МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)

*А.Ю. ТРУСОВА*

# АНАЛИЗ ДАННЫХ. МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Рекомендовано редакционно-издательским советом федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева» в качестве учебного пособия для обучающихся по основной образовательной программе высшего образования по направлению подготовки 38.03.05 Бизнес-информатика

С А М А Р А
Издательство Самарского университета
2023

УДК 519.237(075)
ББК В172.6я7
Т789

Рецензенты: канд. физ.-мат. наук, доц. Л. К. Ш и р я е в а,
канд. техн. наук, доц. З. Ф. К а м а л ь д и н о в а

***Трусова, Алла Юрьевна***

Т789 **Анализ данных. Многомерные статистические методы:** учебное пособие / *А.Ю. Трусова*. – Самара: Издательство Самарского университета, 2023. – 92 с.

**ISBN 978-5-7883-2029-8**

Учебное пособие рассчитано на обучающихся по программе бакалавриата. Призвано помочь обучающимся лучше усвоить содержание и разобраться в основах анализа данных и многомерных статистических методах. Соответствуют требованиям государственного образовательного стандарта высшего профессионального образования по указанному направлению.

Подготовлено на кафедре математики и бизнес-информатики.

УДК 519.237(075)
ББК В172.6я7

ISBN 978-5-7883-2029-8 

## ОГЛАВЛЕНИЕ

# 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 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 трудоемкие алгоритмы реализации вычислительных процедур, практически всегда рассчитанные на привлечение технических средств, и сложная интерпретируемость аналитических результатов. Это требует от пользователя достаточно серьезной подготовки как в области математической статистики, так и в области, в которой проводятся конкретные исследования.

# Глава 1. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ определился как статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для последующего планирования аналогичных экспериментов.

По числу факторов, влияние которых исследуется, различают однофакторный и многофакторный дисперсионный анализ.

**Однофакторный дисперсионный анализ**

Общий вид модели однофакторного дисперсионного анализа имеет вид:

$X_{ij} = \mu + F_i + \varepsilon_{ij}$,

где $X_{ij}$ – значение исследуемой переменной, полученной на $i$-м уровне фактора $\left(i = \overline{1,l}\right)$ с $j$ порядковым номером $\left(j = \overline{1,n}\right)$;

$F_i$ – эффект, обусловленный влиянием $i$-го уровня фактора;

$\mu$ – среднее значение;

$\varepsilon_{ij}$ – случайная компонента, обусловленная влиянием неконтролируемых факторов, т.е. вариацией переменной внутри отдельного уровня.

Под уровнем фактора понимается некая его мера или состояние.

**Основные предпосылки дисперсионного анализа**

1. Математическое ожидание от случайной компоненты ($M(\varepsilon_{ij})$)
2. Случайные компоненты ($\varepsilon_{ij}$) не зависимы
3. Дисперсии $\varepsilon_{ij}$ равны $\sigma^2$, т.е. постоянны для любых $ij$
4. $X_{ij}\ (\varepsilon_{ij}) \in N\ (0, \sigma^2)$

Влияние уровня фактора может быть как фиксированным или систематическим (*модель I*), так и случайным (*модель II*). Например, необходимо выяснить, имеются ли существенные различия между партиями по некоторому показателю качества, т.е. необхо-

димо проверить влияние на качество одного фактора партии изделий. Если включить в исследование все партии сырья, то влияние уровня такого фактора систематическое (*модель I*), а полученные выводы применимы только к тем отдельным партиям, которые изучались при исследовании. Если же включить только отобранную случайную часть партии, то влияние фактора случайное (*модель II*). В многофакторных комплексах возможна смешанная *модель III*, в которой одни факторы имеют случайные уровни, а другие фиксированные.

Рассмотрим задачу. Дано n партий изделий, из каждой партии отобраны $n_1, n_2, n_3, \ldots, n_m$ изделий. Для простоты $n_1 = n_2 = n_3 = \ldots = n_m = n$.

Составим матрицу показателей:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \ldots & x_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{m1} & x_{m2} & x_{m3} & \ldots & x_{mn} \end{pmatrix}$$

$i$ – номер партии, $j$ – номер изделия в составе партии. Необходимо проверить существенность влияния партии изделий на их качество. В предположении, что элементы строк – наблюдения, представляющие собой случайные величины $X_1, X_2, X_3, \ldots, X_m$, имеющие математические ожидания $a_1, a_2, a_3, \ldots, a_m$ и одинаковые дисперсии ($\sigma^2$), данная задача сводится к проверке нулевой гипотезы $H_o$: $a_1 = a_2 = a_3 = \ldots = a_m$, т.е. проверке гипотезы об отсутствии влияния уровней фактора на результат эксперимента.

Введем обозначения для усреднения:

$\overline{x_{i\bullet}}$ – групповая средняя $i$-й партии или групповая средняя $i$-го уровня фактора:

$$\overline{x_{i\bullet}} = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

$\overline{x_{\bullet\bullet}}$ – общая средняя:

$$\overline{x_{\bullet\bullet}} = \frac{\sum_{i=1}^{m}\sum_{i=1}^{n} x_{ij}}{mn} = \frac{\sum_{i=1}^{m} \overline{x_{i\bullet}}}{m}$$

Рассмотрим сумму квадратов отклонения наблюдений $x_{ij}$ от общей средней $\left(\overline{x_{\bullet\bullet}}\right)$.

$$Q = \sum_{i=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_{\bullet\bullet}}\right)^2 = \sum_{i=1}^{m}\sum_{i=1}^{n}\left(\left(x_{ij} - \overline{x_{i\bullet}}\right) + \left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right)\right)^2 =$$

$$\sum_{i=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_{i\bullet}}\right)^2 + 2\sum_{i=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_{i\bullet}}\right)\left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right) +$$

$$+ \sum_{i=1}^{m}\sum_{i=1}^{n}\left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right)^2 = \sum_{i=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_{i\bullet}}\right)^2 + \sum_{i=1}^{m}\sum_{i=1}^{n}\left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right)^2$$

$$Q_1 = \sum_{i=1}^{m}\sum_{i=1}^{n}\left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right)^2 = n\sum_{i=1}^{m}\left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right)^2$$

$$Q_2 = \sum_{i=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_{i\bullet}}\right)^2$$

$Q$ – общая или полная сумма квадратов отклонения;

$Q_1$ – сумма квадратов отклонения групповых средних от общей средней или межгрупповая факторная сумма квадратов отклонений.

Т.о. проверка нулевой гипотезы $H_o$ сводится к проверке существенности различия несмещенных выборочных оценок $S_1^2$ и $S_2^2$ дисперсии $\sigma^2$.

Составим статистику:

$$F_{набл} = \frac{S_1^2}{S_2^2}$$

$F_{кр}(\alpha, k_1, k_2)$, $k_1 = m_1 - 1$, $k_2 = mn - m$

Гипотеза $H_o$ отвергается, если $F_{набл} > F_{кр}$. Применительно к данной задаче это означает наличие существенных различий в качестве изделий различных партий на заданном уровне значимости $\alpha$.

***Пример***

Имеется четыре партии сырья для текстильной промышленности. Из каждой партии отобрано по 5 образцов и проведены испытания на определение величины разрывы. Нагрузки. Результат приведен в таблице.

Необходимо выяснить существенно ли влияние различных партий сырья на величину разрывной нагрузки.

| Номер партии | *I* | *II* | *III* | *IV* | *V* | $\overline{x_{i\bullet}}$ |
|---|---|---|---|---|---|---|
| *1.* | 200 | 140 | 170 | 145 | 165 | 164 |
| *2.* | 190 | 150 | 210 | 150 | 150 | 170 |
| *3.* | 230 | 190 | 200 | 190 | 200 | 202 |
| *4.* | 150 | 170 | 150 | 170 | 180 | 164 |

$m = 4$, $n = 5$.

Чтобы найти $\overline{x_{i\bullet}}$, нужно сумму цифр в средней строке поделить на $n$. $\overline{x_{\bullet\bullet}} = 175$ (суммируется столбик $\overline{x_{i\bullet}}$ и делится на $m$).

$$Q_1 = n\sum_{i=1}^{m}\left(\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}}\right)^2 = 5\left[(164-175)^2 + (170-175)^2 + (202-175)^2 + (164-175)^2\right] = 5 \cdot 996 = 4980$$

$$Q_2 = \sum_{i=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_{i\bullet}}\right)^2 = 7270$$

Сводная таблица:

| Компоненты дисперсий | Суммы квадратов | Число степеней свободы | Средние квадраты |
|---|---|---|---|
| Межгрупповая ($Q_2$) | 4980 | 3 | 1660 |
| Внутригрупповая ($Q_1$) | 7270 | 16 | 454,4 |
| Общая ($Q = Q_1 + Q_2$) | 12250 | 19 | |

$$F_{набл} = \frac{S_1^{\,2}}{S_2^{\,2}} = \frac{1660}{454{,}4} = 3{,}65 \quad F_{кр}\,(0{,}05;\ 3;16) = 3{,}25.$$

$F_{набл} > F_{кр}$ – гипотеза $H_o$ отвергается, т.е. различие между партиями сырья оказывает существенное влияние на величину разрывной нагрузки.

**Понятие о двухфакторном дисперсионном анализе**

Предположим, что в задаче, рассмотренной для однофакторного дисперсионного анализа, изделия изготавливались на разных станках ($l$). Требуется выяснить – имеются ли существенные различия в качестве изделий по каждому фактору.

Фактор $A$ – партия изделия

Фактор $B$ – номер станка

Исходная матрица показателей:

| | $B_1$ | $B_2$ | … | $B_j$ | … | $B_l$ |
|---|---|---|---|---|---|---|
| $A_1$ | $x_{111}$ … $x_{11k}$ | $x_{121}$ … $x_{12k}$ | … | $x_{1j1}$ … $x_{1jk}$ | … | $x_{1l1}$ … $x_{1lk}$ |
| $A_2$ | $x_{211}$ … $x_{21k}$ | $x_{221}$ … $x_{22k}$ | … | $x_{2j1}$ … $x_{2jk}$ | … | $x_{2l1}$ … $x_{2lk}$ |
| … | … | … | … | … | … | … |
| $A_i$ | $x_{i11}$ … $x_{i1k}$ | $x_{i21}$ … $x_{i2k}$ | … | $x_{ij1}$ … $x_{ijk}$ | … | $x_{il1}$ … $x_{ilk}$ |
| … | … | … | … | … | … | … |
| $A_m$ | $X_{m11}$ … $x_{m1k}$ | $X_{m21}$ … $x_{m2k}$ | … | $X_{mj1}$ … $x_{mjk}$ | … | $X_{ml1}$ … $x_{mlk}$ |

По строчкам представлены уровни фактора $A$: $A_i$, $i = \overline{1,m}$

По столбцам представлены уровни фактора $B$: $B_j$, $j = \overline{1,l}$

В ячейках на пересечении $A_i$ *и* $B_j$ находятся значения показателя качества изделия $x_{ijk}$ и таких чисел $ml$.

Двухфакторная дисперсионная модель имеет вид:

$X_{ij} = \mu + F_i + G_j + I_{ij} + \varepsilon_{ij}$.

$X_{ij}$ – значение наблюдения в ячейках $ij$ с индексом $k$.

$F_i$ – эффект, обусловленный влиянием $i$-го уровня фактора $A$.

$\mu$ – общая средняя.

$\varepsilon_{ij}$ – случайная компонента, обусловленная вариацией переменной внутри отдельной ячейки.

$G_j$ – эффект, обусловленный влиянием уровня $j$ фактора $B$.

$I_{ij}$ – эффект, обусловленный влиянием двух факторов, т.е. отклонение от средней по наблюдениям в ячейке $ij$ от суммы первых трех слагаемых в этой модели.

Допущение: $\varepsilon_{ij} \in N(0, \sigma^2)$, $M(F) = M(G) = M(I) = 0$.

Групповые средние вычисляются по формулам:

$\overline{x_{ij\bullet}} = \frac{\sum_{k=1}^{n} x_{ijk}}{n}$, $\overline{x_{i\bullet\bullet}} = \frac{\sum_{j=1}^{l} \overline{x_{ij\bullet}}}{l}$ – среднее значение по строке,

$\overline{x_{\bullet j\bullet}} = \frac{\sum_{i=1}^{m} \overline{x_{ij\bullet}}}{m}$ – среднее значение по столбцу, $\overline{x_{\bullet\bullet\bullet}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{l} \overline{x_{ij\bullet}}}{ml}$ – общая средняя.

Таблица 1. Сводная таблица для двухфакторного дисперсионного анализа

| Компоненты дисперсии | Сумма квадратов | Число степ. свободы | Средние квадраты |
|---|---|---|---|
| Межгрупповая (фактор $A$) | $Q_1 = = l \cdot n \sum_{i=1}^{m} \left(\overline{x_{i\bullet\bullet}} - \overline{x_{\bullet\bullet\bullet}}\right)^2$ | $m - 1$ | $S_1^2 = \frac{Q_1}{m-1}$ |
| Межгрупповая (фактор $B$) | $Q_2 = = mn \sum_{j=1}^{l} \left(\overline{x_{\bullet j\bullet}} - \overline{x_{\bullet\bullet\bullet}}\right)^2$ | $l - 1$ | $S_2^2 = \frac{Q_2}{l-1}$ |

*Окончание табл. 1*

| Компоненты дисперсии | Сумма квадратов | Число степ. свободы | Средние квадраты |
|---|---|---|---|
| Взаимодействие | $Q_3 = n\sum_{i=1}^{m}\sum_{j=1}^{l}\left(\overline{x_{ij\bullet}} - \overline{x_{i\bullet\bullet}} - \overline{x_{\bullet j\bullet}} - \overline{x_{\bullet\bullet\bullet}}\right)^2$ | $(m-1)$ $(l-1)$ | $S_3^2 = \frac{Q_3}{(m-1)(l-1)}$ |
| Остаточная | $Q_4 = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}\left(x_{ijk} - \overline{x_{ij\bullet}}\right)^2$ | $ml$ $(n-1)$ | $S_4^2 = \frac{Q_4}{ml(n-1)}$ |
| Общая | $Q = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}\left(x_{ijk} - \overline{x_{\bullet\bullet\bullet}}\right)^2$ | $mln-1$ | |

Проверка нулевых гипотез для фактора *A* ($H_A$), фактора *B* ($H_B$) и их взаимодействия ($H_{AB}$) об отсутствии влияния на рассматриваемую переменную факторов *A, B* и их взаимодействия осуществляется сравнением отношений $\frac{S_1^2}{S_4^2}$; $\frac{S_2^2}{S_4^2}$ и $\frac{S_3^2}{S_4^2}$ (для *модели I* с фиксированными уровнями факторов) или отношений $\frac{S_1^2}{S_3^2}$; $\frac{S_2^2}{S_3^2}$ и $\frac{S_3^2}{S_4^2}$ (для случайной *модели II*) с соответствующими табличными значениями F критерия Фишера-Снедекора.

## 1.1 Задачи для самостоятельной работы

*Однофакторный дисперсионный анализ*

1. На учебно-опытном участке изучалось влияние различных способов внесения в почву удобрений на урожай зеленой массы некоторой с/х продукции. Каждый вариант опыта имел трехкратную повторяемость. Результаты опыта оказались следующими (кг):

| Номер опыта | Способ внесения удобрения | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| 1 | 21,3 | 23,5 | 24,2 | 29,3 |
| 2 | 28,1 | 22,7 | 30,1 | 28,2 |
| 3 | 31,3 | 28,1 | 29,3 | 27,1 |

С помощью дисперсионного анализа определите влияние фактора способа внесения удобрений со стандартным уровнем значимости.

2. Проведен эксперимент, как изменяется время (мин) решения задачи при различных способах ее предъявления: I – устно, II – письменно, III – в виде текста с графиками и иллюстрациями. Результаты эксперимента представлены в таблице:

| Номер испытуемых | Способы предъявления | | |
|---|---|---|---|
| | I | II | III |
| 1 | 12 | 10 | 10 |
| 2 | 15 | 12 | 10 |
| 3 | 10 | 10 | 9 |
| 4 | 11 | 9 | 8 |
| 5 | 13 | 12 | 10 |

С уровнем значимости $\alpha = 0,05$ установите или отвергните существенность фактора предъявления задания.

**Двухфакторный дисперсионный анализ без повторений**

1. На учебно-опытном участке изучалось влияние различных способов внесения в почву удобрений на урожай зеленой массы некоторой с/х продукции и количества внесенного удобрения. Результаты опыта оказались следующими (кг):

| Количество удобрений | Способ внесения удобрения | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| 100 г | 23,3 | 22,5 | 27,2 | 32,3 |
| 200 г | 25,1 | 29,7 | 32,1 | 30,2 |
| 300 г | 34,3 | 24,1 | 27,3 | 29,1 |

С помощью дисперсионного анализа определите влияние фактора способа внесения удобрений и фактора количества внесенного удобрения на урожай зеленой массы с уровнем значимости 0,05.

2. Проведен эксперимент, как изменяется время (мин) решения задачи при различных способах ее предъявления: I – устно, II – письменно, III – в виде текста с графиками и иллюстрациями и фактора темы: Алгебра, Геометрия, Физика, Химия, Информатика. Результаты эксперимента представлены в таблице:

| Номер испытуемых | Способы предъявления | | |
|---|---|---|---|
| | I | II | III |
| Алгебра | 15 | 11 | 10 |
| Геометрия | 15 | 12 | 10 |
| Физика | 17 | 13 | 9 |
| Химия | 16 | 15 | 13 |
| Информатика | 13 | 12 | 10 |

С уровнем значимости $\alpha = 0,05$ установите или отвергните существенность фактора предъявления задания и фактора темы.

3. Исследуйте влияние различных катализаторов и времени действия их на выход конечного продукта заданной химической реакции. Обозначая катализаторы через $A_1$, $A_2$ … $A_k$ , получим уровни общего «фактора катализа» А. В таблице приведены данные по выходу продукта реакции в граммах.

| Номер наблюдения | Катализаторы | | | | |
|---|---|---|---|---|---|
| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
| 11 мин | 6,2 | 5,6 | 5,9 | 6,7 | 4,0 |
| 9 мин | 6,1 | 5,1 | 4,6 | 7,4 | 4,4 |
| 14 мин | 6,1 | 6,7 | 5,0 | 4,2 | 5,2 |
| 13 мин | 5,8 | 6,9 | 6,1 | 5,3 | 5,5 |
| 11 мин | 6,3 | 5,7 | 6,0 | 4,5 | 5,9 |
| 6 мин | 6,0 | 5,8 | 5,8 | 6,3 | 6,1 |

## Двухфакторный дисперсионный анализ с повторениями

1. В группе из четырех человек измеряется способность к удержанию физического волевого усилия на динамометре (в секундах) правой и левой рукой наедине с экспериментатором в группе однокурсников. С помощью двухфакторного дисперсионного анализа выясните существенность влияния двух факторов – правая, левая рука – в группе и вне группы и их взаимосвязь. Результаты эксперимента представлены в таблице:

| Фактор руки | Фактор группы | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $B_1$ – наедине с экспериментатором | | | | $B_2$ – в группе сокурсников | | | |
| $A_1$ – левая | 10 | 11 | 8 | 10 | 10 | 10 | 5 | 8 |
| $A_2$ – правая | 11 | 13 | 12 | 9 | 15 | 14 | 8 | 7 |

2. Четырем группам испытуемых предъявлялись списки из 10 слов:

1-я группа – короткие слова с большой скоростью,

2-я группа – короткие слова с медленной скоростью,

3-я группа – длинные слова с большой скоростью,

4-я группа – длинные слова с медленной скоростью.

В каждой группе было по 4 испытуемых. Результаты эксперимента представлены в таблице:

| Фактор длины слова | Фактор скорости предъявления слов | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | медленная скорость | | | | большая скорость | | | |
| короткие | 4 | 3 | 3 | 5 | 9 | 8 | 6 | 7 |
| длинные | 7 | 5 | 6 | 7 | 5 | 3 | 3 | 4 |

Установите с помощью двухфакторного дисперсионного анализа наличие или отсутствие значимой взаимосвязи скорости

3. Исследуйте влияние на время (дни) выхода из депрессивного состояния двух факторов – разных уровней интенсивности медикаментозной терапии и уровня интеллекта (IQ) субъектов. Число испытуемых равно 64. В каждую группу входили 4 испытуемых. Результаты эксперимента представлены в таблице:

| Уровень терапии | IQ | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 80 | | | | 90 | | | | 100 | | | | 105 | | | |
| Щадящий | 5 | 0 | 8 | 4 | 6 | 0 | 5 | 8 | 3 | 5 | 1 | 2 | 6 | 4 | 9 | 4 |
| Умеренный | 1 | 3 | 8 | 0 | 0 | 2 | 5 | 9 | 2 | 4 | 8 | 9 | 5 | 6 | 8 | 5 |
| Средний | 17 | 18 | 20 | 15 | 25 | 26 | 19 | 20 | 19 | 24 | 20 | 25 | 35 | 39 | 19 | 25 |
| Интенсивный | 15 | 15 | 18 | 19 | 25 | 19 | 30 | 19 | 30 | 35 | 24 | 22 | 20 | 31 | 24 | 30 |

Установите с помощью двухфакторного дисперсионного анализа значимость ($\alpha$ = 0,05) зависимости времени выхода из депрессии от двух независимых переменных – IQ и интенсивности медикаментозной терапии лечения.

## Глава 2. МНОГОМЕРНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

В многомерном корреляционном анализе изучается связь между группой признаков $X_1, X_2, X_3, \ldots, X_m$. 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:

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

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*) коэф-

фициента корреляции $R_{j0}$, который является обобщением парного коэффициента корреляции $r_{jj}$. Выборочный множественный, или совокупный, коэффициент корреляции вычисляется по формуле:

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

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 $F_{набл} = \frac{R_{j0}^{\ 2}(n-m)}{(1-R_{jo}^{\ 2})(m-1)}$ больше критического значения статистики $F_{кр}$ $(\alpha, k_1, k_2)$, $k_1 = m - 1$, $k_2 = n - m$. 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 элими-

нировании 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 парной 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: $t_{íàáë} = \frac{|r_{ij}|\sqrt{n-m+2}}{\sqrt{1-r_{ij}^2}}$, $t_{кр}(\alpha,k)$ с числом степеней свободы $k=m-n+2$ 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**

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 выборки, имеющих одинаковое значение, называют *связкой*, а количество одинаковых значений в связке – ее *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.

**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. 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 – объем 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 $X_i$, $t_{Xi}$ – количество

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 $\alpha$ 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 $\alpha$, если $t_{набл} > t_{кр}$ – коэффициент ранговой корреляции Спирмена значим на уровне $\alpha$.

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 | 0,5 | 0,5 | 0,5 | 3 | 0 |
| rang $X_j$ | 0,5 | 6 | | 1 | | 7 | | | |

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

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

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

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

**Коэффициент ранговой корреляции Кендалла ($\tau$)**

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 $\tau$ исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными (при n > 10) $\tau$ 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 $\tau$ значим на уровне $\alpha$, 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:

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

где $T_{X_i} = \frac{1}{2}\sum_{i=1}^{m_1}\left(t_{X_i}{}^2 - t_{X_i}\right).$

***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.

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

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

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

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 корреля-

ционной связи при n > 7 статистика $m(n-1)W$ имеет приближенно $\chi^2$ – распределение. Таким образом, $\chi_{набл}^2 = m(n-1)W$, $\chi_{кр}^2 = (\alpha; k)$, $k = n - 1$.

Вывод: $\chi_{набл}^2 > \chi_{кр}^2$ – $W$ 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<br>2<br>3<br>4<br>5 | 1<br>1<br>2<br>1<br>3 | 3<br>2<br>1<br>2<br>1 | 4<br>5<br>7<br>4<br>5 | 2<br>3<br>5<br>6<br>4 | 6<br>6<br>6<br>3<br>2 | 7<br>4<br>4<br>5<br>6 | 5<br>7<br>3<br>7<br>7 | |
| Сумма | 8 | 9 | 25 | 20 | 23 | 26 | 29 | 140<br>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 $\chi^2$ о независимости классификации в таблице сопряженности признаков**

Наблюдаемое значение статистики критерия ХИ-квадрат определяется по формуле:

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

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

$H_o$: не существует различие мнений относительно вопроса $A$ среди разных возрастных групп.

$H_1$: существует различие мнений относительно вопроса $A$ среди разных возрастных групп.

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*.

## 2.1 Задачи для самостоятельной работы

Многомерный корреляционный анализ

1. 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 |

2. 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 $\alpha$=0,05 оценить их значимость.

Вычислить коэффициент конкордации рангов и оценить его значимость на уровне $\alpha$=0,05.

| 2016 | $X_1$ | $X_2$ | $X_4$ | $X_5$ | $X_7$ | $X_8$ |
|---|---|---|---|---|---|---|
| Республика Башкортостан | 7,3 | 1,7 | 0,183 | 0,1 | 0,309 | 0,383 |
| Республика Марий Эл | 5,9 | 1,53 | 0,036 | 0,015 | 0,301 | 0,49 |
| Республика Мордовия | 13,4 | 0,59 | 0,095 | 0,061 | 1 | 1 |
| Республика Татарстан | 21,3 | 1,86 | 0,246 | 0,099 | 0,721 | 0,781 |
| Удмуртская Республика | 7,6 | 0,99 | 0,099 | 0,03 | 0,599 | 0,567 |
| Чувашская Республика | 24,5 | 0,95 | 0,107 | 0,088 | 0,482 | 0,387 |
| Пермский край | 7,9 | 1,52 | 0,341 | 0,195 | 0,57 | 0,711 |
| Кировская область | 9,6 | 0,71 | 0,111 | 0,076 | 0,235 | 0,363 |
| Нижегородская область | 12,8 | 1,4 | 1 | 1 | 0,607 | 0,803 |
| Оренбургская область | 7,1 | 0,61 | 0,06 | 0,014 | 0,147 | 0,131 |

*Окончание табл.*

| 2016 | X1 | X2 | X4 | X5 | X7 | X8 |
|---|---|---|---|---|---|---|
| Пензенская область | 20,1 | 1,2 | 0,294 | 0,17 | 0,283 | 0,289 |
| Самарская область | 3,9 | 1,38 | 0,222 | 0,141 | 0,651 | 0,637 |
| Саратовская область | 4,8 | 0,77 | 0,187 | 0,102 | 0,132 | 0,155 |
| Ульяновская область | 3,6 | 1,61 | 0,345 | 0,406 | 0,452 | 0,656 |

3. Имеются данные по товарообороту (X,тыс. р.) и товарным запасам (Y, тыс. р.) по 10 магазинам области:

| X | 5 | 3 | 24 | 35 | 44 | 55 | 63 | 74 | 82 | 95 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 18 | 12 | 8 | 8 | 8 | 8 | 7 | 6 | 8 | 8 |

Сгруппировать данные по товарообороту в границах 3–35 и 36–95 тыс. р. Найти корреляционное отношение. Составить уравнение регрессии, предварительно определив форму связи.

# Глава 3. 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; $\mu$ – вектор постоянных значений | $H_0$: $\overline{X}_1 = \overline{X}_2$<br>$H_0$: $\overline{X} = \mu$ | $H_1$: $\overline{X}_1 \neq \overline{X}_2$<br>$H_1$: $\overline{X} \neq \mu$ |
| $\Sigma$ – 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 $\mu$**

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: $T^2_{набл} = n\left(\overline{X} - \mu\right)^{\mathrm{T}} S^{-1}\left(\overline{X} - \mu\right)$, 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 $z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$, $\overline{x}_j$ – среднее значение $j$-й компоненты случайной величины $X$, $s_j$ – среднее квадратическое отклонение $j$-й 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 $\alpha$ со степенями свободы $k_1$ и $k_2$ равными $k_1 = m, \quad k_2 = n - m$. 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 $H_0$: $\overline{X} = \mu$, 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:

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

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:

$$C_1 = \begin{pmatrix}1\\0\\0\end{pmatrix},\ C_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}, C_3 = \begin{pmatrix}0\\0\\1\end{pmatrix} \text{ и } \quad C_1^{\text{т}} = \begin{pmatrix}1 & 0 & 0\end{pmatrix},\ C_2^{\text{т}} = \begin{pmatrix}0 & 1 & 0\end{pmatrix},$$

$C_3^{\text{т}} = \begin{pmatrix}0 & 0 & 1\end{pmatrix}$ 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:

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:

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

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, $\widehat{S}$ – несмещенная оценка обобщенной матрицы ковариаций, определяемая соотношением $\hat{S} = \dfrac{n_1 S_1 + n_2 S_2}{n_1 + n_2 - 2}$, $S_1$ и $S_2$ – матрицы ковариаций соответственно первой и второй выборок, $\hat{S}^{-1}$ – обратная матрица обобщенной 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 $\alpha$ 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:

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

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:

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

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**

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:

$$W_{набл} = b \ln \nu,$$

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 $W_{набл} < W_{кр}$.

## 3.1 Задачи для самостоятельной работы

**Проверка многомерных гипотез**

1. 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 $\mu^T$=(1100; 1350; 210; 15) для первого региона и $\mu^T$=(900; 850; 230; 15) для второго региона, проверить гипотезы о равенстве вектора средних значений вектору $\mu$ для каждого региона.

| Область | Среднедушевой денежный доход в месяц, руб. | Среднемесячная заработная плата работников предприятий и организаций, руб. | Величина прожиточного минимума, руб. | Уровень безработицы, % |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| 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 |
| 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 |

*Окончание табл.*

| Область | Среднедушевой денежный доход в месяц, руб. | Среднемесячная заработная плата работников предприятий и организаций, руб. | Величина прожиточного минимума, руб. | Уровень безработицы, % |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| Тверская | 537 | 768 | 133 | 11,3 |
| Тульская | 721 | 755 | 188 | 11,6 |
| Ярославская | 741 | 888 | 173 | 11,1 |

2. 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?

3. 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 |

4. 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 |

5. В таблицах представлены отдельные показатели инновационного потенциала Приволжского Федерального округа за 2016 и 2012 годы. Численность персонала, занятого ИиР, на 10000 населения, занятого в экономике ($X_1$), коэффициент изобретательской

активности ($X_4$), удельный вес инновационной продукции в объеме отгруженной продукции ($X_7$)

На уровне значимости 0,05 проверить гипотезу о статистически значимом различии рассматриваемых показателей.

| 2016 | $X_1$ | $X_4$ | $X_7$ |
|---|---|---|---|
| Республика Башкортостан | 7,3 | 0,183 | 0,309 |
| Республика Марий Эл | 5,9 | 0,036 | 0,301 |
| Республика Мордовия | 13,4 | 0,095 | 1 |
| Республика Татарстан | 21,3 | 0,246 | 0,721 |
| Удмуртская Республика | 7,6 | 0,099 | 0,599 |
| Чувашская Республика | 24,5 | 0,107 | 0,482 |
| Пермский край | 7,9 | 0,341 | 0,57 |
| Кировская область | 9,6 | 0,111 | 0,235 |
| Нижегородская область | 12,8 | 1 | 0,607 |
| Оренбургская область | 7,1 | 0,06 | 0,147 |
| Пензенская область | 20,1 | 0,294 | 0,283 |
| Самарская область | 3,9 | 0,222 | 0,651 |
| Саратовская область | 4,8 | 0,187 | 0,132 |
| Ульяновская область | 3,6 | 0,345 | 0,452 |

# Глава 4. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Задачей дискриминантного анализа является разделить неоднородную совокупность на структурные единицы. Разделение на однородные группы позволяет эффективно использовать моделирование зависимостей между отдельными признаками.

**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.

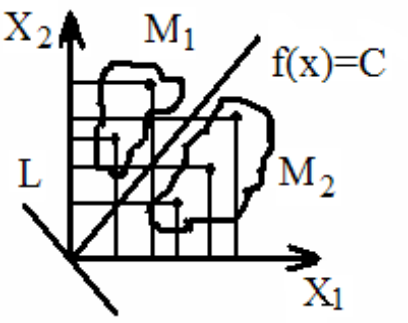


Рис. 1. Геометрическая интерпретация дискриминантной функции и дискриминантных переменных

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 четкую 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*. Введем 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 $a_2$ введем $\overline{x_{ij}}$ – 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).

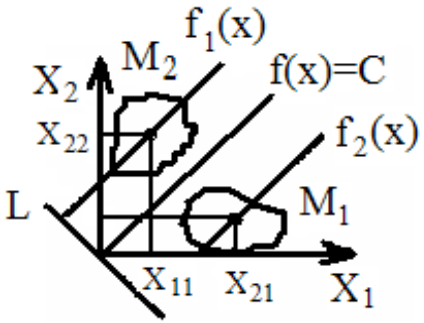


Рис. 2. Центры разделяемых множеств и константа дискриминации

### 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 задается 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 задается 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:

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

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

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

и

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

.

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

$X_{c1}{}^{T} \cdot$

$$X_{c1}=\begin{pmatrix} x_{111}-\overline{x_{11}} & x_{211}-\overline{x_{11}} & x_{311}-\overline{x_{11}} \\ x_{112}-\overline{x_{12}} & x_{212}-\overline{x_{12}} & x_{312}-\overline{x_{12}} \end{pmatrix}.$$

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

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

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

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

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

$X_{c2}{}^{T} \cdot$

$$X_{c2}=\begin{pmatrix} x_{121}-\overline{x_{21}} & x_{221}-\overline{x_{21}} \\ x_{122}-\overline{x_{22}} & x_{222}-\overline{x_{22}} \end{pmatrix}.$$

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

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 мно-

жествах соответственно. Объединенная 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 несмещенной матрицы, характеризующая взаимосвязи между признаками в первом и втором множествах имеет вид: $\widehat{S} = \dfrac{1}{n_1 + n_2 - 2}\left(X_{c1}^T \cdot X_{c1} + X_{c2}^T \cdot X_{c2}\right)$ или

$$\widehat{S} = \frac{1}{n_1 + n_2 - 2}\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

Следовательно, $\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \left(n_1 + n_2 - 2\right)\widehat{S}$. Полученные формулы можно представить в виде несмещенной оценки обобщенной матрицы ковариаций $\widehat{S} = \dfrac{1}{n_1 + n_2 - 2}\left(n_1 S_1 + n_2 S_2\right)$, где $S_1$

и $S_2$ – матрицы ковариаций первой и второй выборок соответственно.

Введем вектор коэффициентов дискриминантной функции $A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$, транспонированный вектор значений коэффициентов $A^T = \begin{pmatrix} a_1 & a_2 \end{pmatrix}$. 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: $A^T (n_1 + n_2 - 2)\widehat{S}A$. Таким образом, оценку вариации дискриминантной функции внутри множеств можно представить в виде:

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

Вариация между множествами может быть оценена как:

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

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

Введем 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:

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

Решением данной задачи является вектор

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

где $\widehat{S}^{-1}$ – обратная матрица к обобщенной 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 несмещенную оценку обобщенной матрицы ковариаций $\widehat{S} = \dfrac{1}{n_1 + n_2 - 2}\left(n_1 S_1 + n_2 S_2\right)$.

4. Вычислить $\widehat{S}^{-1}$.

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 безработных 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 $\left(\overline{X_1} - \overline{X_2}\right)$.

| Высокий уровень | Низкий уровень | Разность |
|---|---|---|
| | | |

2. МатМатрицы ковариаций для обеих групп предприятий:

| $\overline{X_1}$ | $\overline{X_2}$ | $\left(\overline{X_1} - \overline{X_2}\right)$ |
|---|---|---|
| 19,3 | 7,76 | 11,54 |
| 7,75 | 5,52 | 2,23 |

| $S_1$ | $X_1$ | $X_2$ |
|---|---|---|
| $X_1$ | 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 обобщенной матрицы ковариаций $\widehat{S}$ :

| | |
|---|---|
| 5,290286 | 1,396286 |
| 1,396286 | 2,551143 |

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

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

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

| $A$ |
|---|
| 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 трех 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*.

| 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 |

*Окончание табл.*

| Регион | Среднедушевой денежный доход, руб. | Средняя зарплата работников предприятий и организаций, руб. | Величина прожиточного минимума, руб. | Уровень безработицы,% |
|---|---|---|---|---|
| *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 |

## 4.1 Задачи для самостоятельного решения

***Дискриминантный анализ***

1. 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* | | | | |
| 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 |
| Красноярский край | 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 |

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

3. 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 |

1. В таблице представлены объекты – страны СНГ, имеющие высокие и низкие показатели по информационно-коммуникационным технологиям в 2016 году. Рассматриваются следующие показатели классификации: $X_1$ – Численность абонентов фиксированного широкополосного доступа к сети Интернет (на 100 чел. Населения); $X_2$ – Численность абонентов мобильного широкополосного доступа к сети Интернет (на 100 чел. Населения); $X_3$ – Численность персонала, занятого исследованиями и разработками тыс. чел. Провести классификацию стран, относящихся к группе «Подлежат дискриминации»

| Высокий уровень | | | |
|---|---|---|---|
| Объекты | $X_1$ | $X_2$ | $X_3$ |
| 1 | 17 | 64,2 | 829,1 |
| 2 | 19,9 | 46,8 | 23,32 |
| 4 | 28,8 | 55 | 28,93 |
| 5 | 12,9 | 59,8 | 17,58 |
| Низкий уровень уровень | | | |
| 1 | 3 | 2,5 | 4,24 |
| 2 | 0,1 | 0,01 | 3,38 |
| 3 | 0,01 | 0,01 | 3,34 |
| 4 | 2,8 | 0,01 | 35,83 |
| Подлежат дискриминации | | | |
| 1 | 14,7 | 49,4 | 4,14 |
| 2 | 9,1 | 34,2 | 5,62 |
| 3 | 9,3 | 7,5 | 87,39 |

# Глава 5. КЛАСТЕРНЫЙ АНАЛИЗ

## Общая характеристика методов кластерного анализа

Кластерный анализ – совокупность методов, позволяющих классифицировать наблюдения, каждое из которых описывается набором исходных переменных $X_1$, $X_2$, $X_3$,…, $X_m$.

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.

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 $k$-го признака для $i$-го объекта, $x_{jk}$ – значение $k$-го признака для $j$-го 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 $\omega_1$ – вес признака $X_1$, $\omega_2$ – вес признака $X_2$, $\omega_3$– вес признака $X_3$, …, $\omega_m$ – 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 $\omega_k$ устанавливается исследователем самостоятельно, таким образом, что $\sum_{k=1}^{m} \omega_k = 1$. Расстояние *city-block* $d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$.

3. Расстояние Махаланобиса

$d_{ij} = \left(\overline{X_i} - \overline{X_j}\right)^{\mathrm{T}} S^{-1} \left(\overline{X_i} - \overline{X_j}\right)$, 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) $z_{ij} = \dfrac{x_{ij} - \bar{x}}{S_j}$, 2) $z_{ij} = \dfrac{x_{ij}}{x_{\max j}}$, 3) $z_{ij} = \dfrac{x_{ij}}{\overline{x_j}}$,

4) $z_{ij} = \dfrac{x_{ij}}{x_{\min j}}$.

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*-й 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 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 нее 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):

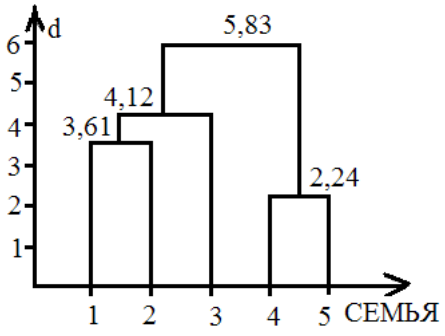


Рис. 3. Дендрограмма
(метод «ближнего соседа»)

| | $n_1$ | $n_2$ | $n_3$ | $S\ (4, 5)$ |
|---|---|---|---|---|
| $n_1$ | 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 |

Продолжим процедуру классификации по методу «дальнего соседа».

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):

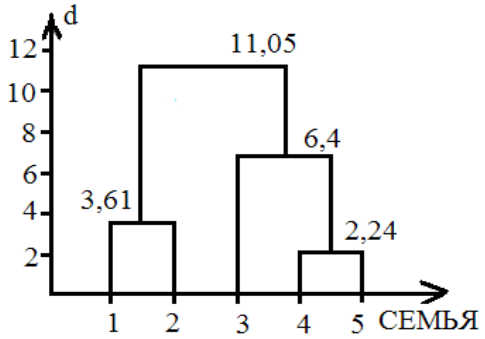


Рис. 4. Дендрограмма (метод «дальнего соседа»)

Проведем 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 нее 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):

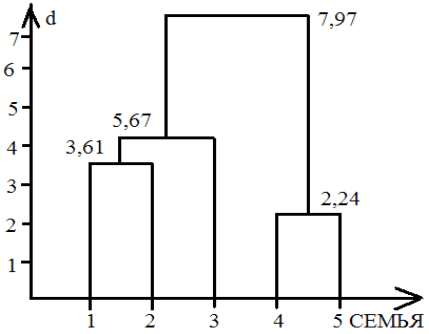


Рис. 5. Дендрограмма (метод «средней связи»)

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):

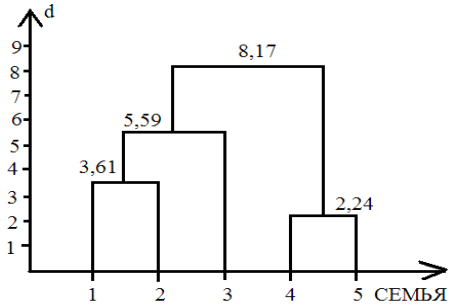


Рис. 6. Дендрограмма (центроидный метод)

Таким образом, сравнивая результаты 4 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 четырех 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 (разделения)**

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. 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 | $F_1$ | $F_{1Б}$ | $F_{1Д}$ | $F_{1С}$ | $F_{1Ц}$ |
| | $F_2$ | $F_{2Б}$ | $F_{2Д}$ | $F_{2С}$ | $F_{2Ц}$ |
| | $F_3$ | $F_{3Б}$ | $F_{3Д}$ | $F_{3С}$ | $F_{3Ц}$ |

Проведем 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)*:

$$d_{1,\overline{X}(1,2,3)} = (2-4,7)^2 + (10-7,7)^2 = 12,58;$$

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

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

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 |
| $\overline{X}(1,2)$ | 3 | 8,5 |

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

Вычисляем координаты центра тяжести каждого кластера. Для кластера *S(1,2)* центр тяжести $\overline{X}(1,2)$=(3;8,5). Для кластера

*S(3,4,5)* центр тяжести $\overline{X}(3,4,5)$=(11;13). Вычислим квадраты расстояний от объектов $n_1$ и $n_2$ до центра тяжести кластера *S(1,2)*:

$$d_{1,\overline{X}(1,2)} = (2-3)^2 + (10-8{,}5)^2 = 3{,}25;$$

$$d_{2,\overline{X}(1,2)} = (4-7)^2 + (7-8{,}5)^2 = 11{,}25.$$

Аналогично вычислим квадраты расстояний от объектов $n_3$, $n_4$ и $n_5$ до центра тяжести кластера *S(3,4,5)*:

$$d_{3,\overline{X}(4,5)} = (8-11)^2 + (6-13)^2 = 57$$

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

$$d_{5,\overline{X}(4,5)} = (13-11)^2 + (9-13)^2 = 18.$$

$F_1$=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).

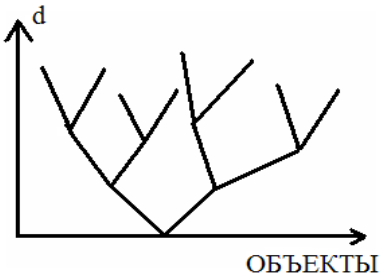


Рис. 7. Дендрограмма дивизимного алгоритма

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.

Проведем 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.

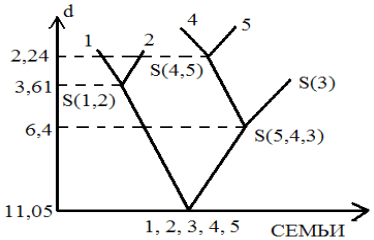


Рис. 8. Дендрограмма дивизимного метода

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.

**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:

$$d_{ij} = \sqrt{\omega_1 \left(x_{i1} - x_{j1}\right)^2 + \omega_2 \left(x_{i2} - x_{j2}\right)^2 + \cdots + \omega_m \left(x_{im} - x_{jm}\right)^2} \ .$$

Естественно предположить, что расходам на питание (признак $X_2$) придается существенно больший вес при классификации семей по потребительскому поведению. Пусть вес $\omega_1$=0,05, а вес $\omega_2$=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 проведем 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.

Проведем 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)*.

Проведем 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 трем 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.

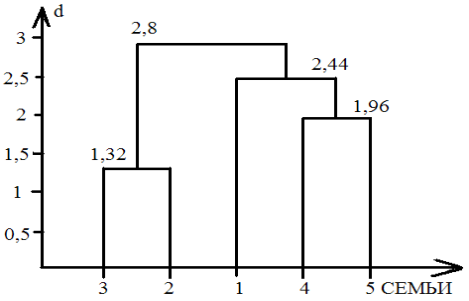


Рис. 9. Дендрограмма
(метод «ближнего соседа», «взвешенная евклидова метрика»)

### 5.1 Задачи для самостоятельной работы

*Кластерный анализ*

1. 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 |

2. 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 |

## Рекомендуемый библиографический список

1. Орлова, И. В. Статистический анализ в экономических задачах: компьютерное моделирование в SPSS [Электронный ресурс] / И. В. Орлова, Н. В. Концевая // Международный журнал прикладных и фундаментальных исследований. – 2014. – № 3. – С. 248–250; URL: https://applied-research.ru/ru/article/view?id=4983 (дата обращения: 24.04.2018).

2. Козлова, А. Ю. Статистический анализ данных в MS Excel: учеб. пособие для вузов / А. Ю. Козлова, В. С. Мхитарян, В. Ф. Шишов. – М.: ИНФРА-М, 2017. – 320 с.

3. Кадочникова Е. И. К вопросу о методах анализа многомерных данных / Е. И. Кадочникова // Путь науки. – 2014. – №5. – С. 64–66.

4. Анализ данных: учеб. пособие для академического бакалавриата / В. С. Мхитарян [и др.], отв. ред. В. С. Мхитарян. – М.: Юрайт, 2016. – 490 с.

5. Миркин, Б. Г. Введение в анализ данных учебник и практикум / Б. Г. Миркин. – М.: Издательство Юрайт, 2018. – 174 с. – (Серия: Авторский учебник). – ISBN 978-5-9916-5009-0.

6. Сидняев, Н. И. Теория планирования эксперимента и анализ статистических данных: учебник и практикум для бакалавриата и магистратуры / Н. И. Сидняев. – 2-е изд., перераб. и доп. – М.: Издательство Юрайт, 2018. – 495 с. – (Серия: Бакалавр и магистр. Академический курс). – ISBN 978-5-534-05070-7.

7. Кремер, Н. Ш. Теория вероятностей и математическая статистика в 2 ч. Часть 1. Теория вероятностей: учебник и практикум для академического бакалавриата / Н. Ш. Кремер. – 4-е изд., перераб. и доп. – М.: Издательство Юрайт, 2018. – 264 с. – (Серия: Бакалавр. Академический курс). – ISBN 978-5-534-01925-4.

8. Теория вероятностей и математическая статистика. Математические модели: учебник для академического бакалавриата /

В. Д. Мятлев, Л. А. Панченко, Г. Ю. Ризниченко, А. Т. Терехин. – 2-е изд., испр. и доп. – М.: Издательство Юрайт, 2018. – 321 с. – (Серия: Университеты России). – ISBN 978-5-534-01698-7.

9. Дубров, А. М. Многомерные статистические методы / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин – М.: Финансы и статистика, 1998. – 352 с.

10. 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.

11. Айвазян, С. А. Прикладная статистика. Основы эконометрики: учебник для вузов в 2 т. / Айвазян С. А., Мхитарян В. С. – М.: ЮНИТИ-ДАНА, 2001.

12.Яковлев, В. Б. СТАТИСТИКА. РАСЧЕТЫ В MICROSOFT EXCEL 2-е изд., испр. и доп. Учебное пособие для СПО М.: Издательство Юрайт, 2018. – 353 с. – ISBN: 978-5-534-02551-4

13. Информационные технологии в маркетинге: учебник и практикум для СПО / С. В. Карпова [и др.]; под общ. ред. С В. Карповой. – М.: Издательство Юрайт, 2018. – 367 с. – (Серия: Профессиональное образование). – ISBN 978-5-9916-9115-4.

Учебное издание

***Трусова Алла Юрьевна***

**АНАЛИЗ ДАННЫХ.**
**МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ**

*Учебное пособие*

Редакционно-издательская обработка А.С. Никитиной

Подписано в печать 28.12.2023. Формат 60x84 1/16.
Бумага офсетная. Печ. л. 5,75.
Тираж 27 экз. Заказ . Арт. – 41 (УП/Р2Д)2023.

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)
443086, САМАРА, МОСКОВСКОЕ ШОССЕ, 34.

---

Издательство Самарского университета.
443086, Самара, Московское шоссе, 34.