое значение критической статистики вычисляется с помощью T^2-критерия Хотеллинга: $T^2_{набл} = n\left(\overline{X} - \mu\right)^T S^{-1}\left(\overline{X} - \mu\right)$, где n – число наблюдений, S – выборочная матрица ковариаций, S^{-1} – обратная матрица к

выборочной матрице ковариаций. Элементы матрицы ковариаций по выборочным данным вычисляются с помощью соотношения

$$S = \frac{1}{n-1}(Z^T Z),$$

где Z – матрица центрированных данных, в которой каждый элемент $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$, $\bar{x}_j$ – среднее значение j-той компоненты случайной величины X, s_j – среднее квадратическое отклонение j-той компоненты случайной величины X. Критическое значение критерия вычисляется с помощью соотношения

$$T^2_{кр}(\alpha; k_1, k_2) = \frac{m(n-1)}{n-m} F(\alpha; k_1, k_2),$$

где $F(\alpha; k_1, k_2)$ – табличное значение F-критерия Фишера-Снедекора для уровня значимости α со степенями свободы k_1 и k_2 равными $k_1 = m$, $k_2 = n - m$. Многомерная гипотеза подтверждается при $T^2_{набл} < T^2_{кр}(\alpha; k_1; k_2)$ и не может быть принята, если $T^2_{набл} > T^2_{кр}(\alpha; k_1; k_2)$.

Приведенная формула T^2-критерия Хотеллинга является общей и рассчитана на проверку гипотезы сразу по всему числу m анализируемых признаков. Однако реально, даже при отрицании гипотезы H_0: $\bar{X} = \mu$, значения одних признаков могут существенно отличаться от некоторых постоянных значений, а другие – несущественно. Возникает необходимость проверки гипотезы по каждому отдельному признаку или нескольким признакам ($k<m$) при условии нивелирования значений остальных признаков. Для решения подобной задачи используется частный критерий Хотеллинга, который вычисляется по формуле:

$$T^2_{набл,j} = \frac{n\left[C_j^{\tau}\left(\bar{X} - \mu\right)\right]^2}{C_j^{\tau} S C^{\tau}},$$

где C_j – нивелирующий вектор. Компоненты вектора C_j – нули и единицы, единицы указывают на признак или признаки, по значениям которых осуществляется проверка гипотезы. Например, если анализируются три признака, то для проверки гипотезы поочередно используются:

$C_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $C_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $C_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ и $C_1^{\tau} = (1 \;\; 0 \;\; 0)$, $C_2^{\tau} = (0 \;\; 1 \;\; 0)$, $C_3^{\tau} = (0 \;\; 0 \;\; 1)$ соответственно.

Расчетные значения $T^2_{набл,j}$ сравниваются с критическим значением $T^2_{кр}(\alpha; k_1; k_2)$. Значения признаков существенно отличаются от некоторых постоянных значений, если $T^2_{набл,j} > T^2_{кр}(\alpha; k_1; k_2)$, и несущественно, если $T^2_{набл,j} < T^2_{кр}(\alpha; k_1; k_2)$.

ПРОВЕРКА ГИПОТЕЗ О РАВЕНСТВЕ ДВУХ ВЕКТОРОВ СРЕДНИХ ЗНАЧЕНИЙ

Пусть исходные матрицы данных имеют вид:

Многомерная случайная величина X_1	X_{11}	X_{21}	...	X_{m1}
1	x_{111}	x_{121}	...	x_{1m1}
2	x_{211}	x_{221}	...	x_{2m1}
...	...	...	...	...
n_1	x_{n111}	x_{n121}	...	x_{n1m1}

Многомерная случайная величина X_2	X_{12}	X_{22}	...	X_{m2}
1	x_{112}	x_{122}	...	x_{1m2}
2	x_{212}	x_{222}	...	x_{2m2}
...	...	...	...	...
n_2	x_{n212}	x_{n222}	...	x_{n2m2}

Векторы средних значений имеют вид: $\overline{X_1} = \begin{pmatrix} \overline{x_{11}} \\ \overline{x_{21}} \\ \dots \\ \overline{x_{m1}} \end{pmatrix}$ и $\overline{X_2} = \begin{pmatrix} \overline{x_{12}} \\ \overline{x_{22}} \\ \dots \\ \overline{x_{m2}} \end{pmatrix}$.

Выдвигаемые гипотезы:

$H_0 : \overline{X_1} = \overline{X_2}$.

$H_1 : \overline{X_1} \neq \overline{X_2}$

Наблюдаемое значение критической статистики вычисляется с помощью T^2 - критерия :

$$T^2_{набл} = \frac{n_1 n_2}{n_1 + n_2} \left(\overline{X_1} - \overline{X_2}\right)^{\mathrm{T}} \hat{S}^{-1} \left(\overline{X_1} - \overline{X_2}\right),$$

где n_1 – число наблюдений в первой таблице, n_2 – число наблюдений во второй таблице, $\overline{X_1}$ – вектор средних значений первой выборки, $\overline{X_2}$ – вектор средних значений второй выборки, $\hat{S}$ – несмещённая оценка обобщённой матрицы ковариаций, определяемая соотношением $\hat{S} = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2 - 2}$, S_1 и S_2 – матрицы ковариаций соответственно первой и второй выборок, $\hat{S}^{-1}$ – обратная матрица обобщённой матрицы ковариаций. Критическое значение вычисляется с помощью соотношения:

$$T^2_{кр}(\alpha; k_1, k_2) = \frac{(n_1 + n_2 - 2)m}{(n_1 + n_2 - m - 1)} F(\alpha; k_1, k_2),$$

где $F(\alpha; k_1, k_2)$ – табличное значение F-критерия Фишера-Снедекора для уровня значимости α со степенями свободы k_1 и k_2, равными $k_1 = m, k_2 = n_1 + n_2 - m - 1$. Многомерная гипотеза подтверждается при $T^2_{набл} < T^2_{кр}(\alpha; k_1; k_2)$ и не может быть принята, если $T^2_{набл} > T^2_{кр}(\alpha; k_1; k_2)$.

При этом также существует возможность расчета частных критериев $T^2_{набл.j}$ для сравнений одного или нескольких средних значений из каждой выборочной совокупности:

$$T^2_{набл.j} = \frac{n_1 n_2 \left(C_j^T \left(\overline{X_1} - \overline{X_2}\right)\right)^2}{(n_1 + n_2) C_j^T \hat{S} C_j},$$

где C_j –вектор, нивелирующий средние значения, не участвующие в сравнении, $1 \le j \le m$. Для частных оценок различий средних значений критические величины определяются формулой:

$$T^2_{кр}(\alpha; k_1; k_2) = \frac{(n_1 + n_2 - 2) j}{(n_1 + n_2 - j - 1)} \cdot F(\alpha; k_1; k_2),$$

где $k_1 = j$, $k_2 = n_1 + n_2 - j - 1$. Расчетные значения $T^2_{набл.j}$ сравниваются с критическим значением $T^2_{кр}(\alpha; k_1; k_2)$. Значения признаков существенно отличаются друг от друга, если $T^2_{набл.j} > T^2_{кр}(\alpha; k_1; k_2)$, и несущественно, если $T^2_{набл.j} < T^2_{кр}(\alpha; k_1; k_2)$.

ПРОВЕРКА ГИПОТЕЗ О РАВЕНСТВЕ КОВАРИАЦИОННЫХ МАТРИЦ

Сравнение ковариационных матриц, отражающих взаимосвязи изучаемых признаков, открывает возможность дополнить и уточнить гипотетические предположения относительно самих признаков. Это приобретает особенное значение, если принять во внимание, что даже специфические, индивидуальные характеристики признаков могут совпадать случайно.

В социальных и экономических исследованиях существует множество задач, требующих идентификации признаковых связей. Особенно часто они возникают при классификации наблюдаемых объектов, распознавании образов и т.п., например, при оценке кредитоспособности клиентов банков, группировке предприятий по уровню устойчивости финансового положения или при оценке эффективности производственной и коммерческой деятельности.

На практике учет ковариаций (корреляций) изучаемого комплекса признаков и проверка равенства матриц ковариаций значительно снижают возможность появления ошибки в выводах. Это происходит из-за весьма малой вероятности случайного совпадения одновременно большого числа сложных характеристик связей признаков.

Выдвигаемые гипотезы: H_0: $\Sigma_1 = \Sigma_2$ и H_1: $\Sigma_1 \neq \Sigma_2$. Наблюдаемое значение критической статистики определяется соотношением:

$$W_{набл} = b \ln \nu,$$

где $b = 1 - \left[\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2}\right] \frac{2m^2 + 3m - 1}{6(m+1)}$,

$$\ln \nu = (n_1 + n_2 - 2)\ln\left|\det \hat{S}\right| - \left((n_1 - 1)\ln\left|\det S_1\right| + (n_2 - 1)\ln\left|\det S_2\right|\right).$$

Критическое значение статистики вычисляется с помощью соотношения

$$W_{кр} = \chi^2(\alpha, k), \quad k = \frac{m(m+1)}{2}.$$

Нулевая гипотеза отвергается, если $W_{набл} > W_{кр}$, и принимается, если $W_{набл} < W_{кр}$.

ГЛАВА 3
ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Дискриминантный анализ – это раздел математики, содержанием которого является разработка методов решения задач разделения (дискриминации) объектов по определенным группам признаков, например, разбиение совокупности предприятий на несколько однородных групп по значениям каких-либо показателей производственно-хозяйственной деятельности. Задачей дискриминантного анализа является разделить неоднородную совокупность на структурные единицы. Разделение на однородные группы позволяет эффективно использовать моделирование зависимостей между отдельными признаками.

Методы дискриминанттного анализа находят применение в различных областях: экономике, медицине, социологии, психологии и т. д. Дискриминантный анализ оказывается очень удобным при обработке результатов тестирования отдельных лиц. Например, при выборе кандидатов на определенную должность можно всех опрашиваемых претендентов разделить на две группы: «подходит» и «не подходит». Еще один пример использования дискриминантного анализа в экономике: для оценки финансового состояния своих клиентов при выдаче им кредита банк классифицирует их на «надежных» и «ненадежных» по ряду признаков. Таким образом, в тех случаях, когда возникает необходимость отнесения того или иного объекта к одному из реально существующих или выделенных определенным способом классов, можно воспользоваться дискриминантным анализом.

ПОНЯТИЕ ДИСКРИМИНАНТНОЙ ФУНКЦИИ, ЕЕ ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ

На рис.1 изображены объекты, принадлежащие двум различным множествам M_1 и M_2 . Каждый объект характеризуется в данном случае двумя переменными X_1 и X_2, которые задают координаты этих объектов.

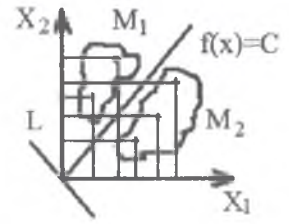

Рис.1. Геометрическая интерпретация дискриминантной функции и дискриминантных переменных

Если рассматривать координаты объектов (точек) по каждой оси, то нетрудно заметить, что эти множества пересекаются, т.е. по каждой переменной отдельно некоторые объекты обоих множеств имеют сходные характеристики. Чтобы наилучшим образом разделить два рассматриваемых множества, нужно иметь чёткую границу, например, в виде прямой, которая разделит данные группы. Для этого необходимо составить функцию, в которой переменные X_1 и X_2 были бы связаны числовыми коэффициентами. Таким образом, задача сводится к определению новой системы координат. Причем новые оси L и C должны быть расположены таким образом, чтобы координаты объектов, принадлежащих разным множествам, на ось L были максимально разделены. Ось C перпендикулярна оси L и разделяет два множества точек наилучшим образом, то есть чтобы множества оказались по разные стороны от этой прямой. Рассмотрим алгоритм нахождения границы C. Введём специальную функцию, которая зависит от начальных координат объектов X_1 и X_2 . Будем предполагать, что граница имеет линейный вид. Это самый простой случай определения границы между множествами. Функция имеет вид: $f(x)=a_1x_1+a_2x_2$.

Функция $f(x)$ называется дискриминантной функцией, а величины x_1 и x_2 – дискриминантными переменными. Как видно, функция линейно связывает координаты точек, коэффициенты a_1 и a_2 необходимо определить.

Для определения a_1 и a_2 введём $\overline{x_{ij}}$ – среднее значение j-й координаты у объектов i-го множества. Тогда для множества M_1 среднее значение функции $f_1(x)$,будет равно: $\overline{f_1}(x)=a_1\overline{x_{11}}+a_2\overline{x_{12}}$; для множества M_2 среднее значение функции $f_2(x)$ равно: $\overline{f_2}(x)=a_1\overline{x_{21}}+a_2\overline{x_{22}}$.

Геометрическая интерпретация этих функций – две параллельные прямые, проходящие через центры множеств (рис.2).

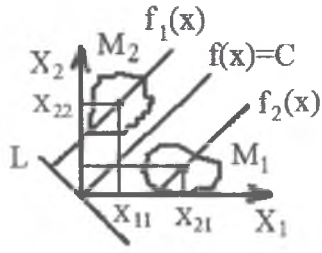

Рис.2. Центры разделяемых множеств и константа дискриминации

РАСЧЕТ КОЭФФИЦИЕНТОВ ДИСКРИМИНАНТНОЙ ФУНКЦИИ

Коэффициенты дискриминантной функции a_1 и a_2 определяются таким образом, чтобы $\overline{f_1}(x)$ и $\overline{f_2}(x)$ как можно больше различались между собой, т.е. чтобы для двух множеств было максимальным выражение:

$$\overline{f_1}(x) - \overline{f_2}(x) = \sum_{i=1}^{n_1} a_i x_{1i} - \sum_{i=1}^{n_2} a_i x_{2i},$$

где n_1 и n_2 – количество точек (объектов) первого и второго множеств соответственно.

Рассмотрим две группы множеств. В первой группе три объекта, во второй – два. Каждый объект задаётся двумя координатами X_1 и X_2. В общем виде таблицы исходных данных имеют вид:

	X_1	X_2
n_1	x_{111}	x_{112}
n_2	x_{211}	x_{212}
n_3	x_{311}	x_{312}

и

	X_1	X_2
n_1	x_{121}	x_{122}
n_2	x_{221}	x_{222}

,

где x_{ikj} – значение j -го признака для i-го объекта k -го множества. Первый индекс означает номер объекта в множестве, второй индекс – номер множества, третий индекс – номер координаты. Например, x_{111} означает значение первой координаты первого объекта для первого множества. Если подставить табличные значения в общую формулу для дискриминантной функции, то можно вычислить значение дискриминантной функции для каждого объекта изучаемых множеств. В общем виде значения дискриминантной функции для каждого объекта изучаемых множеств соответственно равны:

$$f_{11} = a_1 x_{111} + a_2 x_{112},$$
$$f_{12} = a_1 x_{211} + a_2 x_{212},$$
$$f_{13} = a_1 x_{311} + a_2 x_{312},$$
$$f_{21} = a_1 x_{121} + a_2 x_{122},$$
$$f_{22} = a_1 x_{221} + a_2 x_{222},$$

где f_{kt} – дискриминантная функция, в которой первый индекс (k)– номер множества, второй индекс (t) – номер объекта в данном множестве. Например, f_{21} – значение дискриминантной функции первого объекта второго множества. Вычислив значения дискриминантной функции для каждого объекта двух изучаемых множеств, можно рассчитать среднее значение дискриминантной функции для каждого множества по формуле средней арифметической. Таким образом, для каждого множества среднее значение дискриминантной функции задаётся следующими формулами:

$\overline{f_1} = \frac{1}{3}(f_{11} + f_{12} + f_{13})$, $\overline{f_2} = \frac{1}{2}(f_{21} + f_{22})$. Рассмотрим вычисления для первого множества:

$$\overline{f_1} = \frac{1}{3}(f_{11} + f_{12} + f_{13}) = \frac{1}{3}\left[(a_1 x_{111} + a_2 x_{112}) + (a_1 x_{211} + a_2 x_{212}) + (a_1 x_{311} + a_2 x_{312})\right] =$$

$$= \frac{1}{3}\left[a_1(x_{111} + x_{211} + x_{311}) + a_2(x_{112} + x_{212} + x_{312})\right] = a_1 \frac{(x_{111} + x_{211} + x_{311})}{3} +$$

$$+ a_2 \frac{(x_{112} + x_{212} + x_{312})}{3} = a_1 \overline{x_{11}} + a_2 \overline{x_{12}}.$$ Аналогично можно проделать вычисления для второго множества. Таким образом, получим

$$\overline{f_1} = a_1 \overline{x_{11}} + a_2 \overline{x_{12}},$$
$$\overline{f_2} = a_1 \overline{x_{21}} + a_2 \overline{x_{22}},$$

где $\overline{x_{kj}}$ – среднее значение j-го признака в k -м множестве. Вычислим разницу между значениями дискриминантной функции для каждого объекта и соответствующим средним значением дискриминантной функции:

$$f_{11} - \overline{f_1} = a_1(x_{111} - \overline{x_{11}}) + a_2(x_{112} - \overline{x_{12}});$$
$$f_{12} - \overline{f_1} = a_1(x_{211} - \overline{x_{11}}) + a_2(x_{212} - \overline{x_{12}});$$
$$f_{13} - \overline{f_1} = a_1(x_{311} - \overline{x_{11}}) + a_2(x_{312} - \overline{x_{12}});$$
$$f_{21} - \overline{f_2} = a_1(x_{121} - \overline{x_{21}}) + a_2(x_{122} - \overline{x_{22}});$$
$$f_{22} - \overline{f_2} = a_1(x_{221} - \overline{x_{21}}) + a_2(x_{222} - \overline{x_{22}}).$$

Отклонения значений дискриминантной функции для каждого объекта от среднего значения дискриминантной функции для соответствующего множества могут быть как положительными, так и отрицательными. Полученные значения для разницы необходимо возвести в квадрат и просуммировать, что позволит оценить вариацию дискриминантной функции внутри множеств. Таким образом, получим:

$$\sum_{k=1}^{2}\sum_{t=1}^{n_k}\left(f_{kt} - \overline{f_k}\right)^2 = \left(f_{11} - \overline{f_1}\right)^2 + \left(f_{12} - \overline{f_1}\right)^2 + \left(f_{13} - \overline{f_1}\right)^2 + \left(f_{21} - \overline{f_2}\right)^2 + \left(f_{22} - \overline{f_2}\right)^2,$$

следовательно, сумма квадратов отклонений примет вид:

$$\sum_{k=1}^{2}\sum_{t=1}^{n_k}\left(f_{kt} - \overline{f_k}\right)^2 = a_1^2(x_{111} - \overline{x_{11}})^2 + 2a_1(x_{111} - \overline{x_{11}})a_2(x_{112} - \overline{x_{12}}) + a_2^2(x_{112} - \overline{x_{12}})^2 +$$
$$+ a_1^2(x_{211} - \overline{x_{11}})^2 + 2a_1(x_{211} - \overline{x_{11}})a_2(x_{212} - \overline{x_{12}}) + a_2^2(x_{212} - \overline{x_{12}})^2 +$$
$$+ a_1^2(x_{311} - \overline{x_{11}})^2 + 2a_1(x_{311} - \overline{x_{11}})a_2(x_{312} - \overline{x_{12}}) + a_2^2(x_{312} - \overline{x_{12}})^2 +$$
$$+ a_1^2(x_{121} - \overline{x_{21}})^2 + 2a_1(x_{121} - \overline{x_{21}})a_2(x_{122} - \overline{x_{22}}) + a_2^2(x_{122} - \overline{x_{22}})^2 +$$
$$+ a_1^2(x_{221} - \overline{x_{21}})^2 + 2a_1(x_{222} - \overline{x_{21}})a_2(x_{222} - \overline{x_{22}}) + a_2^2(x_{222} - \overline{x_{22}})^2$$

С другой стороны, от исходных таблиц данных можно перейти к таблицам центрированных данных

	X_{1c}	X_{2c}
n_1	$x_{111}-\overline{x_{11}}$	$x_{112}-\overline{x_{12}}$
n_2	$x_{211}-\overline{x_{11}}$	$x_{212}-\overline{x_{12}}$
n_3	$x_{311}-\overline{x_{11}}$	$x_{312}-\overline{x_{12}}$

и

	X_{1c}	X_{2c}
n_1	$x_{121}-\overline{x_{21}}$	$x_{122}-\overline{x_{22}}$
n_2	$x_{221}-\overline{x_{21}}$	$x_{222}-\overline{x_{22}}$

.

Вычислим $X_{c1}{}^T\cdot X_{c1}$ и $X_{c2}{}^T\cdot X_{c2}$.

$$X_{c1}{}^T\cdot X_{c1}=\begin{pmatrix} x_{111}-\overline{x_{11}} & x_{211}-\overline{x_{11}} & x_{311}-\overline{x_{11}} \\ x_{112}-\overline{x_{12}} & x_{212}-\overline{x_{12}} & x_{312}-\overline{x_{12}} \end{pmatrix}\cdot\begin{pmatrix} x_{111}-\overline{x_{11}} & x_{112}-\overline{x_{12}} \\ x_{211}-\overline{x_{11}} & x_{212}-\overline{x_{12}} \\ x_{311}-\overline{x_{11}} & x_{312}-\overline{x_{12}} \end{pmatrix}=\begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix},$$

где $d_{11}=\left(x_{111}-\overline{x}_{11}\right)^2+\left(x_{211}-\overline{x}_{11}\right)^2+\left(x_{311}-\overline{x}_{11}\right)^2$;

$d_{12}=\left(x_{111}-\overline{x}_{11}\right)\cdot\left(x_{112}-\overline{x}_{12}\right)+\left(x_{211}-\overline{x}_{11}\right)\cdot\left(x_{212}-\overline{x}_{12}\right)+\left(x_{311}-\overline{x}_{11}\right)\cdot\left(x_{312}-\overline{x}_{12}\right)$;

$d_{21}=\left(x_{111}-\overline{x}_{11}\right)\cdot\left(x_{112}-\overline{x}_{12}\right)+\left(x_{211}-\overline{x}_{11}\right)\cdot\left(x_{212}-\overline{x}_{12}\right)+\left(x_{311}-\overline{x}_{11}\right)\cdot\left(x_{312}-\overline{x}_{12}\right)$;

$d_{22}=\left(x_{112}-\overline{x}_{12}\right)^2+\left(x_{212}-\overline{x}_{12}\right)^2+\left(x_{312}-\overline{x}_{12}\right)^2$.

$$X_{c2}{}^T\cdot X_{c2}=\begin{pmatrix} x_{121}-\overline{x_{21}} & x_{221}-\overline{x_{21}} \\ x_{122}-\overline{x_{22}} & x_{222}-\overline{x_{22}} \end{pmatrix}\cdot\begin{pmatrix} x_{121}-\overline{x_{21}} & x_{122}-\overline{x_{22}} \\ x_{221}-\overline{x_{21}} & x_{222}-\overline{x_{22}} \end{pmatrix}=\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$

Где $b_{11}=\left(x_{121}-\overline{x_{21}}\right)^2+\left(x_{221}-\overline{x_{21}}\right)^2$;

$b_{12}=\left(x_{121}-\overline{x_{21}}\right)\cdot\left(x_{122}-\overline{x_{22}}\right)+\left(x_{221}-\overline{x_{21}}\right)\cdot\left(x_{222}-\overline{x_{22}}\right)$;

$b_{12}=\left(x_{121}-\overline{x_{21}}\right)\cdot\left(x_{122}-\overline{x_{22}}\right)+\left(x_{221}-\overline{x_{21}}\right)\cdot\left(x_{222}-\overline{x_{22}}\right)$;

$b_{22}=\left(x_{122}-\overline{x_{22}}\right)^2+\left(x_{222}-\overline{x_{22}}\right)^2$.

Вновь полученные матрицы $X_{c1}{}^T\cdot X_{c1}$ и $X_{c2}{}^T\cdot X_{c2}$ характеризуют взаимосвязь между координатами в первом и втором множествах соответственно. Объединённая матрица, характеризующая взаимосвязи между координатами в первом и втором множествах соответственно может быть получена в результате сложения матриц.

Вычислим $X_{c1}{}^T\cdot X_{c1}+X_{c2}{}^T\cdot X_{c2}$. В результате получим:

$$X_{c1}{}^T\cdot X_{c1}+X_{c2}{}^T\cdot X_{c2}=\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix},$$

где $c_{11}=\left(x_{111}-\overline{x_{11}}\right)^2+\left(x_{211}-\overline{x_{11}}\right)^2+\left(x_{311}-\overline{x_{11}}\right)^2+\left(x_{121}-\overline{x_{21}}\right)^2+\left(x_{221}-\overline{x_{21}}\right)^2$;

$c_{12}=\left(x_{111}-\overline{x_{11}}\right)\cdot\left(x_{112}-\overline{x_{12}}\right)+\left(x_{211}-\overline{x_{11}}\right)\cdot\left(x_{212}-\overline{x_{12}}\right)+\left(x_{311}-\overline{x_{11}}\right)\cdot\left(x_{312}-\overline{x_{12}}\right)+$

$+\left(x_{121}-\overline{x_{21}}\right)\cdot\left(x_{122}-\overline{x_{22}}\right)+\left(x_{221}-\overline{x_{21}}\right)\cdot\left(x_{222}-\overline{x_{22}}\right)$;

$c_{21}=\left(x_{111}-\overline{x_{11}}\right)\cdot\left(x_{112}-\overline{x_{12}}\right)+\left(x_{211}-\overline{x_{11}}\right)\cdot\left(x_{212}-\overline{x_{12}}\right)+\left(x_{311}-\overline{x_{11}}\right)\cdot\left(x_{312}-\overline{x_{12}}\right)+$

$+\left(x_{121}-\overline{x_{21}}\right)\cdot\left(x_{122}-\overline{x_{22}}\right)+\left(x_{221}-\overline{x_{21}}\right)\cdot\left(x_{222}-\overline{x_{22}}\right)$;

$$c_{22} = \left(x_{112} - \overline{x_{12}}\right)^2 + \left(x_{212} - \overline{x_{12}}\right)^2 + \left(x_{312} - \overline{x_{12}}\right)^2 + \left(x_{122} - \overline{x_{22}}\right)^2 + \left(x_{222} - \overline{x_{22}}\right)^2;$$

Строгая оценка несмещённой матрицы, характеризующая взаимосвязи между признаками в первом и втором множествах имеет вид: $\tilde{S} = \frac{1}{n_1 + n_2 - 2}\left(X_{c1}^T \cdot X_{c1} + X_{c2}^T \cdot X_{c2}\right)$ или $\tilde{S} = \frac{1}{n_1 + n_2 - 2}\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$. Следовательно, $\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = (n_1 + n_2 - 2)\tilde{S}$. Полученные формулы можно представить в виде несмещённой оценки обобщённой матрицы ковариаций $\hat{S} = \frac{1}{n_1 + n_2 - 2}\left(n_1 S_1 + n_2 S_2\right)$, где S_1 и S_2 – матрицы ковариаций первой и второй выборок соответственно.

Введём вектор коэффициентов дискриминантной функции $A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$, транспонированный вектор значений коэффициентов $A^T = (a_1 \quad a_2)$. Матрицу $\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$ умножим на вектор A и A^T. Учитывая правила умножения матриц, получим $A^T \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} A$. Тогда получим выражение: $A^T (n_1 + n_2 - 2)\hat{S}A$.

Таким образом, оценку вариации дискриминантной функции внутри множеств можно представить в виде:

$$\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left(f_{ki} - \overline{f_k}\right)^2 = A^T\left[(n_1 + n_2 - 2)\tilde{S}\right]A.$$

Вариация между множествами может быть оценена как:

$$\left(\overline{f_1} - \overline{f_2}\right)^2 = \left[\left(a_1\overline{x_{11}} + a_2\overline{x_{12}}\right) - \left(a_1\overline{x_{21}} + a_2\overline{x_{22}}\right)\right]^2 = \left[a_1\left(\overline{x_{11}} - \overline{x_{21}}\right) + a_2\left(\overline{x_{12}} - \overline{x_{22}}\right)\right]^2,$$

$$\left(\overline{f_1} - \overline{f_2}\right)^2 = a_1^2\left(\overline{x_{11}} - \overline{x_{21}}\right)^2 + 2a_1a_2\left(\overline{x_{11}} - \overline{x_{21}}\right)\left(\overline{x_{12}} - \overline{x_{22}}\right) + a_2^2\left(\overline{x_{12}} - \overline{x_{22}}\right)^2.$$

Введём векторы средних значений признаков в каждом множестве: $\overline{X_1} = \begin{pmatrix} \overline{x_{11}} \\ \overline{x_{12}} \end{pmatrix}$ и $\overline{X_2} = \begin{pmatrix} \overline{x_{21}} \\ \overline{x_{22}} \end{pmatrix}$. Вычислим разность векторов $\left(\overline{X_1} - \overline{X_2}\right) = \begin{pmatrix} \overline{x_{11}} - \overline{x_{21}} \\ \overline{x_{12}} - \overline{x_{22}} \end{pmatrix}$, транспонируем $\left(\overline{X_1} - \overline{X_2}\right)^T = \left(\overline{x_{11}} - \overline{x_{21}} \quad \overline{x_{12}} - \overline{x_{22}}\right)$. Вычислим $\left(\overline{X_1} - \overline{X_2}\right)\left(\overline{X_1} - \overline{X_2}\right)^T$. В результате получим $\begin{pmatrix} \left(\overline{x_{11}} - \overline{x_{21}}\right)^2 & \left(\overline{x_{11}} - \overline{x_{21}}\right)\left(\overline{x_{12}} - \overline{x_{22}}\right) \\ \left(\overline{x_{11}} - \overline{x_{21}}\right)\left(\overline{x_{12}} - \overline{x_{22}}\right) & \left(\overline{x_{12}} - \overline{x_{22}}\right)^2 \end{pmatrix}$. Умножим $\left(\overline{X_1} - \overline{X_2}\right)\left(\overline{X_1} - \overline{X_2}\right)^T$ на вектор A и A^T. Учитывая правила умножения матриц, получим $\left(\overline{f_1} - \overline{f_2}\right)^2 = A^T\left(\overline{X_1} - \overline{X_2}\right)\left(\overline{X_1} - \overline{X_2}\right)^T A$, описывающее межгрупповую вариацию.

При нахождении коэффициентов дискриминантной функции a_1 и a_2 необходимо учесть, что для рассматриваемых объектов внутригрупповая

вариация должна быть минимальной, а межгрупповая вариация должна быть максимальной. Тогда наилучшее разделение двух множеств возможно с учетом этих двух условий. Составим функцию F, которая должна быть максимальной:

$$F = \frac{A^T\left(\overline{X_1} - \overline{X_2}\right)\left(\overline{X_1} - \overline{X_2}\right)^T A}{A^T\left[(n_1 + n_2 - 2)\hat{S}\right]A} \to \max.$$

Решением данной задачи является вектор

$$A = \hat{S}^{-1}\left(\overline{X_1} - \overline{X_2}\right),$$

где $\hat{S}^{-1}$ – обратная матрица к обобщённой матрице ковариаций.

Таким образом, вычислив вектор коэффициентов дискриминантной функции, приступают к процедуре дискриминации. Исходные массивы данных по каждой выборке умножаются на вектор A: $U_1=X_1A$, $U_2=X_2A$. Полученные значения усредняются по каждой выборке $\overline{U_1}$ и $\overline{U_2}$. Используя средние значения $\overline{U_1}$ и $\overline{U_2}$, вычисляется константа дискриминации C: $C = \frac{\overline{U_1} + \overline{U_2}}{2}$.

Данная величина представляет собой границу, которая равноудалена от центров двух множеств (рис. 2). Из рис. 1 видно, что дискриминируемые объекты, расположенные выше прямой C, находятся ближе к центру множества M_1 и, следовательно, могут быть отнесены к множеству M_1, а объекты, расположенные ниже прямой C, находятся ближе к центру множества M_2 и, следовательно, могут быть отнесены к множеству M_2.

Алгоритм дискриминантного анализа:

1. Вычислить средние значения признаков для каждого множества (обучающей выборки), записать векторы средних значений $\overline{X_1}$ и $\overline{X_2}$. Вычислить вектор разности $\left(\overline{X_1} - \overline{X_2}\right)$.
2. Вычислить матрицы ковариаций для каждой выборки S_1 и S_2.
3. Вычислить несмещённую оценку обобщённой матрицы ковариаций $\hat{S} = \frac{1}{n_1 + n_2 - 2}\left(n_1 S_1 + n_2 S_2\right)$.
4. Вычислить $\hat{S}^{-1}$.
5. Вычислить вектор коэффициентов дискриминантной функции A.
6. Вычислить константу дискриминации C.
7. Сравнить значение дискриминантной функции тестируемых объектов с величиной C.

Рассмотрим примеры использования дискриминантного анализа для классификации объектов.

Задача 1.

В таблице представлены группы регионов с высоким и низким уровнями безработицы среди мужчин и женщин. Характеризуя регионы долей безра-

ботных среди женщин (X_1) и мужчин (X_2), с помощью дискриминантного анализа требуется классифицировать три последних региона.

№ региона	Показатель / Группа регионов	Безработица среди женщин, % (X_1)	Безработица среди мужчин,% (X_2)
1	Высокий уровень	23,4	9,1
2		19,1	6,6
3		17,5	5,2
4		17,2	10,1
5	Низкий уровень	5,4	4,3
6		6,6	5,5
7		8	5,7
8		9,7	5,5
9		9,1	6,6
10	Подлежат дискриминации	9,9	7,4
11		14,2	9,4
12		12,9	6,7

1. Средние значения признаков для каждого множества, вектор разности $(\overline{X_1} - \overline{X_2})$.

Высокий уровень	Низкий уровень	Разность
$\overline{X_1}$	$\overline{X_2}$	$(\overline{X_1} - \overline{X_2})$
19,3	7,76	11,54
7,75	5,52	2,23

2. Матрицы ковариаций для обеих групп предприятий:

S_1	X_1	X_2
X_1	6,125	1,355
X_2	1,355	3,7925

S_2	X_1	X_2
X_1	2,5064	0,8708
X_2	0,8708	0,5376

3. Несмещенная оценка обобщённой матрицы ковариаций $\overline{S}$:

5,290286	1,396286
1,396286	2,551143

4. $\hat{S}^{-1}$

0,220942	-0,12093
-0,12093	0,458166

5. Вектор оценок коэффициентов дискриминантной функции $A = \hat{S}^{-1}(\overline{X}_1 - \overline{X}_2)$:

A
2,280007
-0,37377

6. Рассчитать оценки векторов значений дискриминантной функции для матриц исходных данных X_1 и X_2

№	U_1	№	U_2
1	49,95086	1	10,70483
2	41,08126	2	12,99231
3	37,95652	3	16,10957
4	35,44105	4	20,06033
		5	18,28118
Среднее значение	41,10742		15,62964

7. Константа дискриминации C=28,36853

8. Значение дискриминантной функции для предприятий группы Z:

№ предприятия	Z		u_z	Группа
	X_1	X_2		
10	9,9	7,4	19,80617	Низкий уровень, Y
11	14,2	9,4	28,86266	Высокий уровень, X
12	12,9	6,7	26,90783	Низкий уровень, Y

Процедура дискриминантного анализа закончена. В результате установлено, что два из трёх регионов попадают в множество регионов низкого уровня безработицы, так как величина дискриминантной функции этих регионов меньше, чем полученное значение константы дискриминации C, а один регион попадает в множество высокого уровня безработицы, так как величина дискриминантной функции этого региона больше, чем значение константы дискриминации C.

ГЛАВА 4
КЛАСТЕРНЫЙ АНАЛИЗ

Необходимость анализа и формализации задач, связанных со сравнением и классификацией объектов, сознавали учёные далекого прошлого. «Его (Аристотеля) величайшим и в то же время чреватым наиболее опасными последствиями вкладом в науку была идея классификации, которая проходит через все его работы. Аристотель ввел или, по крайней мере, кодифицировал способ классификации предметов, основанный на сходстве и различии…», – писал Дж. Бернал в «Науке истории общества». После Аристотеля ещё в докомпьютерной эре имеется ряд интересных примеров, прекрасно построенных классификаций, как в естественных, так и в общественных науках. Две из них широко известны: иерархическая классификация (основанная на понятии сходства) растений и видов М. Адансона (1757 г.) и знаменитая периодическая система элементов Д.И. Менделеева (1869 г.).

До разработки математического аппарата проблемы теории и практики классификации относились не к разработке методов и алгоритмов, а к полноте и тщательности отбора и теоретического анализа изучаемых объектов, характеризующих их признаков, смысла и числа градаций по каждому из них. Все методы классификации сводились к комбинационной группировке, при которой два объекта относятся к одной группе только при точном совпадении зарегистрированных на них градаций одновременно по всем характеризующим их признакам. Однако все данные о социально-экономических явлениях носят многомерный и разнотипный характер, и до их анализа обычно бывает неясно, насколько существенно то или иное свойство для конкретной цели. В этих условиях на первый план выходят проблемы построения группировок и классификаций.

ОБЩАЯ ХАРАКТЕРИСТИКА МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА

Под классификацией понимается разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы либо отнесение каждого из заданной группы объектов к одному из заранее известных классов. Кластерный анализ – совокупность методов, позволяющих классифицировать наблюдения, каждое из которых описывается набором исходных переменных $X_1, X_2, X_3, \ldots, X_m$.

Целью кластерного анализа является образование групп, схожих между собой объектов, которые принято называть кластерами. Слово *кластер* английского происхождения (*cluster*), переводится как сгусток, пучок, группа. Родственные понятия, используемые в литературе, – класс, таксон, сгущение.

В кластерном анализе используется политетический подход при образовании групп. Все группировочные признаки одновременно участвуют в группировке (классификации), т.е. они учитываются все сразу при отнесении наблюдения в ту или иную группу. При этом, как правило, не указаны четкие границы каждой группы, а также неизвестно заранее, сколько же групп целесообразно выделить в исследуемой совокупности.

Особо важное место кластерный анализ занимает в тех отраслях науки, которые связаны с изучением массовых явлений и процессов. Необходимость развития методов кластерного анализа и их использования продиктована, прежде всего, тем, что они помогают построить научно обоснованные классификации, выявить внутренние связи между единицами наблюдаемой совокупности. Методы кластерного анализа могут использоваться с целью сжатия информации, что является важным фактором в условиях постоянного увеличения и усложнения потоков статистических данных.

Первые публикации по кластерному анализу появились в конце 30-х годов прошлого столетия, но активное развитие этих методов и их широкое использование началось в конце 60-х – начале 70-х годов. Первоначально методы кластерного анализа использовались в психологии, археологии, биологии. В настоящее время они стали активно применяться в социологии, политологии, экономике.

Методы кластерного анализа позволяют решать следующие задачи:

- Проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение такой задачи приводит к углублению знаний о совокупности классифицируемых объектов.
- Проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов.
- Построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться внести в нее структуру.

Методы кластерного анализа делятся на две большие группы:

1) агломеративные (объединяющие);
2) дивизимные (разделяющие)

Агломеративные методы последовательно объединяют отдельные объекты в группы (кластеры), а дивизимные методы расчленяют группы на отдельные объекты. В свою очередь каждый метод как объединяющего, так и разделяющего типа может быть реализован при помощи различных алгоритмов.

МЕРЫ СХОДСТВА

Для проведения классификации вводится понятие сходства объектов по наблюдаемым переменным. В каждый кластер должны попасть объекты, имеющие сходные характеристики.

В кластерном анализе для количественной оценки сходства вводится понятие метрики. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается m признаками, то он может быть представлен как точка в m-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние. В кластерном анализе используются различные меры расстояния между объектами:

1. Евклидово расстояние:

$$d_{ij} = \sqrt{\sum_{k=1}^{m}\left(x_{ik} - x_{jk}\right)^2} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{im} - x_{jm}\right)^2},$$

где x_{ik} – значение k-ого признака для i-го объекта, x_{jk} – значение k-го признака для j-ого объекта. Например, пусть нам даны три объекта n_1, n_2, n_3, каждый из которых описывается четырьмя признаками X_1, X_2, X_3, X_4.

	X_1	X_2	X_3	X_4
n_1	x_{11}	x_{12}	x_{13}	x_{14}
n_2	x_{21}	x_{22}	x_{23}	x_{24}
n_3	x_{31}	x_{32}	x_{33}	x_{34}

Расстояния между парами объектов определяются как:

$$d_{12} = \sqrt{\left(x_{11} - x_{21}\right)^2 + \left(x_{12} - x_{22}\right)^2 + \left(x_{13} - x_{23}\right)^2 + \left(x_{14} - x_{24}\right)^2},$$
$$d_{13} = \sqrt{\left(x_{11} - x_{31}\right)^2 + \left(x_{12} - x_{32}\right)^2 + \left(x_{13} - x_{33}\right)^2 + \left(x_{14} - x_{34}\right)^2},$$
$$d_{23} = \sqrt{\left(x_{21} - x_{31}\right)^2 + \left(x_{22} - x_{32}\right)^2 + \left(x_{23} - x_{33}\right)^2 + \left(x_{24} - x_{34}\right)^2},$$

где d_{12} – евклидово расстояние между первым и вторым объектами, d_{13} и d_{23} – между первым и третьим и вторым и третьим соответственно.

2. Взвешенное евклидово расстояние:

$$d_{ij} = \sqrt{\sum_{k=1}^{m}\omega_k\left(x_{ik} - x_{jk}\right)^2} = \sqrt{\omega_1\left(x_{i1} - x_{j1}\right)^2 + \omega_2\left(x_{i2} - x_{j2}\right)^2 + \cdots + \omega_m\left(x_{im} - x_{jm}\right)^2},$$

где ω_1 – вес признака X_1, ω_2 – вес признака X_2, ω_3 – вес признака X_3, …, ω_m – вес признака X_m. Вопрос о придании переменным соответствующих весов должен решаться после проведения исследователем анализа изучаемой совокупности и социальной сущности классифицирующих переменных. Вес

задается пропорционально степени важности элементов. Значение ω_k устанавливается исследователем самостоятельно, таким образом, что $\sum_{k=1}^{m} \omega_k = 1$.

3. Расстояние *city-block* $d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$.

4. Расстояние Махаланобиса

$d_{ij} = \left(\overline{X_i} - \overline{X_j}\right)' S^{-1} \left(\overline{X_i} - \overline{X_j}\right)$, где $\overline{X_i}$ и $\overline{X_j}$ – векторы средних значений, S – матрица ковариаций

Оценка сходства между объектами сильно зависит от абсолютного значения признака и от степени его вариации в совокупности. Чтобы устранить подобное влияние на процедуру классификации, значения переменных нормируют одним из следующих способов:

1) $z_{ij} = \frac{x_{ij} - \overline{x}}{S_j}$, 2) $z_{ij} = \frac{x_{ij}}{x_{\max j}}$, 3) $z_{ij} = \frac{x_{ij}}{\overline{x_j}}$, 4) $z_{ij} = \frac{x_{ij}}{x_{\min j}}$.

Иногда в качестве меры сходства используются парные коэффициенты корреляции, коэффициент ранговой корреляции. Если исходные переменные являются альтернативными признаками, т.е. принимают значения 0 и 1, то в качестве меры сходства используются меры ассоциативности.

Используя любую из перечисленных мер сходства, от таблицы исходных данных необходимо перейти к матрице, содержащей меры сходства, т.е. расстояния. В общем виде такая матрица имеет вид:

	n_1	n_2	n_3	...	n_n
n_1	0	d_{12}	d_{13}	...	d_{1m}
n_2	d_{21}	0	d_{23}	...	d_{2m}
n_3	d_{31}	d_{32}	0	...	d_{3m}
...	...	...	...	...	...
n_n	d_{n1}	d_{n2}	d_{n3}		0

На пересечении i-ой строки и j-го столбца матрицы находится расстояние от i-го объекта до j-го объекта. На главной диагонали матрицы расположены нули. Матрица симметрична относительно главной диагонали, так как $d_{ij} = d_{ji}$.

ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ

Из всех методов кластерного анализа самыми распространенными являются иерархические агломеративные методы. Сущность этих методов заключается в том, что на первом шаге каждый объект выборки рассматривается как отдельный кластер. Процесс объединения кластеров происходит последовательно:

1) в таблице, содержащей расстояния, находится минимальное число d_{ij}, это означает, что на данном расстоянии объединяются в один кластер i

и j объекты; таблица расстояний пересчитывается с учетом вновь образовавшегося кластера;

2) во вновь полученной матрице находится минимальное расстояние – в результате возможно:
   а) два других объекта объединятся в новый кластер;
   б) третий объект будет присоединен к первому кластеру;
3) два предыдущих пункта повторяются.

Пересчет таблиц расстояний зависит от метода кластеризации. Используются четыре основных метода: метод «ближнего соседа», метод «дальнего соседа», метод «средней связи», центроидный метод.

В методе «ближнего соседа» после объединения i-го и j-го объектов в кластер новое расстояние $d(k;S(i,j))$ от k-го объекта до кластера, содержащего i-й и j-й объекты, выбирается как минимальное расстояние из двух расстояний от k-го объекта до i-го объекта $d(k;i)$ и от k-го объекта до j-го объекта $d(k;j)$, т.е. $d(k;S(i,j))=min\{d(k;i);d(k;j)\}$.

В методе «дальнего соседа» после объединения i-го и j-го объектов в качестве расстояния от k-го объекта до кластера, состоящего из i-го и j-го объектов $d(k;S(i,j))$, выбирается максимальное расстояние из двух расстояний от k-го объекта до i-го объекта $d(k;i)$ и от k-го объекта до j-го объекта $d(k;j)$, т.е. $d(k;S(i,j))=max\{d(k;i);d(k;j)\}$.

В методе «средней связи» расстояние от k-го объекта до кластера, состоящего из i-го и j-го объектов $d(k;S(i,j))$, рассчитывается как среднее арифметическое двух расстояний $d(k;i)$ и $d(k;j)$, т.е. $d(k;S(i,j))=\{d(k;i)+d(k;j)\}/2$.

Центроидный метод предполагает пересчет тех значений матрицы расстояний, которые связаны с новым кластером. Кластеру $S(i,j)$ присваиваются новые значения признаков X_1, X_2, X_3 , X_4, которые рассчитываются как средние арифметические $(X_{i1}+ X_{j1})/2$. Для нашего примера, в котором три объекта и четыре признака, например, после объединения в кластер $S(2,3)$ объектов n_2 и n_3, исходная матрица значений принимает вид:

	X_1	X_2	X_3	X_4
n_1	x_{11}	x_{12}	x_{13}	x_{14}
$S(2,3)$	$(x_{21}+ x_{31})/2$	$(x_{22}+ x_{32})/2$	$(x_{23} + x_{33})/2$	$(x_{24} +x_{34})/2$

По вновь полученной таблице пересчитывается расстояние между объектом n_1 и кластером $S(2,3)$. Далее повторяются операции пунктов 1) – 3), т.е. находится минимальное расстояние, на котором новый объект или добавляется в кластер, или образует новый кластер.

Рассмотрим процедуру классификации на *примере*.

Потребительское поведение 5 семей характеризуется удельными (на душу) расходами за летние месяцы на культуру, спорт, отдых (признак X_1 – тыс. руб.) и питание (признак X_2 – тыс.руб.).

Значения показателей представлены в таблице.

№ семьи	1	2	3	4	5
X_1	2	4	8	12	13
X_2	10	7	6	11	9

Используя евклидову метрику, были рассчитаны расстояния между объектами (семьями). Например, расстояние между 1 и 2 объектами

$$d_{12} = \sqrt{(2-4)^2 + (10-7)^2} = 3{,}61.$$

Матрица расстояний имеет вид:

	n_1	n_2	n_3	n_4	n_5
n_1	0	3,61	7,21	10,05	11,05
n_2	3,61	0	4,12	8,94	9,22
n_3	7,21	4,12	0	6,4	5,83
n_4	10,05	8,94	6,4	0	**2,24**
n_5	11,05	9,22	5,83	2,24	0

Из матрицы видно, что минимальное расстояние 2,24 – это расстояние между объектами n_4 и n_5. Следовательно, эти объекты образуют первый кластер *S(4,5)*. Далее необходимо пересчитать расстояния от объектов n_1, n_2 и n_3 до первого кластера *S(4,5)*. В методе «ближнего соседа» *d(1;S(4,5))=min{10,05; 11,05}=10,05*. В методе «дальнего соседа» *d(1;S(4,5))=max{10,05; 11,05}=11,05*. В методе средней связи *d(1;S(4,5))=(10,05+11,05)/2=10,55*.

	Методы		
	«ближнего соседа»	«дальнего соседа»	средняя связь
$d_{1,S(4,5)}$	*min*{10,05;11,05}=10,05	*max*{10,05;11,05}=11,05	(10,05+11,05)/2=10,55
$d_{2,S(4,5)}$	*min*{8,94; 9,22}=8,94	*max* {8,94; 9,22}=9,22	(8,94+ 9,22)=9,08
$d_{3,S(4,5)}$	*min* {6,45, 5,83}=5,83	*max* {6,45, 5,83}=6,4	(6,45+5,83)=6,12

Таким образом, матрица расстояний для метода «ближнего соседа» принимает вид:

	n_1	n_2	n_3	*S (4, 5)*
n_1	0	**3,61**	7,21	10,05
n_2	3,61	0	4,12	8,94
n_3	7,21	4,12	0	5,83
S (4, 5)	10,05	8,94	5,83	0

Из неё видно, что минимальное расстояние 3,61 – это расстояние между объектами n_1 и n_2. Следовательно, эти объекты образуют второй кластер *S(1,2)*. Пересчитаем расстояния от объекта n_3 до кластера *S(1,2)* и от кластера *S(4,5)* до кластера *S(1,2)*: *d(3;S(1,2))=min{7,21; 4,12}=4,12*; *d(S(4,5);S(1,2))=min{8,94;10,05}=8,94*.

Матрица расстояний для метода «ближнего соседа» после пересчета принимает вид:

	S (1, 2)	n_3	*S (4, 5)*
S (1, 2)	0	**4,12**	8,94
n_3	4,12	0	5,83
S (4, 5)	8,94	5,83	0

На минимальном расстоянии 4,12 объект n_3 присоединяется к кластеру *S(1,2)*, в результате образуется кластер *S(1,2,3)*. Вновь пересчитываем расстояние между кластерами *S(1,2,3)* и *S(4,5)*: *d(S(1,2,3);S(4,5))=min{8,94; 5,83}=5,83*.

Окончательно, таблица расстояний имеет вид:

	S (1, 2, 3)	*S (4, 5)*
S (1, 2,3)	0	**5,83**
S (4, 5)	5,83	0

Объединение кластеров *S(1,2,3)* и *S(4,5)* возможно на расстоянии 5,83. На этом процедура классификации по методу «ближнего соседа» заканчивается.

Графические результаты процедуры классификации изображаются в виде дендрограммы. По оси абсцисс откладываются объекты (семьи), по оси ординат – расстояния, на которых происходило объединение. Для метода «ближнего соседа» дендрограмма имеет вид (рис. 3):

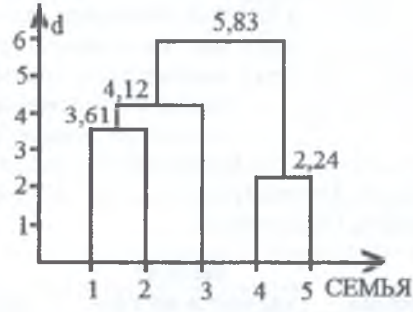

Рис. 3. Дендрограмма (метод «ближнего соседа»)

Продолжим процедуру классификации по методу «дальнего соседа».

	n_1	n_2	n_3	*S (4, 5)*
n_1	0	**3,61**	7,21	11,05
n_2	3,61	0	4,12	9,22
n_3	7,21	4,12	0	6,4
S (4, 5)	11,05	9,22	6,4	0

Из матрицы видно, что минимальное расстояние 3,61 – это расстояние между объектами n_1 и n_2. Следовательно, эти объекты образуют второй кластер *S(1,2)*. Пересчитаем расстояния от объекта n_3 до кластера *S(1,2)* и от кластера *S(4,5)* до кластера *S(1,2)*:

d(3;S(1,2))=max{7,21; 4,12}=7,21; *d(S(4,5);S(1,2))=max{9,22;11,05}=11,05.*

Матрица расстояний для метода «дальнего соседа» после пересчета принимает вид:

	S (1, 2)	n_3	*S (4, 5)*
S (1, 2)	0	7,21	11,05
n_3	7,21	0	**6,4**
S (4, 5)	11,05	6,4	0

Видно, что минимальное расстояние 6,4 – это расстояние между объектами n_3 и кластером *S(4,5)*. Следовательно, объект n_3 присоединяется к кластеру *S(4,5)*, в результате образуется кластер *S(3,4,5)*. Вновь пересчитываем расстояние между кластерами *S(1,2)* и *S(3,4,5)*: *d(S(1,2);S(3,4,5))=max{7,21; 11,05}=11,05.*

Окончательно таблица расстояний имеет вид:

	S (1, 2)	*S (3,4, 5)*
S (1, 2)	0	**11,05**
S (3,4, 5)	11,05	0

Объединение кластеров *S(1,2)* и *S(3,4,5)* возможно на расстоянии 11,05. На этом процедура классификации по методу «дальнего соседа» заканчивается. Для метода «дальнего соседа» дендрограмма имеет вид (рис. 4):

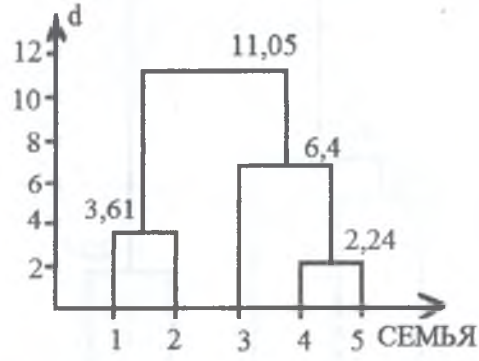

Рис. 4. Дендрограмма (метод «дальнего соседа»)

Проведём процедуру классификации, используя метод «средней связи».

	n_1	n_2	n_3	*S (4, 5)*
n_1	0	**3,61**	7,21	10,55
n_2	3,61	0	4,12	9,08
n_3	7,21	4,12	0	6,12
S (4, 5)	10,55	9,08	6,12	0

Из матрицы видно, что минимальное расстояние 3,61 – это расстояние между объектами n_1 и n_2. Аналогично методу «ближнего соседа» эти объекты образуют второй кластер *S(1,2)*. Пересчитаем расстояния от объекта n_3 до кластера *S(1,2)* и от кластера *S(4,5)* до кластера *S(1,2)*:
d(3;S(1,2))=(7,21+ 4,12)/2=5,67; *d(S(4,5);S(1,2))=(10,55+9,08)/2=9,82.*

Матрица расстояний для метода «средней связи» после пересчета принимает вид:

	S (1, 2)	n_3	*S (4, 5)*
S (1, 2)	0	**5,67**	9,82
n_3	5,67	0	6,12
S (4, 5)	9,82	6,12	0

Видно, что минимальное расстояние 5,67– это расстояние между объектами n_3 и кластером *S(1,2)*. Следовательно, объект n_3 присоединяется к кластеру *S(1,2)*, в результате образуется кластер *S(1,2,3)*. Вновь пересчитываем расстояние между кластерами *S(1,2,3)* и *S(4,5)*: *d(S(1,2,3); S(4,5)=(9,82+6,12)/2=7,97.*

Окончательно, матрица расстояний имеет вид:

	S (1, 2, 3)	*S (4, 5)*
S (1, 2,3)	0	**7,97**
S (4,5)	7,97	0

Из неё видно, что объединение кластеров *S(1,2,3)* и *S(4,5)* возможно на расстоянии 7,97. На этом процедура классификации по методу «средней связи» заканчивается.

Для метода «средней связи» дендрограмма имеет вид (рис. 5):

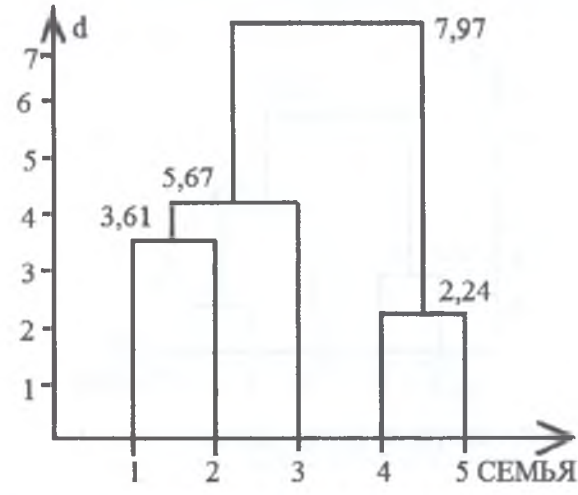

Рис. 5. Дендрограмма (метод «средней связи»)

Рассмотрим центроидный метод. Начальный этап классификации совпадает с рассмотренными выше методами. Так как, минимальное расстояние в таблице расстояний 2,24 – это расстояние между объектами n_4 и n_5. Эти объекты образуют первый кластер *S(4,5)*. Чтобы пересчитать расстояния, необходимо вычислить координаты центра тяжести образовавшегося кластера. Для этого необходимо вычислить среднее значение по каждому признаку: $X_{1ц}$=(12+13)/2=12,5; $X_{2ц}$=(11+9)/2=10. Кластер *S(4,5)* характеризуется в дальнейшем его центром тяжести.

Таблица первоначальных данных принимает вид:

№ семьи	1	2	3	*S(4,5)*
X_1	2	4	8	12,5
X_2	10	7	6	10

Далее необходимо пересчитать расстояния от кластера *S(4,5)* до объектов n_1, n_2 и n_3. В частности,

$$d_{1,S(4,5)} = \sqrt{(12{,}5-2)^2+(10-10)^2} = 10{,}5;$$

$$d_{2,S(4,5)} = \sqrt{(12{,}5-4)^2+(10-7)^2} = 9{,}01;\ d_{3,S(4,5)} = \sqrt{(12{,}5-8)^2+(10-6)^2} = 6{,}02.$$

	n_1	n_2	n_3	$S(4, 5)$
n_1	0	**3,61**	7,21	10,5
n_2	3,61	0	4,12	9,01
n_3	7,21	4,12	0	6,02
$S(4, 5)$	10,5	9,01	6,02	0

Из матрицы расстояний видно, что минимальное расстояние 3,61 – это расстояние между объектами n_1 и n_2. Следовательно, эти объекты образуют второй кластер $S(1,2)$. Вычисляем координаты центра тяжести образовавшегося кластера: $X_{1ц}$=(2+4)/2=3; $X_{2ц}$=(10+7)/2=8,5. Кластер $S(1,2)$ характеризуется в дальнейшем его центром тяжести (3; 8,5).

Таблица первоначальных данных принимает вид:

№ семьи	$S(1,2)$	3	$S(4,5)$
X_1	3	8	12,5
X_2	8,5	6	10

Пересчитываем расстояния от кластера $S(1,2)$ до объекта n_3 и кластера $S(4,5)$, используя евклидову метрику:

$$d_{3,S(1,2)} = \sqrt{(3-8)^2 + (8{,}5-6)^2} = 5{,}59;$$

$$d_{S(4,5),S(1,2)} = \sqrt{(3-12{,}5)^2 + (8{,}5-10)^2} = 9{,}62.$$

Матрица расстояний имеет вид:

	$S(1, 2)$	n_3	$S(4, 5)$
$S(1, 2)$	0	**5,59**	9,62
n_3	5,59	0	6,02
$S(4, 5)$	9,62	6,02	0

Видно, что минимальное расстояние 5,59 – это расстояние между объектами n_3 и кластером $S(1,2)$. Следовательно, объект n_3 присоединяется к кластеру $S(1,2)$, в результате образуется кластер $S(1,2,3)$. Пересчитываем координаты центра тяжести нового кластера $S(1,2,3)$: $X_{1ц}$=(2+4+8)/3=4,67; $X_{2ц}$=(10+7+6)/3=7,67. Кластер $S(1,2,3)$ характеризуется в дальнейшем его центром тяжести (4,67;7,67).

Таблица первоначальных данных принимает вид:

№ семьи	$S(1,2,3)$	$S(4,5)$
X_1	4,67	12,5
X_2	7,67	10

Расстояние между кластерами $S(1,2,3)$ и $S(4,5)$

$$d_{S(4,5),S(1,2,3)} = \sqrt{(4{,}67-12{,}5)^2 + (7{,}67-10)^2} = 8{,}17$$

Окончательно таблица расстояний имеет вид:

	$S(1, 2, 3)$	$S(4, 5)$
$S(1, 2,3)$	0	**8,17**
$S(4,5)$	8,17	0

Из таблицы видно, что объединение кластеров $S(1,2,3)$ и $S(4,5)$ возможно на расстоянии 8,17. На этом процедура классификации по центроидному методу заканчивается. Для центроидного метода дендрограмма имеет вид (рис. 6):

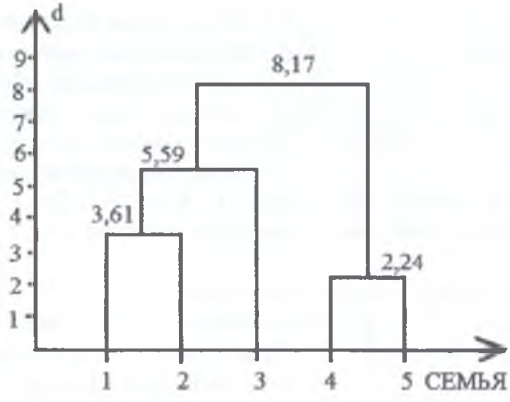

Рис. 6. Дендрограмма (центроидный метод)

Таким образом, сравнивая результаты 4-х разбиений пяти семей на однородные группы, можно отметить, что наиболее устойчивым является разбиение на два кластера *S(1,2,3)* и *S(4,5)*. Только в одном случае из четырёх при использовании метода «дальнего соседа» получено разбиение *S(1,2)* и *S(3,4,5)*. В общем случае, если в результате классификации различными методами получаются различные разбиения на однородные группы, используют строгие математические критерии для выбора окончательного разбиения. К таким критериям относятся критерии качества классификации. Рассмотрим данные критерии.

КРИТЕРИИ КАЧЕСТВА КЛАССИФИКАЦИИ

При использовании различных методов кластеризации для одной и той же совокупности могут быть получены различные варианты разбиения. Существенное влияние на характеристики кластерной структуры оказывают набор признаков, по которым осуществляется классификация, тип выбранного алгоритма и выбор меры сходства.

После завершения процедур классификации необходимо оценить полученные результаты. Для этой цели используется мера качества классификации, которую называют функционалом или критерием качества классификации. Наилучшим по выбранному функционалу следует считать такое разбиение, при котором достигается *минимальное* значение функционала качества. Рассмотрим три наиболее распространенных функционала качества классификации (разбиения).

Первый функционал или критерий определяется суммой квадратов расстояний от каждого объекта кластера до его центра. В результате суммируются результирующие квадраты расстояний по всем сформированным кластерам:

$$F_1 = \sum_{l=1}^{k} \sum_{i=1}^{p} d^2\left(x_i; \overline{x_l}\right),$$

где l – номер кластера; $\overline{x_l}$ – центр тяжести l-го кластера; $d^2\left(x_i; \overline{x_l}\right)$ – расстояние от i-го объекта l-го кластера до центра тяжести кластера l; p – количество объектов в кластере l. Величина критерия F_1 должна быть минимальной.

Второй функционал определяется суммой квадратов внутри кластерных расстояний $F_2 = \sum_{l=1}^{k} \sum_{i,j \in S_l} d_{ij}^2$. В этом случае наилучшим следует считать такое разделение, при котором F_2 также минимально, т.е. получены кластеры большой плотности, и объекты, попавшие в один кластер, близки между собой по значениям тех переменных, которые использовались для классификации.

Третий функционал определяется суммарной внутриклассовой вариацией признаков, т.е. предполагает вычисление суммы квадратов отклонений значений признаков от их средних значений для всех объектов, входящих в кластер, а также по всем кластерам вместе. Наилучшим считается разбиение, при котором F_3 также минимально. Таким образом, третий функционал представляет собой суммарную внутриклассовую дисперсию:

$$F_3 = \sum_{l=1}^{k} \sum_{i \in S_l} \sigma_{ij}^2 .$$

Численные значения функционалов можно представить в сводной таблице, которая позволяет принять окончательное решение о выборе оптимального разбиения на кластеры.

Методы		«ближнего соседа»	«дальнего соседа»	«средней связи»	центроидный
Функционалы	F_1	$F_{1Б}$	$F_{1Д}$	$F_{1С}$	$F_{1Ц}$
	F_2	$F_{2Б}$	$F_{2Д}$	$F_{2С}$	$F_{2Ц}$
	F_3	$F_{3Б}$	$F_{3Д}$	$F_{3С}$	$F_{3Ц}$

Проведём расчет критериев качества классификации для рассматриваемого примера с пятью семьями. Рассчитаем значения F_1, F_2 и F_3 для разбиений на кластеры *S(1,2,3)* и *S(4,5)*.

Чтобы вычислить критерий F_1, необходимо создать две таблицы исходных данных, соответствующих кластерам *S(1,2,3)* и *S(4,5)*.

№ семьи	X_1	X_2
1	2	10
2	4	7
3	8	6
$\overline{X}$(1,2,3)	4,7	7,7

№ семьи	X_1	X_2
4	12	11
5	13	9
$\overline{X}$(4,5)	12,5	10

Вычисляем координаты центра тяжести каждого кластера (аналогично центроидному методу). Для кластера *S(1,2,3)* центр тяжести $\overline{X}(1,2,3)=(4,7;7,7)$. Для кластера *S(4,5)* центр тяжести $\overline{X}(4,5)=(12,5;\ 10)$. Вычислим квадраты расстояний от объектов n_1, n_2 и n_3 до центра тяжести кластера *S(1,2,3)*:

$d_{1,\overline{X}(1,2,3)}=(2-4,7)^2+(10-7,7)^2=1258;$

$d_{2,\overline{X}(1,2,3)}=(4-4,7)^2+(7-7,7)^2=0,98;$

$d_{3,\overline{X}(1,2,3)}=(8-4,7)^2+(6-7,7)^2=13,78.$

Аналогично вычислим квадраты расстояний от объектов n_4 и n_5 до центра тяжести кластера *S(4,5)*:

$d_{4,\overline{X}(4,5)}=(12-12,5)^2+(11-10)^2=1,25;\ d_{5,\overline{X}(4,5)}=(13-12,5)^2+(9-10)^2=1,25.$

F_1=12,58+0,98+13,78+1,25+1,25=29,84.

Вычислим критерий F_2. Для этого необходимо просуммировать квадраты расстояний внутри каждого кластера. Для первого кластера *S(1,2,3)* необходимо вычислить

$d^2_{12}+d^2_{13}+d^2_{23}=(3,61)^2+(7,21)^2+(4,12)^2=81,99$; для второго кластера *S(4,5)* используется только одно расстояние $d^2_{45}=(2,24)^2=5,02$. Таким образом, значение F_2=81,99+5,02=87,01.

Вычислим критерий F_3. Для этого вычислим вариацию каждой переменной (X_1 и X_2) по двум кластерам. Вариация переменной X_1 в кластере *S(1,2,3)*: $(2-4,7)^2+(4-4,7)^2+(8-4,7)^2=18,67$. Вариация переменной X_2 в кластере *S(1,2,3)*: $(10-7,7)^2+(7-7,7)^2+(6-7,7)^2=8,67$. Вариация переменной X_1 в кластере *S(4,5)*: $(12-12,5)^2+(13-12,5)^2=0,5$. Вариация переменной X_2 в кластере *S(4,5)*: $(11-10)^2+(9-10)^2=2$. F_3=18,67+8,67+0,5+2=29,84.

Рассчитаем значения F_1, F_2 и F_3 для разбиений на кластеры *S(1,2)* и *S(3,4,5)*. Чтобы вычислить критерий F_1, необходимо создать две таблицы исходных данных, соответствующих кластерам *S(1,2)* и *S(3,4,5)*.

№ семьи	X_1	X_2
1	2	10
2	4	7
$\bar{X}(1,2)$	3	8,5

№ семьи	X_1	X_2
3	8	6
4	12	11
5	13	9
$\bar{X}(3,4,5)$	11	13

Вычисляем координаты центра тяжести каждого кластера. Для кластера *S(1,2)* центр тяжести $\bar{X}(1,2)=(3;8,5)$. Для кластера *S(3,4,5)* центр тяжести $\bar{X}(3,4,5)=(11;13)$. Вычислим квадраты расстояний от объектов n_1 и n_2 до центра тяжести кластера *S(1,2)*: $d_{1,\bar{X}(1,2)} = (2-3)^2 + (10-8,5)^2 = 3,25$; $d_{2,\bar{X}(1,2)} = (4-7)^2 + (7-8,5)^2 = 11,25$. Аналогично вычислим квадраты расстояний от объектов n_3, n_4 и n_5 до центра тяжести кластера *S(3,4,5)*: $d_{3,\bar{X}(4,5)} = (8-11)^2 + (6-13)^2 = 57$; $d_{4,\bar{X}(4,5)} = (12-11)^2 + (11-13)^2 = 5$; $d_{5,\bar{X}(4,5)} = (13-11)^2 + (9-13)^2 = 18$. F_1=3,25+11,25+57+5+18=94,5.

Вычислим критерий F_2. Просуммируем квадраты расстояний внутри каждого кластера. Для первого кластера *S(1,2)* необходимо вычислить $d^2_{12}=(3,61)^2=13,03$; для второго кластера *S(3,4,5)* $d^2_{34}+d^2_{35}+d^2_{45}=(6,4)^2+(5,83)^2+(2,24)^2=79,97$. Таким образом, значение F_2=79,97+13,03=93.

Вычислим критерий F_3. Для этого вычислим вариацию каждой переменной (X_1 и X_2) по двум кластерам. Вариация переменной X_1 в кластере *S(1,2)*: $(2-3)^2 + (4-3)^2 = 2$. Вариация переменной X_2 в кластере *S(1,2)*: $(10-8,5)^2 + (7-8,5)^2 = 4,5$. Вариация переменной X_1 в кластере *S(3,4,5)*: $(8-11)^2 + (12-11)^2 + (13-11)^2 = 14$. Вариация переменной X_2 в кластере *S(3,4,5)*: $(6-13)^2 + (11-13)^2 + (9-13)^2 = 69$. F_3=2+4,5+14+69=89,5.

Составим сводную таблицу для функционалов, рассчитанных для различных методов. Так как в методах «ближнего соседа», «средней связи» и центроидного классификация совпадает, то оставим две колонки в сводной таблице.

Методы		«ближнего соседа», «средней связи», центроидный (кластеры *S(1,2,3)* и *S(4,5)*)	«дальнего соседа» (кластеры *S(1,2)* и *S(3,4,5)*)
Функционалы	F_1	29,84	94,5
	F_2	87,01	93
	F_3	29,84	89,5

Из сводной таблицы видно, что разбиение на два кластера *S(1,2,3)* и *S(4,5)* является самым оптимальным, так как все критерии классификации имеют наименьшие значения.

ДИВИЗИМНЫЙ АЛГОРИТМ КЛАСТЕРНОГО АНАЛИЗА

Кроме рассмотренных агломеративных методов иерархического кластерного анализа, существуют методы, противоположные им по логическому построению процедур классификации. Они называются иерархическими дивизимными методами. Основной исходной посылкой дивизимного метода является то, что первоначально все объекты принадлежат одному кластеру. В процессе классификации по определенным правилам постепенно от этого кластера отделяются группы схожих между собой объектов. Таким образом, на каждом шаге количество кластеров возрастает, а мера расстояния между кластерами уменьшается. Дендрограмма дивизимного метода представлена в виде дерева (рис. 7).

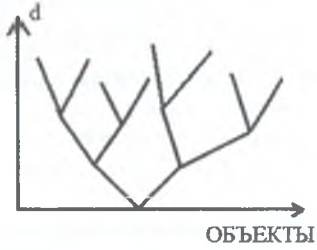

Рис. 7. Дендрограмма дивизимного алгоритма

Итак, первоначально все объекты принадлежат одному кластеру. По таблице расстояний необходимо найти наибольшее расстояние, предположим d_{ij} – максимальное, это означает, что на расстоянии d_{ij} i-й и j-й объекты разделяются.

Далее необходимо выяснить, как распределяются остальные объекты. Для этого необходимо сравнить расстояния от каждого из объектов до i-го и j-го объектов. Если расстояние от произвольного k-го объекта до i-го объекта меньше, чем до j-го, то k-й объект присоединяется к i-му объекту. Если же расстояние от k-го объекта до i-го объекта больше,чем до j-го, то k-й объект присоединяется к j-му объекту. Т.е., условии $d_{ki} < d_{kj} \Rightarrow k$-й объект присоединяется к i-му объекту, при $d_{ki} > d_{kj} \Rightarrow k$-й объект присоединяется к j-му объекту.

В каждом образовавшемся кластере необходимо выбрать наибольшее расстояние из всех возможных расстояний между объектами кластера и повторить процедуру, рассмотренную выше.

Проведём классификацию пяти семей по дивизимному алгоритму.

	n_1	n_2	n_3	n_4	n_5
n_1	0	3,61	7,21	10,05	11,05
n_2	3,61	0	4,12	8,94	9,22
n_3	7,21	4,12	0	6,4	5,83
n_4	10,05	8,94	6,4	0	2,24
n_5	11,05	9,22	5,83	2,24	0

Из таблицы расстояний видно, что максимальное расстояние 11,05 – расстояние между объектами n_1 и n_5. Следовательно, на расстоянии $d_{1,5}$=11,05 данные объекты разделяются и образуют кластеры *S(1)* и *S(5)*.

Выясним, как разделятся оставшиеся объекты n_2, n_3 и n_4. Выделим из таблицы расстояний расстояния от объектов n_2, n_3 и n_4 до кластеров *S(1)* и *S(5)*.

	n_1	n_5	Сравнение расстояний	Вывод
n_2	3,61	9,22	3,61<9,22	n_2 присоединяется к *S(1)*
n_3	7,21	5,83	7,21>5,83	n_3 присоединяется к *S(2)*
n_4	10,05	2,24	10,05>2,24	n_4 присоединяется к *S(2)*

Таким образом, образовались два кластера *S(1,2)* и *S(3,4,5)*. Если в результате классификации необходимо оставить два кластера, то на этом дивизимный алгоритм заканчивается. Если же исследователь должен получить три кластера, то дивизимный алгоритм продолжается для кластера *S(3,4,5)*. В исходной таблице расстояний остаются расстояния между объектами кластера *S(3,4,5)*.

S(3,4,5)	n_3	n_4	n_5
n_3	0	6,4	5,83
n_4	6,4	0	2,24
n_5	5,83	2,24	0

Видно, что максимальное расстояние 6,4 – расстояние между объектами n_3 и n_4. Следовательно, на расстоянии $d_{3,4}$=6,4 данные объекты разделяются и образуют кластеры *S(3)* и *S(4)*. Выясним, к какому кластеру присоединится объект n_5. Сравним расстояния от объекта n_5 до кластеров *S(3)* и *S(4)*: $d_{5,3}$=5,83 > $d_{5,4}$=2,24. Таким образом, объект n_5 присоединяется к кластеру *S(4)*. В результате сформированы три кластера: *S(1,2)*, *S(4,5)* и *S(3)*. На рис. 8 представлена дендрограмма.

Интерпретация полученной дендрограммы дивизимного алгоритма: видно, что два кластера *S(1,2)*, *S(3,4,5)* имеют максимальную меру. Разделение кластера *S(3,4,5)* происходит на значительно меньшем расстоянии, поэтому исследователь вправе оставить в рассмотрении два кластера.

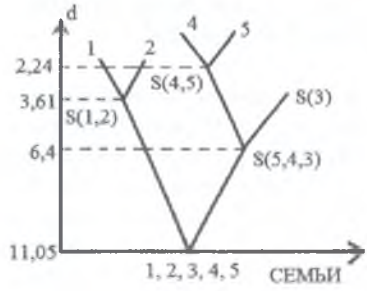

Рис. 8. Дендрограмма дивизимного метода

КЛАССИФИКАЦИЯ НА ОСНОВЕ «ВЗВЕШЕННОЙ» ЕВКЛИДОВОЙ МЕТРИКИ

Рассмотрим классификацию семей на основе «взвешенного евклидова расстояния». Как рассматривалось ранее, «взвешенное евклидово расстояние» между объектами определяется с помощью соотношения: $d_{ij}=\sqrt{\omega_1(x_{i1}-x_{j1})^2+\omega_2(x_{i2}-x_{j2})^2+\cdots+\omega_m(x_{im}-x_{jm})^2}$. Естественно предположить, что расходам на питание (признак X_2) придается существенно больший вес при классификации семей по потребительскому поведению. Пусть вес ω_1=0,05, а вес ω_2=0,95. Учитывая численные значения для веса признаков X_1 и X_2, вычислим меру сходства между объектами:

$d_{12}=\sqrt{(2-4)^2\,0{,}05+(10-7)^2\,0{,}95}=2{,}96$; $d_{13}=\sqrt{(2-8)^2\,0{,}05+(10-6)^2\,0{,}95}=4{,}12$;

$d_{14}=\sqrt{(2-12)^2\,0{,}05+(10-11)^2\,0{,}95}=2{,}44$; $d_{15}=\sqrt{(2-13)^2\,0{,}05+(10-9)^2\,0{,}95}=2{,}65$;

$d_{23}=\sqrt{(4-8)^2\,0{,}05+(7-6)^2\,0{,}95}=1{,}32$; $d_{24}=\sqrt{(4-12)^2\,0{,}05+(7-11)^2\,0{,}95}=4{,}29$;

$d_{25}=\sqrt{(4-13)^2\,0{,}05+(7-9)^2\,0{,}95}=2{,}8$; $d_{34}=\sqrt{(8-12)^2\,0{,}05+(6-11)^2\,0{,}95}=4{,}95$;

$d_{35}=\sqrt{(8-13)^2\,0{,}05+(6-9)^2\,0{,}95}=3{,}13$; $d_{45}=\sqrt{(12-13)^2\,0{,}05+(11-9)^2\,0{,}95}=1{,}96$.

Составим таблицу «взвешенных расстояний» и **проведём классификацию методом «ближнего соседа».**

	n_1	n_2	n_3	n_4	n_5
n_1	0	2,96	4,12	2,44	2,65
n_2	2,96	0	**1,32**	4,29	2,8
n_3	4,12	1,32	0	4,95	3,13
n_4	2,44	4,29	4,95	0	1,96
n_5	2,65	2,8	3,13	1,96	0

Из таблицы видно, что минимальное расстояние 1,32 – это расстояние между объектами n_2 и n_3. Следовательно, эти объекты образуют первый кластер *S(2,3)*. Далее необходимо пересчитать расстояния от объектов n_1, n_4 и n_5 до первого кластера *S(2,3)*. В методе «ближнего соседа» $d(1;S(2,3))=min\{2{,}96;4{,}12\}=2{,}96$; . $d(4;S(2,3))=min\{4{,}29;4{,}95\}=4{,}29$; $d(5;S(2,3))=min\{2{,}8;3{,}13\}=2{,}8$.

Таблица расстояний после пересчета расстояний принимает вид:

	n_1	*S(2,3)*	n_4	n_5
n_1	0	2,96	2,44	2,65
S(2,3)	2,96	0	4,29	2,8
n_4	2,44	4,29	0	**1,96**
n_5	2,65	2,8	1,96	0

Минимальное расстояние 1,96. Следовательно, эти объекты n_4 и n_5 образуют второй кластер *S(4,5)*. Пересчитаем расстояния от объекта n_1 и кластера *S(2,3)* до нового кластера *S(4,5)*: $d(1;S(4,5))=min\{2{,}44;2{,}65\}=2{,}44$; $d(S(2,3);S(4,5))=min\{4{,}29;\ 2{,}8\}=2{,}8$.

Таблица расстояний после пересчета расстояний принимает вид.

	n_1	$S(2,3)$	$S(4,5)$
n_1	0	2,96	**2,44**
$S(2,3)$	2,96	0	2,8
$S(4,5)$	2,44	2,8	0

Минимальное расстояние 2,44 – это расстояние между объектом n_1 и кластером $S(4,5)$. Следовательно, первый объект присоединяется к кластеру $S(4,5)$. Образуется новый кластер $S(1,4,5)$.$d(S(2,3);S(1,4,5))=min\{2,96;2,8\}=2,8$.

	$S(2,3)$	$S(1,4,5)$
$S(2,3)$	0	2,8
$S(1,4,5)$	2,8	0

Два кластера могут объединиться на расстоянии 2,8. На этом классификация по методу «ближнего соседа» заканчивается.

Проведём классификацию методом «дальнего соседа».

Объекты n_2 и n_3 образуют первый кластер $S(2,3)$ на расстоянии 1,32.
$d(1;S(2,3))=max\{2,96;4,12\}=4,12$; $d(4;S(2,3))= max\ \{4,29; 4,95\}=4,95$;
$d(5;S(2,3))=max\{2,8;3,13\}=3,13$.

Таблица расстояний после пересчета расстояний принимает вид:

	n_1	$S(2,3)$	n_4	n_5
n_1	0	4,12	2,44	2,65
$S(2,3)$	4,12	0	4,95	3,13
n_4	2,44	4,95	0	**1,96**
n_5	2,65	3,13	1,96	0

На расстоянии 1,96 объекты n_4 и n_5 образуют второй кластер S(4,5). Пересчитываем расстояния от всех объектов до нового кластера:
$d(1;S(4,5))=max\ \{2,44;2,65\}=2,65$;
$d(S(2,3);S(4,5))=max\{4,95; 3,13\}=4,95$.

Таблица расстояний после пересчета расстояний принимает вид.

	n_1	$S(2,3)$	$S(4,5)$
n_1	0	4,12	**2,65**
$S(2,3)$	4,12	0	4,95
$S(4,5)$	2,65	4,95	0

Минимальное расстояние 2,65 – это расстояние между объектом n_1 и кластером $S(4,5)$. Следовательно, первый объект присоединяется к кластеру $S(4,5)$. Образуется новый кластер $S(1,4,5)$. Расстояние от кластера $S(2,3)$ до кластера $S(1,4,5)$: $d(S(2,3);S(1,4,5))=max\{4,12;4,95\}=4,95$.

	S(2,3)	S(1,4,5)
S(2,3)	0	4,95
S(1,4,5)	4,95	0

Два кластера могут объединиться на расстоянии 4,95. На этом классификация по методу «дальнего соседа» заканчивается. Результаты классификации по двум методам совпали. Пять семей разбиваются на два однородных по свойству кластера $S(2,3)$ и $S(1,4,5)$.

Проведём классификацию методом «средней связи».

На расстоянии 1,32 объекты n_2 и n_3 образуют первый кластер $S(2,3)$:
$d(1;S(2,3))=(2,96+4,12)/2=3,54$; $d(4;S(2,3))=(4,29+4,95)/2=4,62$;
$d(5;S(2,3))=(2,8+3,13)/2=2,97$.

Таблица расстояний после пересчета расстояний принимает вид:

	n_1	$S(2,3)$	n_4	n_5
n_1	0	3,54	2,44	2,65
$S(2,3)$	3,54	0	4,62	2,97
n_4	2,44	4,62	0	**1,96**
n_5	2,65	2,97	1,96	0

На расстоянии 1,96 объекты n_4 и n_5 образуют второй кластер $S(4,5)$.
$d(1;S(4,5))$=(2,44+2,65)/2=2,55;
$d(S(2,3);S(4,5))$=(4,62+2,97)/2=3,8.

Таблица расстояний после пересчета расстояний принимает вид:

	n_1	S(2,3)	S(4,5)
n_1	0	3,54	**2,65**
S(2,3)	3,54	0	3,8
S(4,5)	2,65	3,8	0

Минимальное расстояние – 2,65. Следовательно, первый объект присоединяется к кластеру $S(4,5)$. Образуется новый кластер $S(1,4,5)$. Пересчет расстояний: $d(S(2,3);S(1,4,5))$=(3,54+2,55)/2=3,05. Два кластера могут объединиться на расстоянии 3,05.

	$S(2,3)$	$S(1,4,5)$
$S(2,3)$	0	3,05
$S(1,4,5)$	3,05	0

На этом классификация по методу «средней связи» заканчивается. Результаты классификации по трём методам совпали. Пять семей разбиваются на два однородных по свойству кластера $S(2,3)$ и $S(1,4,5)$. Структура дендрограмм совпадает, различны только расстояния, соответствующие объединению объектов. Результаты классификации для метода «ближнего соседа» представлены графически в виде дендрограммы на рис 9.

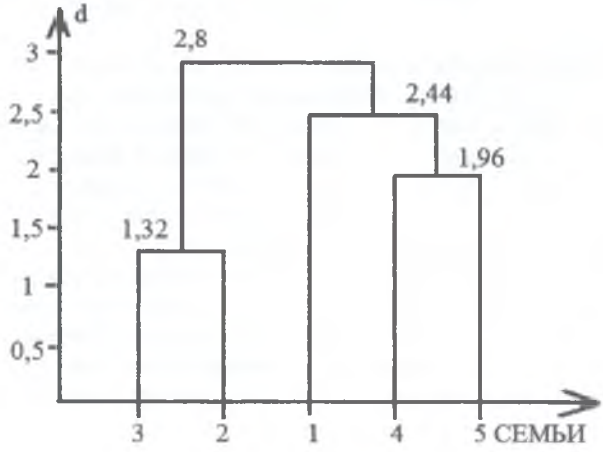

Рис. 9. Дендрограмма (метод «ближнего соседа», «взвешенная евклидова метрика»)

Многомерный корреляционный анализ

1. В таблице представлены результаты исследований миграции населения некоторого региона. Построить матрицу корреляций. Вычислить множественный и частные коэффициенты корреляции. Проверить их значимость на уровне α=0,05.

Причины миграции населения

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Миграционный фактор	Поиск работы	Поиск жилья	Смена профессии	Вступление в брак	Развод	Рождение ребёнка	Уход на пенсию
18,7	14,6	9,2	15	10,4	16,6	15,4	16,8
8,3	16,4	3,4	15	24,7	20,8	22,4	13,8
9,5	8,1	5,3	13,2	18	17,2	11,8	7,3
10,6	12	8,2	10,9	13,3	12,2	12,4	7,3
6,2	7,6	4,3	6,3	14,2	13	11,8	9
9,5	5,4	4,3	8,6	6,6	7,6	7,8	14,7
8,8	4,8	6,8	13,8	15,2	9,4	7,6	15,5
8,8	3,9	5,3	4,0	10,4	6,2	5,2	3,0
9	3,5	4,3	4	2,8	2,2	3	9,5
3,4	4,1	4,2	4,6	7,6	10	7	6,4
6,2	3	3,9	3,4	2,8	8	6	2,1
2,5	3	3,4	4,6	5,7	3	3,8	7,7
4,8	3,7	6,3	1,7	0,9	2	2,6	3,8
2	3,7	1,4	8	2,8	4,8	2,4	7,3
0,9	2,4	0,9	1,1	5,7	4,8	4,8	5,6
1,3	3,7	3,9	0,5	5,7	3,4	3,4	2,1
1,1	1,3	0,4	2,3	1,9	2,6	2,2	5,1
2,7	3,5	0,6	2,3	4,7	3,6	3	1,2
0,2	3,9	2,9	1,7	1,9	3	5,2	1,2
1,8	2,8	0,4	5,7	1,9	4,8	2,8	3,4
0,9	0,2	0,4	1,7	1,9	0,6	0,6	3,8

2. В таблице представлены результаты социологического исследования. Построить матрицу корреляций. Вычислить множественный и частные коэффициенты корреляции. Проверить гипотезы о статистической значимости данных коэффициентов.

	X_1	X_2	X_3	X_4	X_5
N_1	50	12,9	9,9	2,6	24,6
N_2	42,4	17,7	15,3	1,5	23,1
N_3	68,8	0,8	11,9	2,7	15,8
N_4	49,6	13,5	13,5	7,5	15,9
N_5	50,4	9,3	12	5,9	22,4
N_6	44,5	21,9	12,7	2,2	18,7
N_7	57,4	10,3	9,8	1,8	20,7
N_8	59	6,9	9,1	1,9	23,1
N_9	54,6	15,2	12,6	2,4	15,2

3. Имеются данные, характеризующие показатели качества жизни, выделенной по группе стран, представленных в таблице:

Страна	Продолжительность предстоящей жизни, лет	Уровень грамотности взрослого населения, %	Доля учащихся среди молодежи, %	Реальный ВВП на душу населения, $
Аргентина	72,6	96,2	79	8498
Бразилия	66,6	83,3	61	5928
Венесуэла	72,3	91,1	67	8090
Сингапур	77,1	91,1	68	22604
Колумбия	70,3	91,3	69	6347
Таиланд	69,5	93,8	55	7742
Малайзия	71,4	83,5	61	9572
Мексика	72,1	89,6	67	6769
Турция	68,5	82,3	60	5516
Оман	70,7	59	60	9383
Кувейт	75,4	78,6	58	23848
Гонконг	79	92,2	67	22950
Чили	75,1	95,2	73	9930
Бахрейн	72,2	85,2	84	16751
Фиджи	72,1	91,6	78	6159

Задание:

1. Вычислить компонентные индексы, составляющие индекс развития человеческого потенциала (ИРЧП), по каждой стране: индекс ожидаемой продолжительности жизни $i_{p_1} = \frac{p_{1факт} - 25}{85 - 25}$; индекс достигнутого уровня

образования $i_{обр}=[2i_{p2}+i_{p3}]:3$, где $i_{p_2}=\frac{p_{2факт}-0}{100-0}$ – индекс грамотности населения, $i_{p_3}=\frac{p_{3факт}-0}{100-0}$ – индекс числа поступивших в учебные заведения первого, второго и третьего уровней; индекс скорректированного реального ВВП на душу населения, ППС в долл. США $i_{p_4}=\frac{\ln p_{4факт}-\ln 100}{\ln 40000-\ln 100}$.

2. Вычислить ИРЧП по каждой стране: ИРЧП=$(i_{p1}+i_{обр}+i_{p4})/3$.

3. Вычислить коэффициенты ранговой корреляции Спирмена, Кендалла, характеризующие зависимость рангов по ИРЧП от рангов по каждому его компоненту. Какой компонент оказывает наибольшее влияние на значение ИРЧП по анализируемой группе стран?

4. Имеются показатели, характеризующие обездоленность населения в развивающихся африканских странах (%):

Страна	Население, которое не доживет до 40 лет	Уровень грамотности населения	Население, не имеющее доступа к		Дети в возрасте до 5 лет с пониженной массой тела
			доброкачественной воде	медицинскому обслуживанию	
	p_1	p_2	p_{31}	p_{32}	p_{33}
1	26	63,4	50	20	14
2	23	64,5	35	40	27
3	31	57,1	50	49	36
4	33	51,7	45	39	19
5	32	40,1	18	70	24
6	29	37,7	26	37	23
7	21	45,8	66	62	34
8	42	54,9	41	60	23
9	38	35,9	54	20	26
10	36	31	34	60	27
11	34	35,5	75	54	48
12	38	19,2	22	10	30
13	36	13,6	52	1	36
14	50	31,4	66	62	29
15	38	40,1	37	61	27

Вычислите индекс нищеты населения (ИНН) для каждой страны $ИНН = \sqrt[3]{\frac{p_1^3 + p_2^3 + p_3^3}{3}}$, где $p_3 = \frac{p_{31} + p_{32} + p_{33}}{3}$. Вычислите парные и частные коэффициенты корреляции между ИНН и его компонентами. Вычислите множественный коэффициент корреляции, характеризующий зависимость ИНН от его составляющих.

5. При приеме на работу семи кандидатам на вакантные должности было предложено два теста. Результаты тестирования в баллах приведены в таблице:

Тест	Кандидаты						
	1	2	3	4	5	6	7
1	31	82	25	26	53	30	29
2	21	55	8	27	32	42	26

Вычислить ранговые коэффициенты корреляции Спирмена и Кендалла между результатами тестирования по двум тестам и на уровне α=0,05 оценить их значимость.

6. На соревнованиях по фигурному катанию 9 судей выставили следующие балльные оценки 10 фигуристам:

Фигурист	Судья								
	1	2	3	4	5	6	7	8	9
1	6,0	5,8	5,7	5,8	6,0	5,9	5,9	5,9	5,8
2	5,4	5,3	5,2	5,3	5,4	5,5	5,6	5,3	5,1
3	5,2	5,0	4,9	5,1	5,2	5,0	4,8	5,3	4,9
4	5,9	5,9	5,8	5,7	5,9	5,8	6,0	5,8	5,7
5	5,0	4,9	4,9	4,9	5,1	5,0	5,0	4,8	4,7
6	5,6	5,5	5,4	5,4	5,5	5,5	5,7	5,6	5,5
7	4,8	4,7	4,6	4,6	4,8	4,9	5,0	4,6	4,5
8	5,4	5,6	5,4	5,5	5,6	5,7	5,4	5,3	5,2
9	5,8	5,7	5,6	5,7	5,8	5,9	5,6	5,7	5,8
10	5,3	5,2	5,1	5,4	5,5	5,4	5,2	5,3	5,2

Вычислить коэффициент конкордации рангов и оценить его значимость на уровне α=0,05.

7. В таблицах приведены данные, характеризующие некоторые экономические параметры регионов. Проверить гипотезу о равенстве векторов средних значений этих регионов, а также гипотезу о равенстве матриц ковариаций. Считая, что векторы μ^T=(1100; 1350; 210; 15) для первого региона и μ^T=(900; 850; 230; 15) для второго региона, проверить гипотезы о равенстве вектора средних значений вектору μ для каждого региона.

Северный и Северо-Западный районы				
Регион	Среднедушевой денежный доход в месяц, руб.	Среднемесячная заработная плата работников предприятий и организаций, руб.	Величина прожиточного минимума, руб	Уровень безработицы, %
	X_1	X_2	X_3	X_4
Республика Карелия	1052	1228	175	16,6
Республика Коми	1156	1699	219	17,8
Архангельская область	711	1168	123	14,9
Вологодская область	800	1187	168	12,7
Мурманская область	1533	1711	221	21
Санкт-Петербург	1060	1148	173	11,3
Ленинградская область	649	961	128	15
Новгородская область	896	862	204	15,4
Псковская область	544	710	134	16,1

Центральный район				
Область	Среднедушевой денежный доход в месяц, руб.	Средне-месячная заработная плата работ-ников предприятий и организаций, руб.	Величина прожиточного минимума, руб	Уровень безработицы, %
	X_1	X_2	X_3	X_4
Брянская	554	606	156	15,7
Владимирская	589	740	151	12
Ивановская	530	629	144	18,8
Калужская	640	794	158	10,2
Костромская	586	771	152	11,2

Окончание таблицы

Москва	4017	1522	595	4,8
Московская	703	1036	157	9,9
Орловская	693	686	180	13,2
Рязанская	568	704	146	7,1
Смоленская	712	775	185	16,4
Тверская	537	768	133	11,3
Тульская	721	755	188	11,6
Ярославская	741	888	173	11,1

8. Чтобы оценить производственную эффективность предложенной к внедрению технологии, проведена проверка качества продукции, выпущенной на старой и новой автоматических линиях, при этом получены следующие данные об удельном весе продукции высшего качества в %:

Партия	Старая линия		
№	X_1	X_2	X_3
1	58	14	3,6
2	62	18	4,4
3	51	12	4,2
4	67	16	3,9
5	41	11	3,4
6	53	9	2,8

Партия	Новая линия		
№	X_1	X_2	X_3
1	74	4	2,8
2	59	7	2,6
3	69	12	4,1
4	78	6	2,3
5	82	8	3,5
6	75	11	3,8
7	86	5	2,2
8	63	11	3,7

При уровне значимости 0,01 установить, действительно ли новая линия, налаженная на передовую технологию, позволяет получать более высокий уровень качества продукции? Выяснить, имеют ли данные линии одинаковую взаимосвязь признаков в выборке?

9. Для оценки существенности воздействия состояния окружающей среды на здоровье людей в районе с неблагоприятной экологической обстановкой проведены медицинские обследования 12 отобранных случайных групп населения. Известно, что средний по республике уровень продолжительности жизни составляет 69 лет, заболеваемости онкологическими болезнями – 580 случаев на 100 000 жителей, уровень младенческой смертности 12%. На уровне значимости 0,02 определить, действительно ли факторы окружающей среды оказывают существенное негативное влияние на уровень здоровья населения. После проверки гипотезы по всем трем характерным признакам проверьте значимость каждого признака в отдельности, сделайте выводы.

Половозрастная группа населения	Средний уровень продолжительности жизни, лет	Заболеваемость онкологическими болезнями, на 100 000 жителей	Уровень младенческой смертности, %
	X_1	X_2	X_3
1	64	590	18
2	58	604	17
3	67	598	15
4	66	610	17
5	71	690	14
6	56	540	21
7	58	624	18
8	62	670	16
9	64	656	14
10	61	711	15
11	63	630	16
12	68	705	11

10. Проверьте гипотезу о равенстве матриц ковариаций предприятий двух отраслей «А» и «В» по следующим данным (уровень значимости 0,01).

Отрасль А		
Предприятия	Рентабельность производства, %	Среднегодовая выработка на одного работника, тыс. руб.
№	X_1	X_2
1	14	3,6
2	18	4,4
3	12	4,2
4	16	3,9
5	11	3,4
6	9	2,8

Отрасль В		
Предприятия	Рентабельность производства, %	Среднегодовая выработка на одного работника, тыс. руб.
№	X_1	X_2
1	4	2,8
2	7	2,6
3	12	4,1
4	6	2,3
5	8	3,5
6	11	3,8
7	5	2,2
8	11	3,7

11. В таблицах представлены основные социально-экономические показатели Поволжского региона за 1997 и 1998 годы. Проверить гипотезу о равенстве векторов средних значений для 1997 и 1998 годов. Уровень значимости 0,05. Основные показатели: X_1 – среднедушевой денежный доход в месяц, руб.; X_2 – среднемесячная начисленная зарплата работников предприятий и организаций, руб.; X_3 – величина прожиточного минимума, руб.; X_4 – уровень безработицы, %.

Поволжский регион 1997 г.

Регион	X_1	X_2	X_3	X_4
Республика Калмыкия	431	542	132	26,1
Республика Татарстан	691	885	242	7,9
Астраханская область	589	713	165	14,6
Волгоградская область	672	736	201	14,4
Пензенская область	488	546	133	12,0
Самарская область	928	1075	256	9,3
Ульяновская область	594	670	219	9,8
Саратовская область	620	604	169	15,8

Поволжский регион 1998 г.

Регион	X_1	X_2	X_3	X_4
Республика Калмыкия	431	610	108	30,8
Республика Татарстан	748	949	217	10,9
Астраханская область	636	849	139	15,9
Волгоградская область	639	817	152	14,7
Пензенская область	453	598	103	18,1
Самарская область	1164	1161	265	8,6
Ульяновская область	646	681	157	16,1
Саратовская область	617	740	203	11,1

Для данных за 1997 год проверить гипотезу о равенстве вектора средних значений вектору (520; 780; 200; 13). В случае неравенства данных векторов проверить частные критерии Хотеллинга.

12. В таблицах представлены две обучающие выборки. Провести классификацию объектов с помощью дискриминантного анализа.

Регион	Среднедушевой денежный доход, руб.	Средняя зарплата работников предприятий и организаций, руб.	Величина прожиточного минимума, руб.	Уровень безработицы,%
Высокий уровень				
Республика Карелия	1023	1097	208	11,9
Республика Коми	1260	1485	266	13,9
Архангельская область	792	1074	168	12,4
Владимирская область	568	661	168	11,6
Калужская область	639	701	198	11,2
Костромская область	605	667	189	9,4
Низкий уровень				
Псковская область	534	632	164	14,2
Брянская область	595	532	206	12,9
Ивановская область	546	547	177	16,9
Орловская область	651	610	209	9,8
Рязанская область	603	614	194	10,1
Смоленская область	647	644	218	12,9
Подлежат дискриминации				
Вологодская область	831	1094	206	10,5
Мурманская область	1300	1655	233	18,5
Санкт-Петербург	1022	1037	224	9,9
Ленинградская область	601	870	167	12,8
Новгородская область	757	758	213	13,5
Москва	3516	1250	664	4,8
Московская область	662	927	182	8,8
Пермская область	534	654	170	9,9
Тульская область	709	678	234	10
Ярославская область	727	787	210	8,8

13. В таблицах представлены две обучающие выборки. Провести классификацию объектов с помощью дискриминантного анализа.

Регион	Среднедушевой денежный доход, в руб.	Средняя зарплата работников предприятий и организаций, руб.	Величина прожиточного минимума, руб.	Уровень безработицы, %
Высокий уровень				
Иркутская область	983	1281	208	14,4
Приморский край	843	1191	168	13,3
Хабаровский край	899	1292	179	12,7
Амурская область	873	1135	183	15,6
Низкий уровень				
Республика Бурятия	738	943	179	21,3
Республика Хакасия	758	1021	167	13
Еврейская авт. область	666	890	141	25,7
Подлежат дискриминации				
Республика Тыва	590	772	105	22
Красноярскийкрай	1042	1401	249	13,3
Читинская область	570	996	102	18,5
Республика Саха	1741	2270	187	12,6
Чукотский авт.окр.	1872	2816	140	8,4
Камчатская область	1649	2096	190	12,5
Магаданская область	1516	2018	175	13,6
Сахалинская область	1127	1665	151	15
Калининградская область	595	718	173	11,5

14. В таблицах представлены две обучающие выборки. Провести классификацию объектов с помощью дискриминантного анализа.

	Показатель	Объем реализованной продукции	
№ района	Уровень использования земли	Растениеводства	Животноводства
1	Низкий	0,25	0,41
2		0,51	0,51
3		0,27	0,42
4		0,33	0,56
5	Высокий	1,17	0,28
6		4,99	0,67
7		5,18	0,45
8		2,49	0,38
9		2,73	0,33
10	Подлежат дискриминации	0,32	0,45
11		0,67	0,32
12		4,6	0,56

Кластерный анализ

15. Провести классификацию городов, используя агломеративные методы с алгоритмами «ближнего соседа», «дальнего соседа», «средней связи», «центроидного». Построить дендрограммы. Вычислить функционалы качества разбиения. Провести классификацию, используя дивизимный метод. Провести классификацию, используя взвешенную евклидову метрику методом «средней связи». Вес указан в таблице.

Города	Минимальная заработанная плата, руб. (0,4)	Среднедушевой доход в месяц, руб. (0,5)	Место в России (0,1)
	X_1	X_2	X_3
Москва	2269	1908	19
Белгород	1717	1382	44
Иваново	1184	912	76
Брянск	1213	1150	64
Орел	1335	1325	49
Тамбов	1234	1433	40
Ярославль	1906	1683	29

16. Провести классификацию регионов, используя агломеративные методы с алгоритмами «ближайшего соседа», «дальнего соседа», «средней связи», «центроидного». Построить дендрограммы. Вычислить функционалы качества разбиения. Провести классификацию, используя дивизимный метод. Провести классификацию, используя взвешенную евклидову метрику методом «средней связи». Вес указан в таблице.

Область	Оплата труда (0,75)	Доходы от собственности (0,25)
Брянская	33,6	2,4
Владимирская	44,2	3,0
Ивановская	41,1	3,6
Калужская	40,8	2,5
Костромская	44,4	2,0
Москва	17,6	11,7
Московская	43,9	3,8

17. Провести классификацию электоратов ведущих партий, используя агломеративные методы с алгоритмами «ближайшего соседа», «дальнего соседа», «средней связи», «центроидного». Построить дендрограммы. Вычислить функционалы качества разбиения. Провести классификацию, используя дивизимный метод. Провести классификацию, используя взвешенную евклидову метрику методом «ближайшего соседа». Вес указан в таблице.

	Партии					
Индекс	КПРФ	ЛДПР	ОВР	Единство	Яблоко	СПС
Настроение	0,8	0,9	2,3	1,9	1,8	2,9
Удовлетворенность жизнью (0,2)	0,1	0,5	0,5	0,5	0,4	1,2
Оценка материального положения семьи (0,05)	0,9	1,3	0,9	1,5	2	2,3
Оценка материального положения в городе, сел. районе (0,03)	0,5	0,7	0,9	1,2	0,8	2,9
Оценка экономического положения в России (0,3)	0,3	0,6	0,3	0,8	0,4	0,8
Отношение к реформам (0,1)	0,5	1,7	2,5	3,3	3,2	8,1
Оценка политической обстановки (0,1)	0,1	0,3	0,4	0,3	0,4	0,5
Прогноз политического развития (0,125)	1,3	1,9	4,3	6	3,1	4,8
Прогноз экономического развития (0,125)	1,3	1,9	3,1	4,1	2,5	6,3

Рекомендуемый библиографический список

1. Дубров, А.М. Многомерные статистические методы / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин – М.: Финансы и статистика, 1998. – 352 с.
2. Голуб, Л.А. Социально-экономическая статистика: учеб. пособие для вузов / Л.А. Голуб. – М.: Гуманит. изд. центр ВЛАДОС, 2001. – 272 с.
3. Сошникова, Л.А. Многомерный статистический анализ в экономике: учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тимашевич, Г. Уебе, М. Шеффер; под общ. ред. В.Н. Тимашевича; – М.: ЮНИТИ-ДАНА, 1999. – 598 с.
4. Айвазян, С.А. Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. / С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ-ДАНА, 2001.

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$ *Приложение 1*

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,33	0,1293	0,66	0,2454	0,99	0,3389
0,01	0,0040	0,34	0,1331	0,67	0,2486	1,00	0,3413
0,02	0,0080	0,35	0,1368	0,68	0,2517	1,01	0,3438
0,03	0,0120	0,36	0,1406	0,69	0,2549	1,02	0,3461
0,04	0,0160	0,37	0,1443	0,70	0,2580	1,03	0,3485
0,05	0,0199	0,38	0,1480	0,71	0,2611	1,04	0,3508
0,06	0,0239	0,39	0,1517	0,72	0,2642	1,05	0,3531
0,07	0,0279	0,40	0,1554	0,73	0,2673	1,06	0,3554
0,08	0,0319	0,41	0,1591	0,74	0,2703	1,07	0,3577
0,09	0,0359	0,42	0,1628	0,75	0,2734	1,08	0,3599
0,10	0,0398	0,43	0,1664	0,76	0,2764	1,09	0,3621
0,11	0,0438	0,44	0,1700	0,77	0,2794	1,10	0,3643
0,12	0,0478	0,45	0,1736	0,78	0,2823	1,11	0,3665
0,13	0,0517	0,46	0,1772	0,79	0,2852	1,12	0,3686
0,14	0,0557	0,47	0,1808	0,80	0,2881	1,13	0,3708
0,15	0,0596	0,48	0,1844	0,81	0,2910	1,14	0,3729
0,16	0,0636	0,49	0,1879	0,82	0,2939	1,15	0,3749
0,17	0,0675	0,50	0,1915	0,83	0,2967	1,16	0,3770
0,18	0,0714	0,51	0,1950	0,84	0,2995	1,17	0,3790
0,19	0,0753	0,52	0,1985	0,85	0,3023	1,18	0,3810
0,20	0,0793	0,53	0,2019	0,86	0,3051	1,19	0,3830
0,21	0,0832	0,54	0,2054	0,87	0,3078	1,20	0,3849
0,22	0,0871	0,55	0,2088	0,88	0,3106	1,21	0,3869
0,23	0,0910	0,56	0,2123	0,89	0,3133	1,22	0,3883
0,24	0,0948	0,57	0,2157	0,90	0,3159	1,23	0,3907
0,25	0,0987	0,58	0,2190	0,91	0,3186	1,24	0,3925
0,26	0,1026	0,59	0,2224	0,92	0,3212	1,25	0,3944
0,27	0,1064	0,60	0,2257	0,93	0,3238	1,26	0,3962
0,28	0,1103	0,61	0,2291	0,94	0,3264	1,27	0,3980
0,29	0,1141	0,62	0,2324	0,95	0,3289	1,28	0,3997
0,30	0,1179	0,63	0,2357	0,96	0,3315	1,29	0,4015
0,31	0,1217	0,64	0,2389	0,97	0,3340	1,30	0,4032
0,32	0,1255	0,65	0,2422	0,98	0,3365	1,31	0,4049

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,32	0,4066	1,65	0,4505	1,98	0,4761	2,62	0,4956
1,33	0,4082	1,66	0,4515	1,99	0,4767	2,64	0,4959
1,34	0,4099	1,67	0,4525	2,00	0,4772	2,66	0,4961
1,35	0,4115	1,68	0,4535	2,02	0,4783	2,68	0,4963
1,36	0,4131	1,69	0,4545	2,04	0,4793	2,70	0,4965
1,37	0,4147	1,70	0,4554	2,06	0,4803	2,72	0,4967
1,38	0,4162	1,71	0,4564	2,08	0,4812	2,74	0,4969
1,39	0,4177	1,72	0,4573	2,10	0,4821	2,76	0,4971
1,40	0,4192	1,73	0,4582	2,12	0,4830	2,78	0,4973
1,41	0,4207	1,74	0,4591	2,14	0,4838	2,80	0,4974
1,42	0,4222	1,75	0,4599	2,16	0,4846	2,82	0,4976
1,43	0,4236	1,76	0,4608	2,18	0,4854	2,84	0,4977
1,44	0,4251	1,77	0,4616	2,20	0,4861	2,86	0,4979
1,45	0,4265	1,78	0,4625	2,22	0,4868	2,88	0,4980
1,46	0,4279	1,79	0,4633	2,24	0,4875	2,90	0,4981
1,47	0,4292	1,80	0,4641	2,26	0,4881	2,92	0,4982
1,48	0,4306	1,81	0,4649	2,28	0,4887	2,94	0,4984
1,49	0,4319	1,82	0,4656	2,30	0,4893	2,96	0,4985
1,50	0,4332	1,83	0,4664	2,32	0,4898	2,98	0,4986
1,51	0,4345	1,84	0,4671	2,34	0,4904	3,00	0,49865
1,52	0,4357	1,85	0,4678	2,36	0,4909	3,20	0,49931
1,53	0,4370	1,86	0,4686	2,38	0,4913	3,40	0,49966
1,54	0,4382	1,87	0,4693	2,40	0,4918	3,60	0,499841
1,55	0,4394	1,88	0,4699	2,42	0,4922	3,80	0,499928
1,56	0,4406	1,89	0,4706	2,44	0,4927	4,00	0,499968
1,57	0,4418	1,90	0,4713	2,46	0,4931	4,50	0,499997
1,58	0,4429	1,91	0,4719	2,48	0,4934	5,00	0,499997
1,59	0,4441	1,92	0,4726	2,50	0,4938		
1,60	0,4452	1,93	0,4732	2,52	0,4941		
1,61	0,4463	1,94	0,4738	2,54	0,4945		
1,62	0,4474	1,95	0,4744	2,56	0,4948		
1,63	0,4484	1,96	0,4750	2,58	0,4951		
1,64	0,4495	1,97	0,4756	2,60	0,4953		

Таблица значений распределения Стъюдента *Приложение 2*

Число степеней свободы	Уровень значимости						
k	**0,001**	**0,01**	**0,02**	**0,05**	**0,1**	**0,15**	**0,2**
1	636,578	63,656	31,821	12,706	6,314	4,165	3,078
2	31,6	9,925	6,965	4,303	2,92	2,282	1,886
3	12,924	5,841	4,541	3,182	2,353	1,924	1,638
4	8,61	4,604	3,747	2,776	2,132	1,778	1,533
5	6,869	4,032	3,365	2,571	2,015	1,699	1,476
6	5,959	3,707	3,143	2,447	1,943	1,65	1,44
7	5,408	3,499	2,998	2,365	1,895	1,617	1,415
8	5,041	3,355	2,896	2,306	1,86	1,592	1,397
9	4,781	3,25	2,821	2,262	1,833	1,574	1,383
10	4,587	3,169	2,764	2,228	1,812	1,559	1,372
11	4,437	3,106	2,718	2,201	1,796	1,548	1,363
12	4,318	3,055	2,681	2,179	1,782	1,538	1,356
13	4,221	3,012	2,65	2,16	1,771	1,53	1,35
14	4,14	2,977	2,624	2,145	1,761	1,523	1,345
15	4,073	2,947	2,602	2,131	1,753	1,517	1,341
16	4,015	2,921	2,583	2,12	1,746	1,512	1,337
17	3,965	2,898	2,567	2,11	1,74	1,508	1,333
18	3,922	2,878	2,552	2,101	1,734	1,504	1,33
19	3,883	2,861	2,539	2,093	1,729	1,5	1,328
20	3,85	2,845	2,528	2,086	1,725	1,497	1,325
21	3,819	2,831	2,518	2,08	1,721	1,494	1,323
22	3,792	2,819	2,508	2,074	1,717	1,492	1,321
23	3,768	2,807	2,5	2,069	1,714	1,489	1,319
24	3,745	2,797	2,492	2,064	1,711	1,487	1,318
25	3,725	2,787	2,485	2,06	1,708	1,485	1,316
26	3,707	2,779	2,479	2,056	1,706	1,483	1,315
27	3,689	2,771	2,473	2,052	1,703	1,482	1,314
28	3,674	2,763	2,467	2,048	1,701	1,48	1,313
29	3,66	2,756	2,462	2,045	1,699	1,479	1,311
30	3,646	2,75	2,457	2,042	1,697	1,477	1,31
40	3,551	2,704	2,423	2,021	1,684	1,468	1,303
60	3,46	2,66	2,39	2	1,671	1,458	1,296
120	3,373	2,617	2,358	1,98	1,658	1,449	1,289
∞	3,291	2,526	2,326	1,96	1,645	1,442	1,282

Число степеней свободы	Уровень значимости						
k	**0,001**	**0,01**	**0,02**	**0,05**	**0,1**	**0,15**	**0,2**
1	10,827	6,635	5,412	3,841	2,706	2,072	1,642
2	13,815	9,21	7,824	5,991	4,605	3,794	3,219
3	16,266	11,345	9,837	7,815	6,251	5,317	4,642
4	18,466	13,277	11,668	9,488	7,779	6,745	5,989
5	20,515	15,086	13,388	11,07	9,236	8,115	7,289
6	22,457	16,812	15,033	12,592	10,645	9,446	8,558
7	24,321	18,475	16,622	14,067	12,017	10,748	9,803
8	26,124	20,09	18,168	15,507	13,362	12,027	11,03
9	27,877	21,666	19,679	16,919	14,684	13,288	12,242
10	29,588	23,209	21,161	18,307	15,987	14,534	13,442
11	31,264	24,725	22,618	19,675	17,275	15,767	14,631
12	32,909	26,217	24,054	21,026	18,549	16,989	15,812
13	34,527	27,688	25,471	22,362	19,812	18,202	16,985
14	36,124	29,141	26,873	23,685	21,064	19,406	18,151
15	37,698	30,578	28,259	24,996	22,307	20,603	19,311
16	39,252	32	29,633	26,296	23,542	21,793	20,465
17	40,791	33,409	30,995	27,587	24,769	22,977	21,615
18	42,312	34,805	32,346	28,869	25,989	24,155	22,76
19	43,819	36,191	33,687	30,144	27,204	25,329	23,9
20	45,314	37,566	35,02	31,41	28,412	26,498	25,038
21	46,796	38,932	36,343	32,671	29,615	27,662	26,171
22	48,268	40,289	37,659	33,924	30,813	28,822	27,301
23	49,728	41,638	38,968	35,172	32,007	29,979	28,429
24	51,179	42,98	40,27	36,415	33,196	31,132	29,553
25	52,619	44,314	41,566	37,652	34,382	32,282	30,675
26	54,051	45,642	42,856	38,885	35,563	33,429	31,795
27	55,475	46,963	44,14	40,113	36,741	34,574	32,912
28	56,892	48,278	45,419	41,337	37,916	35,715	34,027
29	58,301	49,588	46,693	42,557	39,087	36,854	35,139
30	59,702	50,892	47,962	43,773	40,256	37,99	36,25
35	66,619	57,342	54,244	49,802	46,059	43,64	41,778
40	73,403	63,691	60,436	55,758	51,805	49,244	47,269
45	80,078	69,957	66,555	61,656	57,505	54,81	52,729
50	86,66	76,154	72,613	67,505	63,167	60,346	58,164
55	93,167	82,292	78,619	73,311	68,796	65,855	63,577
60	99,608	88,379	84,58	79,082	74,397	71,341	68,972
65	105,988	94,422	90,501	84,821	79,973	76,807	74,351
70	112,317	100,425	96,387	90,531	85,527	82,255	79,715

0,05	k_1					
k_2	1	2	3	4	5	6
1	161,446	199,499	215,707	224,583	230,16	233,988
2	18,513	19	19,164	19,247	19,296	19,329
3	10,128	9,552	9,277	9,117	9,013	8,941
4	7,709	6,944	6,591	6,388	6,256	6,163
5	6,608	5,786	5,409	5,192	5,05	4,95
6	5,987	5,143	4,757	4,534	4,387	4,284
7	5,591	4,737	4,347	4,12	3,972	3,866
8	5,318	4,459	4,066	3,838	3,688	3,581
9	5,117	4,256	3,863	3,633	3,482	3,374
10	4,965	4,103	3,708	3,478	3,326	3,217
11	4,844	3,982	3,587	3,357	3,204	3,095
12	4,747	3,885	3,49	3,259	3,106	2,996
13	4,667	3,806	3,411	3,179	3,025	2,915
14	4,6	3,739	3,344	3,112	2,958	2,848
15	4,543	3,682	3,287	3,056	2,901	2,79
16	4,494	3,634	3,239	3,007	2,852	2,741
17	4,451	3,592	3,197	2,965	2,81	2,699
18	4,414	3,555	3,16	2,928	2,773	2,661
19	4,381	3,522	3,127	2,895	2,74	2,628
20	4,351	3,493	3,098	2,866	2,711	2,599
21	4,325	3,467	3,072	2,84	2,685	2,573
22	4,301	3,443	3,049	2,817	2,661	2,549
23	4,279	3,422	3,028	2,796	2,64	2,528
24	4,26	3,403	3,009	2,776	2,621	2,508
25	4,242	3,385	2,991	2,759	2,603	2,49
26	4,225	3,369	2,975	2,743	2,587	2,474
27	4,21	3,354	2,96	2,728	2,572	2,459
28	4,196	3,34	2,947	2,714	2,558	2,445
29	4,183	3,328	2,934	2,701	2,545	2,432
30	4,171	3,316	2,922	2,69	2,534	2,421
35	4,121	3,267	2,874	2,641	2,485	2,372
40	4,085	3,232	2,839	2,606	2,449	2,336
45	4,057	3,204	2,812	2,579	2,422	2,308
50	4,034	3,183	2,79	2,557	2,4	2,286
55	4,016	3,165	2,773	2,54	2,383	2,269
60	4,001	3,15	2,758	2,525	2,368	2,254
∞	3,846	3	2,609	2,376	2,219	2,103

0,05	k_1					
k_2	**8**	**12**	**20**	**24**	**30**	∞
1	238,884	243,905	248,016	249,052	250,096	253,465
2	19,371	19,412	19,446	19,454	19,463	19,489
3	8,845	8,745	8,66	8,638	8,617	8,545
4	6,041	5,912	5,803	5,774	5,746	5,652
5	4,818	4,678	4,558	4,527	4,496	4,392
6	4,147	4	3,874	3,841	3,808	3,698
7	3,726	3,575	3,445	3,41	3,376	3,26
8	3,438	3,284	3,15	3,115	3,079	2,959
9	3,23	3,073	2,936	2,9	2,864	2,739
10	3,072	2,913	2,774	2,737	2,7	2,572
11	2,948	2,788	2,646	2,609	2,57	2,439
12	2,849	2,687	2,544	2,505	2,466	2,332
13	2,767	2,604	2,459	2,42	2,38	2,243
14	2,699	2,534	2,388	2,349	2,308	2,169
15	2,641	2,475	2,328	2,288	2,247	2,105
16	2,591	2,425	2,276	2,235	2,194	2,049
17	2,548	2,381	2,23	2,19	2,148	2,001
18	2,51	2,342	2,191	2,15	2,107	1,958
19	2,477	2,308	2,155	2,114	2,071	1,92
20	2,447	2,278	2,124	2,082	2,039	1,886
21	2,42	2,25	2,096	2,054	2,01	1,855
22	2,397	2,226	2,071	2,028	1,984	1,827
23	2,375	2,204	2,048	2,005	1,961	1,802
24	2,355	2,183	2,027	1,984	1,939	1,779
25	2,337	2,165	2,007	1,964	1,919	1,757
26	2,321	2,148	1,99	1,946	1,901	1,738
27	2,305	2,132	1,974	1,93	1,884	1,719
28	2,291	2,118	1,959	1,915	1,869	1,702
29	2,278	2,104	1,945	1,901	1,854	1,686
30	2,266	2,092	1,932	1,887	1,841	1,672
35	2,217	2,041	1,878	1,833	1,786	1,61
40	2,18	2,003	1,839	1,793	1,744	1,564
45	2,152	1,974	1,808	1,762	1,713	1,527
50	2,13	1,952	1,784	1,737	1,687	1,498
55	2,112	1,933	1,764	1,717	1,666	1,473
60	2,097	1,917	1,748	1,7	1,649	1,453
∞	1,943	1,757	1,576	1,523	1,465	1,207

Оглавление

Учебное издание

Трусова Алла Юрьевна

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Часть 1

Учебное пособие

Редактор Т.И. Кузнецова
Компьютерная верстка, макет Т.В. Кондратьевой

Подписано в печать 28.04.08. Формат 60x84/16. Бумага офсетная. Печать офсетная.
Усл.-печ. л. 4,0. Гарнитура Times.
Тираж 250 экз. Заказ №1521
Издательство «Самарский университет», 443011, г. Самара, ул. Акад. Павлова, 1.
Тел. 8 (846) 334-54-23
Отпечатано на УОП СамГУ