

ББК  
22.17  
Д 458

---

# **АНАЛИЗ МНОГОМЕРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ**

**С. В. ДРОНОВ**

---

С. В. ДРОНОВ

# АНАЛИЗ МНОГОМЕРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Монография

Москва Вологда  
«Инфра-Инженерия»  
2025

УДК 519.23  
ББК 22.172  
Д75

Рецензенты:

д. ф.-м. н., профессор, ведущий научный сотрудник  
(Алтайский государственный технический университет)

*Пышинограй Григорий Владимирович;*

д. ф.-м. н., профессор  
(Алтайский государственный университет)

*Родионов Евгений Дмитриевич*

**Дронов, С. В.**

**Д75** Анализ многомерных статистических данных : монография / С. В. Дронов. – Москва ; Вологда : Инфра-Инженерия, 2025. – 308 с. : ил., табл.  
ISBN 978-5-9729-2524-7

Подробно рассмотрен ряд методов обработки многомерных данных, способов сокращения размерности и некоторых продвинутых методик визуализации этих данных. При этом приводится математическое обоснование и неформальное объяснение механизма действия методов. В частности, изучаются проблемы исключения грубых ошибок наблюдения, кластерный анализ, методы главных компонент, экстремальной группировки признаков и корреляционных плеяд, а также многомерного шкалирования и анфолдинга. Обсуждаются подходы к переводу качественных данных в числовую форму. Уделено значительное внимание выявлению и оцениванию силы связи как между числовыми, так и качественными показателями. Кроме этого, в книгу включены материалы о post-hoc задачах кластерного анализа и метрической структуре семейства кластерных разбиений конечного множества.

Для тех исследователей, кто, имея математическую подготовку в объеме университетских курсов высшей математики, хочет не просто использовать многомерные статистические методы, а разобраться в механизме их действия. Может быть использовано как основа для чтения специальных курсов магистрантам и аспирантам направлений, использующих в своих предметных областях статистические методы.

УДК 519.23  
ББК 22.172

ISBN 978-5-9729-2524-7

© Дронов С. В., 2025

© Издательство «Инфра-Инженерия», 2025

© Оформление. Издательство «Инфра-Инженерия», 2025



# Оглавление

<b>1</b>	<b>Предварительные сведения</b>	<b>7</b>
1.1	Теория вероятностей и математическая статистика . . . . .	8
1.2	Нечисловые данные . . . . .	14
1.3	Нормальное распределение . . . . .	16
<b>2</b>	<b>Изучение связи числовых показателей</b>	<b>22</b>
2.1	Понятие статистической связи . . . . .	22
2.2	Проверка независимости . . . . .	24
2.3	Коэффициент парной корреляции Пирсона . . . . .	26
2.4	Корреляционная теория . . . . .	31
<b>3</b>	<b>Регрессионный анализ</b>	<b>42</b>
3.1	Постановка задачи . . . . .	42
3.2	Нормальное уравнение регрессии . . . . .	47
3.3	Задачи с ограничениями . . . . .	53
3.4	Оптимальный выбор матрицы плана . . . . .	55
3.5	Задача статистического прогноза . . . . .	58
<b>4</b>	<b>Оценка силы связи нечисловых показателей</b>	<b>62</b>
4.1	Вводные замечания . . . . .	62
4.2	Ранговые шкалы . . . . .	64
4.2.1	Экспертные оценки . . . . .	72
4.2.2	Согласованные группы экспертов . . . . .	74
4.3	Два категоризованных показателя . . . . .	78
4.3.1	Таблицы сопряженности общего вида . . . . .	78
4.3.2	Четырехпольные таблицы . . . . .	79
4.3.3	Коэффициент относительного риска . . . . .	83
4.4	Ступенчатая связь. Ледж-коэффициент . . . . .	86

<b>5</b>	<b>Дисперсионный анализ</b>	<b>92</b>
5.1	Вводные замечания . . . . .	92
5.1.1	Проверка гипотезы нормальности . . . . .	94
5.1.2	Однородность дисперсий . . . . .	97
5.2	Однофакторный анализ . . . . .	98
5.3	Несколько факторов . . . . .	101
5.4	Некоторые примеры . . . . .	103
<b>6</b>	<b>Группировка и цензурирование</b>	<b>109</b>
6.1	Группировка в одномерном случае . . . . .	110
6.2	Одномерное цензурирование . . . . .	113
6.3	Многомерная группировка . . . . .	115
6.4	О цензурировании многомерных данных . . . . .	117
<b>7</b>	<b>Алгоритмы кластерного анализа</b>	<b>119</b>
7.1	Постановка задачи . . . . .	119
7.2	Иерархические алгоритмы . . . . .	121
7.3	Оцифровка иерархий . . . . .	129
7.4	Коэффициент кластерных различий . . . . .	132
7.5	Два классических алгоритма . . . . .	136
7.5.1	Алгоритм $k$ -средних . . . . .	136
7.5.2	Алгоритм ФОРЕЛЬ . . . . .	137
<b>8</b>	<b>Дискриминантный анализ</b>	<b>139</b>
8.1	Постановка задачи для двух классов . . . . .	140
8.2	Линейное прогностическое правило . . . . .	144
8.3	Рекомендации для нелинейного случая . . . . .	148
8.4	Дискриминация на практике . . . . .	150
8.5	Более двух классов . . . . .	153
8.6	Проверка качества дискриминации . . . . .	155
<b>9</b>	<b>Post-hoc задача в кластерном анализе</b>	<b>158</b>
9.1	Вводные замечания . . . . .	158
9.2	Применение дисперсионного анализа . . . . .	159
9.3	Учет искажений кластерной структуры . . . . .	161
9.4	Кластерная переменная . . . . .	164
9.4.1	Ранжирование кластеров . . . . .	165
9.4.2	Применение анализа соответствий . . . . .	167



9.4.3	Специализированные методы: регрессия . . . . .	169
<b>10</b>	<b>Методы анализа соответствий</b>	<b>170</b>
10.1	Случай двух показателей . . . . .	170
10.1.1	Расстояние хи-квадрат . . . . .	172
10.1.2	Матрицы рассеивания . . . . .	175
10.1.3	Оцифровка для задач кластерного анализа . . . . .	179
10.2	Множественный анализ соответствий . . . . .	182
10.2.1	Бинарная матрица в роли таблицы сопряженности . . . . .	183
10.2.2	Максимальные корреляции . . . . .	185
10.3	О выборе размерности меток . . . . .	187
10.4	Пример с профессиями . . . . .	189
10.5	Случай смешанных данных . . . . .	193
<b>11</b>	<b>Снижение размерности</b>	<b>197</b>
11.1	Сущность задачи снижения размерности . . . . .	197
11.2	Метод главных компонент . . . . .	200
11.2.1	Формальная постановка задачи . . . . .	201
11.2.2	Вычисление главных компонент . . . . .	203
11.2.3	Численная иллюстрация метода . . . . .	207
11.3	Метод экстремальной группировки признаков . . . . .	209
11.3.1	Критерий квадратов . . . . .	210
11.3.2	Критерий модулей . . . . .	213
11.4	Метод корреляционных плеяд . . . . .	216
11.4.1	Построение плеяд . . . . .	216
11.4.2	Дерево зависимостей показателей . . . . .	216
<b>12</b>	<b>Факторный анализ</b>	<b>219</b>
12.1	Постановка задачи . . . . .	219
12.1.1	Интерпретация решения . . . . .	222
12.2	Математическая модель . . . . .	224
12.2.1	Подходы к решению основного уравнения ФА . . . . .	227
12.2.2	Центроидный метод . . . . .	229
12.3	Алгоритм факторного анализа по шагам . . . . .	233
12.4	Вращение решений . . . . .	236
12.5	Оценивание значений латентных факторов . . . . .	238
12.5.1	Метод Бартлетта . . . . .	239
12.5.2	Метод Томсона . . . . .	240

12.6 Практика выявления латентных факторов . . . . .	242
<b>13 Многомерное шкалирование</b>	<b>247</b>
13.1 Вводные замечания . . . . .	248
13.2 Модель Торгерсона . . . . .	251
13.2.1 Стресс-критерий . . . . .	254
13.3 Алгоритм Торгерсона . . . . .	255
13.4 О шкалировании индивидуальных различий . . . . .	258
13.5 Многомерный анфолдинг с одной или двумя целями . . . .	261
<b>14 О геометрии семейства кластерных разбиений</b>	<b>274</b>
14.1 Кластерная метрика на решетке разбиений . . . . .	274
14.2 Ближайшее разбиение . . . . .	279
14.3 Удаленные разбиения. Коллигативный коэффициент . . . .	281
14.4 Согласованность метрики со структурой решетки . . . . .	286
14.5 Геометрия отрезка в семействе $\Xi(X)$ . . . . .	290
<b>Приложение: некоторые важные распределения</b>	<b>297</b>
<b>Библиографический список</b>	<b>300</b>

# Предисловие

Эта книга написана для тех, кто хочет не просто применить методы статистической обработки многомерных данных, но и разобраться в том, почему и как эти методы работают. Я много лет преподаю в университете вероятностные и статистические курсы студентам разных специальностей и постоянно отмечаю, что стремление разобраться в сути предмета у большинства студентов с годами все быстрее уменьшается. Это, видимо, отчасти связано с тем, что сегодня имеется достаточно большое количество разнообразных библиотек алгоритмов и стандартных программ работы с данными, легко доступных через интернет. Более того, очень популярный нынче предмет «Машинное обучение» лозунг «Вы только намекайте, что вам нужно, а соответствующий алгоритм мы подберем автоматически» сделал своим главным девизом. Поэтому многие склонны рассматривать любой курс, посвященный методам обработки данных как большой справочник с приложенным к нему автоматическим поиском. А, поскольку пытаться изучать справочник – заведомо бессмысленная затея, то и отношение к моим предметам возникает соответственное.

Тем не менее, сколько бы самых замечательных методов обработки данных не было бы создано, практик неизбежно столкнется с ситуацией, которая адекватно не обрабатывается ни одним из этих методов. Я неоднократно убеждался в этом, помогая специалистам различных конкретных предметных областей в анализе их данных. Чаще всего это были специалисты-медики, которые оказались плотно привязаны к структурированию и обработке своих данных в силу всеобщего принятия концепции доказательной медицины (нужно не просто предложить, например, новую методику лечения, но и убедительно показать, что она работает лучше существующих). Единственный шанс не спастись в ситуации, когда классические методики отказываются работать, это предложить собственную модификацию метода, которая уже сработает на конкретном



практическом примере. Конечно же, такую модификацию вы сможете предложить только тогда, когда сумеете полностью разобраться в сути проблемы и возможных альтернативных подходах к ее решению.

Целью, которую я видел перед собой, занимаясь написанием настоящей книги, является именно объяснение причин и механизмов действия основных статистических методов в надежде, что это поможет внимательному читателю сделать что-то подобное в своей конкретной задаче.

При этом, поскольку предмет наш, разумеется, математический, для адекватных объяснений приходится привлекать некоторые математические конструкции и факты. Поэтому лучше, если у читателя имеется соответствующая математическая база. В самой большой степени нам придется использовать университетские курсы теории вероятностей и математической статистики. Коротко важнейшие для нашей тематики сведения из них я попытался напомнить в первой главе книги.

Не ждите здесь описания каких-то новейших продвинутых методов. Теория анализа данных слишком быстро развивается, и попытка успеть за ней в печатном тексте не может вызвать ничего, кроме саркастической усмешки. Посмотрите, например, как выглядят в старых книгах и справочниках по статистике обзоры программного обеспечения! В предлагаемой вам книге рассмотрен лишь набор проверенных временем методов обработки данных и предпринята попытка сделать их анализ в заявленном ключе.

Тем не менее, я счел возможным включить в текст некоторые собственные результаты, в частности, посвященные обработке нечисловых данных, ступенчатых зависимостей, а также главу, посвященную геометрии семейств кластерных разбиений, которые ранее не публиковались нигде, кроме специализированных журналов.

Можно, видимо, эту книгу использовать и как справочник, снабженный примерами, но ее ценность от этого значительно снизится. Впрочем, о том, насколько удачной и ценной получилась книга, судить не мне, а вам – ее читателям.