ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ

РОССИЙСКАЯ АКАДЕМИЯ НАУК

Институт аналитического приборостроения
Учебно-научный центр
“Приборы и средства автоматизации для научных исследований”

А. Л. Буляница, В. Е. Курочкин, И. С. Кноп

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКОЛОГИЧЕСКОЙ ИНФОРМАЦИИ:
ДИСКРИМИНАНТНЫЙ, КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Учебное пособие

Санкт-Петербург
2005

УДК 519.2(075)
ББК 22.172
Б90

Буляница А. Л., Курочкин В. Е., Кноп И. С.

Б90 Методы статистической обработки экологической информации: дискриминантный, корреляционный и регрессионный анализ: Учеб. пособие /СПбГУАП. СПб; РАН. Ин-т аналитич. приб-я, 2005. 48 с.

В учебном пособии рассмотрены классические методы статистической обработки информации – дискриминантный, корреляционный, факторный и регрессионный анализы и их современные модификации. Даются рекомендации по их применению при решении различных задач обработки экспериментальных данных.

Приложение содержит уникальную медико-экологическую статистическую информацию, которая в большой степени применима к регионам Северо-Запада России.

Предназначено для студентов старших курсов и является базовым пособием при изучении дисциплины “Физический эксперимент и обработка его результатов”. Может быть полезно специалистам в областях обработки информации, организации здравоохранения, промышленной и медицинской экологии.

Рецензенты:
кафедра физической оптики и спектроскопии СПбГУИТМО;
кандидат технических наук, доцент *Л. В. Новиков*

Утверждено
редакционно-издательским советом университета
в качестве учебного пособия

ПРЕДИСЛОВИЕ

В учебном пособии описываются методы решения широкого круга задач прикладной статистики, например исключение выбросов и оценивание однородности последовательности измерений, выявление характера связи между различными группами данных, оценивание и компенсация детерминированных составляющих сигнала и т.д. Применение этих методов иллюстрируется с помощью уникальной статистики смертности от сердечно-сосудистых заболеваний в г. Архангельске в 1983 г., собранной врачом Белой Н.С. Сама статистика, представленная в Приложении, содержит информацию о температуре, силе ветра, перепадах давления, показателях магнитной и солнечной активности, влияние которых на уровень смертности может представлять интерес, прежде всего, для медиков и экологов.

Использование учебного пособия предполагает наличие у студентов лишь необходимых базовых знаний, не выходящих за рамки учебных курсов "Теория вероятностей и статистика", "Обработка результатов эксперимента" или аналогичных курсов. Библиографический список содержит необходимый минимум ссылок на работы, большинство из которых считаются в настоящее время классическими.

Авторы выражают глубокую благодарность доктору физико-математических наук, профессору кафедры "Высшая математика" Санкт-Петербургского государственного политехнического университета Георгию Леонидовичу Шевлякову, совместная работа с которым способствовала написанию пособия, и кандидату технических наук, доценту кафедры "Промышленная и экологическая безопасность" Санкт-Петербургского государственного университета аэрокосмического приборостроения Вадиму Петровичу Котову, чья помощь, выразившаяся во внимательном прочтении материала пособия и высказанных замечаниях и предложениях, сделанных в ходе его обсуждения, была весьма существенной.

1. ПЕРВИЧНЫЙ АНАЛИЗ ИСХОДНЫХ ДАННЫХ. УЧЕТ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ

В качестве исходных данных использованы величины, имеющие различную физическую природу и, следовательно, различные единицы и систему измерения. Рассмотрим следующие типы величин:

Метрические, оцененные количественно с помощью физических единиц измерения (температура, перепад давления, скорость ветра, площадь солнечных пятен). Единицами измерения числа случаев смерти и числа вспышек на Солнце являются "штуки" или "разы", т. е. безразмерные величины. Однако эти величины также оценены количественно и потому могут быть отнесены к этому же типу метрических величин **(А)**.

Балльные величины представляют собой также монотонную количественную характеристику, но измеряются в некоторых условных единицах. К таковым признакам относятся *ak*-индекс возмущенности магнитного поля Земли, число Вольфа, представляющее собой достаточно сложную зависимость от числа вспышек на Солнце как имеющихся, так и вновь образуемых, и интегральный показатель солнечной активности **(Б)**.

Кластерные величины представляют собой набор номеров каких-либо классов. Их отличие от балльных состоит в невозможности в ряде случаев связать номер класса и какую-либо однозначную монотонную количественную характеристику. Например, каждого человека можно отнести к одному из классов: *0* – практически непьющий; *1* – слабо пьющий; *2* – умеренно употребляющий алкоголь; *3* – сильно пьющий и *4* – хронический алкоголик. В этой ситуации возрастание номера класса, с некоторыми оговорками, характеризует количественное увеличение дозы потребляемого алкоголя. В то же время введение кластерного признака "социальная группа" по типу *1* – рабочий, *2* – служащий и *3* – неработающий уже не позволяет предложить какую-либо достоверную количественную характеристику, обусловливающую такую нумерацию классов **(В)**.

Индикаторные – величины, значение которых предполагает разрешение простейшей альтернативы в терминах "наличие–отсутствие", "да–нет" и т. д. Очевидно, что такими величинами являются наличие

гипертонической болезни, инфаркта, избыточного питания у пациента (или их отсутствие). Однако и другие величины можно отнести к индикаторным, например пол пациента (*1* – мужской, *0* – женский). За исключением крайне редких случаев индикаторные величины принимают значения 0 или 1 (**Г**).

В представленных исходных данных величины типа **В** и **Г** отсутствуют.

Традиционная методика заполнения отсутствующих значений базируется на построении интерполяционного полинома, в простейшем случае полинома первой степени. Так, при наличии обоих соседних по времени измерений в качестве оценки отсутствующего значения может быть принято $x_i = (x_{i-1} + x_{i+1})/2$. В ряде случаев можно использовать большее число измерений и строить интерполяционный полином более высокой степени, либо использовать для построения интерполяционного полинома большее число точек, а коэффициенты полинома (параметр сдвига и параметр положения) оценивать по методу наименьших квадратов. Подобные методики можно использовать и при отсутствии нескольких измерений подряд. Очевидно, что достоверность такой интерполяции будет тем ниже, чем больше имеется пропущенных значений и чем меньше измерений используется для построения интерполяционного полинома.

2. ДИСКРИМИНАНТНЫЙ (КЛАСТЕРНЫЙ) АНАЛИЗ

В Приложении представлены гелио- и метеофакторы в дни, когда происходили (или не происходили) случаи внезапной смерти от сердечно-сосудистых патологий. Важнейший вопрос: имеются ли различия в показателях гелио- (гео-, метео-) признаков в эти дни. Объединим все измерения, выполненные в те дни, в которые не наблюдались случаи внезапной смерти в группу (класс), называемый в дальнейшем "Класс 0", а измерения, осуществленные в дни, когда регистрировались случаи внезапной смерти в другую группу, называемую далее "Класс 1". Рассмотрим следующую задачу: оценим возможность разделения (дискриминации) классов на основе измерений признаков. Для решения указанной задачи следует оценить значимость различий между средними значениями признаков $t = (M_1 - M_0)/\delta$, где M_i – выборочные средние значений признаков для Классов 1 и 0 соответственно; δ – взвешенное стандартное отклонение средних, рассчитанное по формуле

$$\delta = \sqrt{\frac{(N_0 - 1)\sigma_0^2 + (N_1 - 1)\sigma_1^2}{(N_0 + N_1)(N_0 + N_1 - 2)}}.$$

Здесь N_i – число измерений в Классах 0 или 1, а σ_i^2 – дисперсия измерений заданного признака в рамках Классов 0 или 1 соответственно.

Известно, что если сам признак представляет собой нормально распределенную случайную величину, то величина t удовлетворяет распределению Стьюдента с $N_0 + N_1 - 2$ степенями свободы. Эта величина табулирована [1], т. е. всегда можно определить вероятность того, что при заданном значении t измерения из обоих классов принадлежат одной выборке. Очевидно, что чем меньше указанная вероятность P, тем сильнее дискриминация классов. Как правило, говорить о достоверном разделении классов можно, если $P < 0{,}01$ (иногда $P < 0{,}05$).

Даже если исходная случайная величина распределена по закону, далекому от нормального, распределение выборочного среднего (при достаточно больших N_i) весьма близко к нормальному. Поэтому количественная ошибка от использования критерия Стьюдента в этом случае будет невелика.

Для альтернативной оценки дискриминации классов можно вместо выборочного среднего использовать медиану – средний член (или полусумму средних членов) вариационного ряда. Разумеется, в отличие от выборочного среднего, распределение медианы будет отлично от нормального. Известно, распределение любой порядковой статистики, в том числе и медианы, удовлетворяет так называемому "бета-распределению" [1,2]. В этом случае связать величину t, построенную на основе различия медиан выборок, с вероятностью несколько сложнее. Соответствующие методики в рамках так называемых "непараметрических статистик" имеются. Однако в данном пособии предложим упрощенную трактовку использования медианного критерия дискриминации: если разность между медианами соизмерима с разностью между соответствующими выборочными средними, то эффект возможной дискриминации классов (при больших t) *не вызван* одиночными статистическими выбросами измерений, т. е. действительно отражает статистически обоснованное различие между измерениями в рамках различных классов. Оценка дискриминации классов по измерениям признаков приведена в табл. 1.

Таблица 1

Дискриминация Классов 0 и 1

Признак	Число измерений	Среднее	Медиана	δ	t	P
t	237	1,66	2,3			
	128	1,82	3,4	0,79	0,20	>0,20
p	237	5,59	4,5			
	128	13,49	11,8	0,52	15,1	<0,001
vcp	237	2,96	2,9			
	128	3,03	2,9	0,09	0,78	>0,20
ak	237	29,6	18,0			
	128	29,8	18,5	2,4	0,06	>0,20
s	237	642,6	590			
	128	613,7	620	35,7	–0,81	>0,20
w	237	102,2	101			
	128	101,7	106	3,3	–0,15	>0,20
ps	237	7,29	6			
	128	7,02	5	0,50	–0,54	>0,20
as	237	1,08	1			
	128	1,19	1	0,04	2,66	<0,01

Примечание: В Класс 0 входит 237 измерений, в Класс 1 – 128.

Анализ данных таблицы позволяет сделать выводы:

1. Имеется четкая дискриминация Классов 0 и 1 по признаку p – перепад давления в течение суток. Уровень значимости очень высок – 0,001.

2. Данные о дискриминации Классов по признаку (*as*-интегральному показателю) солнечной активности сомнительны. Уровень значимости достаточно высок – 0,01. Вместе с тем медианы измерений полностью совпадают, что может свидетельствовать о "ложной" дискриминации, вызванной редкими выбросами.

3. В некоторых случаях выбросы приводят не к дискриминации, а к "антидискриминации" классов. Так, в случаях с признаками t, w и *ps* разность между медианами существенно больше, чем между выборочными средними. Вместе с тем при формальной замене разности выборочных средних на разности медиан получатся значения 1,39; 1,52 и 2,00, что при указанных объемах выборки не может свидетельствовать об уровне значимости выше, чем $P = 0,05$.

Таким образом, следует признать достоверным разделение классов только по признаку p.

Сама дискриминация классов может выполняться по следующему одно- или двухпороговому алгоритму.

Однопороговый алгоритм формулируется следующим образом: если значение признака меньше (или равно) пороговому, измерение относится к Классу 0, если больше – к Классу 1. В этом случае возможны ошибки дискриминации первого и второго рода: ошибка первого рода – ошибочное отнесение измерения к Классу 1 при превышении порога, несмотря на то, что это измерение принадлежало Классу 0, и ошибка второго рода – отнесение измерения из Класса 1 к Классу 0 на основании того, что оно не превосходит порога. Очевидно, что критерием выбора порога должна быть минимизация суммарной вероятности ошибок первого и второго рода или какая-либо взвешенная комбинация этих ошибок.

Двухпороговый алгоритм дискриминации предполагает:

– отнесение измерения к Классу 0 при недостижении первого порога (меньшего);

– отнесение измерения к Классу 1 при превышении второго (большего) порога;

– отнесение измерения, заключенного между пороговыми значениями, к области нечувствительности (т. е. не допускающего отнесения к какому-либо из Классов). В этом случае, помимо минимизации ошибок первого и второго рода, требуется и уменьшение доли неклассифицированных измерений.

Рассмотрим методику подбора порога по признаку p. Из Приложения следует, что к Классу 0 должны быть отнесены $N_0 = 237$ измерений, к Классу $1 - N_1 = 128$. В табл. 2а–ж рассмотрены схемы выбора различных порогов.

Таблица 2а

Дискриминация измерений по признаку p при пороге 6,0

p	Класс 0	Класс 1	Сумма
≤6	152	26	178
>6	85	102	187
Сумма	237	128	365

Таблица 2б

Дискриминация измерений по признаку p при пороге 7,0

p	Класс 0	Класс 1	Сумма
≤7	164	27	191
>7	73	101	174
Сумма	237	128	365

Таблица 2в

Дискриминация измерений по признаку p при пороге 7,5

p	Класс 0	Класс 1	Сумма
≤7,5	175	31	206
>7,5	62	97	159
Сумма	237	128	365

Таблица 2г

Дискриминация измерений по признаку p при пороге 8,0

p	Класс 0	Класс 1	Сумма
≤8	184	36	220
>8	53	92	145
Сумма	237	128	365

Таблица 2д

Дискриминация измерений по признаку p при пороге 8,5

p	Класс 0	Класс 1	Сумма
≤8,5	190	44	234
>8,5	47	84	131
Сумма	237	128	365

Таблица 2е

Дискриминация измерений по признаку *p* при пороге 9,0

p	Класс 0	Класс 1	Сумма
≤9	194	50	244
>9	43	78	121
Сумма	237	128	365

Таблица 2ж

Дискриминация измерений по признаку *p* при пороге 10,0

p	Класс 0	Класс 1	Сумма
≤10	203	55	258
>10	34	73	107
Сумма	237	128	365

Рассмотрим две возможные целевые функции (показателя качества): 1) суммарная относительная ошибка дискриминации $P_1 = (\alpha + \beta)/(N_0 + N_1)$ и 2) сумма относительных ошибок дискриминации (средняя по классам ошибка дискриминации) в форме $P_2 = \alpha/N_0 + \beta/N_1$. Критерием будет, соответственно, $P_{1,2} \to \min$. Здесь α и β – ошибки первого и второго рода.

Основываясь на данных табл. 2а–ж, в табл. 3 сведены результаты дискриминации Классов 0 и 1 по измерениям признака *p* для различных порогов.

Таблица 3

Выбор порога при дискриминации классов измерений по признаку *p*

Порог	α	β	$\alpha + \beta$	P_1	P_2
6	85	26	111	0,304	0,562
7	73	27	100	0,274	0,519
7,5	62	31	93	0,255	0,504
8	53	36	89	0,244	0,505
8,5	47	44	91	0,249	0,542
9	43	50	93	0,255	0,572
10	34	55	89	0,244	0,573

Анализ данных табл. 3 свидетельствует:

– минимум по критерию 1 реализуется при пороговых значениях 8 и 10;

– минимум по критерию 2 реализуется при пороговых значениях 7,5–8;

– выбор какого-либо иного критерия оптимального выбора порога (например, учитывающего различную значимость ошибок первого и второго рода) может привести к какому-либо иному пороговому значению.

Двухпороговая схема, предполагающая выбор двух пороговых значений, ограничивающих область неопределенности (нечувствительности), также допускает различные критерии оптимизации. Ограничимся рассмотрением критерия в форме $P_3 = \alpha + \beta + \gamma \to \min$, где γ – число измерений, входящих в зону нечувствительности. Методика подбора пороговых значений иллюстрируется данными табл. 4а–в.

Таблица 4а

Дискриминация классов по признаку *p* при порогах 6 и 9 мм рт. ст.

p	Класс 0	Класс 1	Сумма
≤ 6	152	26	178
$6 < p < 9$	42	24	66
≥ 9	43	78	121
Всего	237	128	365

Здесь $\alpha = 43$, $\beta = 26$ и $\gamma = 66$, т. е. ошибочно классифицированных или неклассифицированных измерений – 135 (или 37,0%).

Таблица 4б

Дискриминация классов по признаку *p* при порогах 7 и 10 мм рт. ст.

p	Класс 0	Класс 1	Сумма
≤ 7	164	27	191
$7 < p < 10$	39	28	67
≥ 10	34	73	107
Всего	237	128	365

Здесь $\alpha = 34$, $\beta = 27$ и $\gamma = 67$, т. е. ошибочно классифицированных или неклассифицированных измерений – 128 (или 35,1%).

Таблица 4в

Дискриминация классов по признаку *p* при порогах 6 и 8 мм рт. ст.

p	Класс 0	Класс 1	Сумма
≤ 6	152	26	178
$6 < p < 8$	32	10	42
≥ 8	53	92	145
Всего	237	128	365

Здесь $\alpha = 53$, $\beta = 26$ и $\gamma = 42$, т. е. ошибочно классифицированных или неклассифицированных измерений – 121 (или 33,2%).

Если значимость ошибок соизмерима со значимостью неклассифицированных измерений, наилучший подбор порогов из трех предложенных вариантов будет 6 и 8; если значимость ошибок первого и второго рода существенно выше, чем отсутствие классификации, наилучший подбор порогов из вышеперечисленных – 7 и 10.

Пороговые алгоритмы дискриминации просты, легко формализуются и легко встраиваются как этапы более сложных алгоритмов идентификации или дискриминации.

3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Как известно, корреляционная связь (КС) является "односторонним" индикатором причинно-следственной связи (ПСС) по механизму: наличие ПСС приводит к КС, обратное следствие может отсутствовать (наличие КС не всегда говорит о ПСС).

При этом наличие причинно-следственной связи не следует понимать однозначно, как "изменение одного из признаков влечет за собой соответствующее изменение другого". Возможно также наличие *общей причины* подобной динамики признаков.

3.1. Выборочный коэффициент корреляции

Значения выборочного коэффициента корреляции между X_1 и X_2, определяются в соответствии с формулой

$$r = \sum_{i=1}^{n}(X_i^1 - < X^1 >)(X_i^2 - < X^2 >) / \sqrt{\sum_{i=1}^{n}(X_i^1 - < X^1)^2 \sum_{i=1}^{n}(X_i^2 - < X^2 >)^2}. \quad (1)$$

Статистически значимыми следует признать значения $|r| > 0{,}25$. Величина предложенного порога 0,25 базируется на следующих положениях:

1. Выборочный коэффициент корреляции r есть лишь оценка генерального коэффициента корреляции ρ, рассчитанная по конечной выборке объема n (КС должна характеризоваться величиной генерального коэффициента корреляции).

2. ρ – случайная величина, возможный размах которой определяется: оценкой r; объемом выборки n; доверительной вероятностью.

Иными словами, с достаточно большой вероятностью (не менее 99%) можно утверждать, что при $|r| > 0{,}25$ и при n – достаточно больших (более 20), диапазон возможной величины генерального коэффициента корреляции не будет включать ноль, т. е. будет КС.

3.2. Робастные модификации выборочного коэффициента корреляции

Выражение (1) свидетельствует, что оценка r не является робастной (устойчивой к выбросам), поскольку включает все, в том числе и

резко отстоящие измерения. Таким образом, КС может явиться следствием одиночных (или редких) выбросов (погодных, климатических, геомагнитных аномалий). Вследствие этого, рекомендуется дополнить исследование некоторыми робастными аналогами выборочного коэффициента корреляции.

К этим аналогам следует отнести: $R_{\text{м}}$ – медианный коэффициент корреляции, предложенный профессором Санкт-Петербургского государственного политехнического университета Г. Л. Шевляковым; Sp – ранговый коэффициент корреляции Спирмена; Kn – знаковый коэффициент корреляции Кендалла.

Медианный коэффициент корреляции вычисляется по следующей процедуре:

1. Признаки X_1 и X_2 делятся на величины отклонения Хемпеля (т. е. нормируются на отклонение Хемпеля, равное единице).

2. На основе полученных нормированных признаков X_1^*, X_2^* строятся величины $X^+ = X_1^* + X_2^*, X^- = X_1^* - X_2^*$.

3. Для полученных величин рассчитываются отклонения Хемпеля H^+, H^-.

4. Коэффициент корреляции вычисляется как $R_{\text{м}} = \dfrac{(H^+)^2 - (H^-)^2}{(H^+)^2 + (H^-)^2}$.

Ранговый и знаковые коэффициенты корреляции вычисляются, как указано в [1].

При расчете рангового и знакового коэффициента корреляции требуется предварительное ранжирование признаков. Например, один из признаков упорядочивается по возрастанию (т. е. ранг 1 присваивается наименьшему, ранг n – наибольшему из значений признака, либо наоборот). Соответственно определяются ранги измерений второго признака – $r_1, r_2, \ldots, r_n$, и на основе этих данных производится вычисление.

Ранговый коэффициент корреляции Спирмена:

$$\text{Sp} = 1 - \frac{6S}{n^3 - n}, \quad S = \sum_{i=1}^{n} (r_i - i)^2. \tag{2}$$

Знаковый коэффициент корреляции Кендалла:

$$\text{Kn} = \frac{2S}{n^2 - n}, \quad S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{sign}(r_j - i). \tag{3}$$

Заметим, что Kn и S так же, как и выборочный коэффициент корреляции, изменяются в пределах от -1 до $+1$. В случае полной корреляции $r_i = i$, $i = 1, 2, \ldots, n$ (т. е. все измерения обоих признаков упорядоче-

ны одинаково). В этом случае, очевидно, сумма S, входящая в выражение (2) для Sp, равна нулю, так как содержит только нулевые слагаемые.

Подробнее рассмотрим противоположный случай: он описывается соответствием рангов $r_1 = n, r_2 = n-1, \ldots, r_n = 1$, т. е. возрастание измерений одного признака сопровождается убыванием соответствующего измерения другого признака. Проще говоря, $r_i = n+1-i,\ i = 1,2,\ \ldots, n$.

Тогда, использовав выражения для конечных сумм (см. [3]), можно доказать, что $\mathrm{Sp} = -1$.

Заметим, во-первых, что при замене отклонения Хемпеля на стандартное отклонение, получим выборочный коэффициент корреляции r. Во-вторых, медианный коэффициент корреляции в принципе не может быть рассчитан, если для какого-либо из двух признаков отклонение Хемпеля будет ноль. В рассматриваемом случае таким признаком будет *as*.

Проиллюстрируем простым модельным примером процедуры вычисления коэффициентов корреляции, прежде всего, менее распространенные (медианный, ранговый и знаковый). Исходные данные представлены в табл. 5а.

Таблица 5а

Исходные данные для модельного примера

№ п/п	Измерения		Ранги	
	X_1	X_2	r_i	r_j
1	100	8	4	3
2	90	6	5	4
3	110	5	3	5
4	150	11	1	1
5	120	9	2	2

Выборочный коэффициент корреляции между признаками X_1 и X_2, вычисленный по (1), дает оценку $+0{,}791$, т. е. наблюдается достаточно сильная положительная корреляция.

Упорядочив ранги первого признака, получим ранговую связь вида:

$$i:\ 1\ \ 2\ \ 3\ \ 4\ \ 5$$
$$r_i:\ 1\ \ 2\ \ 5\ \ 3\ \ 4.$$

Выражение для S (2) приводит к сумме $S = 0+0+4+1+1 = 6$. Так как $n = 5$, получим $\mathrm{Sp} = 1-36/120 = 0{,}700$. Расчет суммы S по (3) даст: $i = 1-1+1+1+1 = 4$, $i = 2-1+1+1 = 3$, $i = 3-0+1 = 1$, $i = 4-0$. Таким образом, $S = 8$ и $\mathrm{Kn} = 16/20 = 0{,}800$.

Подробнее проиллюстрируем процедуру вычисления медианного коэффициента корреляции. Как известно, медиана – средний (или полусумма двух средних) член вариационного ряда. В нашем случае, так как $n = 5$ – это измерение, имеющее ранг 3 (т. е. 110 и 8, соответственно). Для поиска отклонения Хемпеля нужно найти абсолютные отклонения каждого измерения от медианы, и далее медиану этих величин. Отклонение Хемпеля первого признака H_1 будет 10, для второго – $H_2 = 2$. Делением на $H_{1,2}$ можно нормализовать эти признаки, приведя к единичному отклонению Хемпеля. Это можно сделать, если отклонение Хемпеля отлично от нуля. В табл. 5б приведены значения нормализованных признаков и их комбинации.

Таблица 5б

Нормализованные значения признаков

№ п/п	X_1^*	X_2^*	$X_1^* + X_2^*$	$X_1^* - X_2^*$
1	10	4	**14**	6
2	9	3	12	6
3	11	2,5	13,5	8,5
4	15	5,5	20,5	9,5
5	12	4,5	16,5	**7,5**

В табл. 5б выделены медианы признаков $X_1^* \pm X_2^*$. Соответствующие величины отклонений Хемпеля будут $H^+ = 2{,}0$ и $H^- = 1{,}5$. Далее (согласно п. 4 алгоритма) $R_м = (4-2{,}25)/(4+2{,}25) = 0{,}280$.

Видно, что использование медиан "сглаживает" в данном модельном примере корреляции. Однако и значение 0,28 подтверждает наличие корреляционных связей между признаками. Само "сглаживание" может быть вызвано тем, что медианы не учитывают большие отклонения от среднего экстремальных измерений (наибольшей и наименьшей порядковых статистик), которые могут вносить основной вклад, прежде всего, в выборочный коэффициент корреляции.

3.3. Выявление и интерпретация значимых корреляционных связей

В табл. 6 приведены близкие к значимым коэффициенты корреляции между рассматриваемыми в Приложении признаками и их возможная интерпретация.

В принципе выявление причинно-следственных связей на основе корреляционных является достаточно сложной задачей. Так, некоторые из сделанных выводов можно считать достаточно тривиальными:

Таблица 6

Корреляционный анализ признаков и их интерпретация

X_1	X_2	r	$R_м$	Sp	Kn	Интерпретация результата
t	p	–0,171	–0,273	–0,183	–0,123	В более холодное время года большие перепады давления (погода менее стабильна)
t	s	0,214	0,294	0,257	0,171	В летнее время Солнце более активно
t	w	0,345	0,466	0,348	0,233	Аналогично предыдущему
t	ps	0,267	0,304	0,323	0,215	То же
p	vcp	0,283	0,210	0,210	0,139	При больших перепадах давления, большая сила ветра
s	w	0,700	0,793	0,759	0,569	Показатели солнечной активности сильно коррелируют между собой
s	ps	0,597	0,737	0,623	0,448	Аналогично предыдущему
s	as	0,475	–	0,173	0,113	То же
w	ps	0,579	0,664	0,624	0,444	–"–
w	as	0,454	–	0,179	0,116	–"–
ps	as	0,484	–	0,299	0,217	–"–

– показатели s и ps непосредственно связаны с активностью Солнца (с различными параметрами, характеризующими активность Солнца) и по причине общего происхождения являются зависимыми друг от друга;

– w непосредственно связано с числом вспышек Солнца и солнечными пятнами;

– as представляет собой некоторую взвешенную комбинацию всех вышеуказанных показателей солнечной активности;

– связь между перепадом давления и средней скоростью ветра достаточно тривиальна;

– большая устойчивость циклонов и антициклонов в летнее время может иметь какое-либо метеорологическое объяснение (возможно, только применительно к г. Архангельску или аналогичным регионам).

Интерес представляет отсутствие значимых корреляций между солнечной активностью и возмущенностью магнитного поля Земли. Однако, во-первых, возмущение Солнца может иметь отклик в виде возмущения магнитного поля Земли, значительно задержанный по времени (на несколько дней), во-вторых, активность Солнца в рассматриваемый период не слишком велика, а Архангельск может являться регионом с малым возмущением магнитного поля Земли.

Заметим, что в том случае, когда признак X_1 (X_2) принимает малое число значений (например, всего два, как as) ранговый и знаковый

коэффициенты корреляции могут давать сильно уменьшенные по абсолютной величине значения, так как обладают “сглаживающим” эффектом (реагируют на большие отклонения от среднего так же, как и на малые).

4. ФАКТОРНЫЙ АНАЛИЗ

С помощью факторного анализа возможно оценить взаимное влияние различных признаков друг на друга. Процедура факторного анализа связана с анализом собственных чисел и исследование структуры собственных векторов корреляционной матрицы. В принципе решение полной проблемы собственных чисел хорошо известно. В нашем случае матрица симметрична и собственные числа будут вещественны. Методы поиска собственных чисел различны (например, сингулярное разложение Лоусона и Хенсона [4] или процедура, описанная в работе Агно [5] и т.д.). Основная вычислительная трудность может быть связана с плохой обусловленностью корреляционной матрицы в случае, когда размерность матрицы велика (много исследуемых признаков) и имеются значительные корреляции между некоторыми из них. Очевидно, что в противном случае малых корреляций матрица будет близка к единичной, собственные числа близки к единице и, следовательно, степень обусловленности также близка к единице.

Корреляционная матрица между признаками, пронумерованными в следующем порядке – *t*, *p*, *vcp*, *ak*, *s*, *w*, *ps*, *as* представлена ниже (все величины, меньшие 0,10 по абсолютной величине, обнулены):

1	–0,17	0	–0,11	0,21	0,35	0,27	0,32
	1	0,28	0	–0,12	–0,10	–0,11	–0,10
		1	–0,13	–0,12	0	0	–0,12
			1	0,12	0,11	0	0
				1	0,70	0,60	0,48
					1	0,58	0,45
						1	0,48
							1

Расчет собственных чисел λ и собственных векторов, называемых факторами, иллюстрируется данными табл. 7 (длина собственных векторов равна единице).

Таблица 7

Собственные числа и нормированные собственные векторов корреляционной матрицы

№ п/п	Собственные числа λ	Координаты собственных векторов							
		v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
1	2,885	0,296	–0,146	–0,083	0,049	0,491	0,493	0,467	0,425
2	1,276	–0,091	–0,570	–0,723	0,330	–0,001	–0,115	–0,143	–0,021
3	1,127	–0,486	0,414	0,033	0,719	0,219	0,144	0,047	–0,066
4	0,770	0,632	–0,203	0,362	0,558	–0,185	0,096	–0,183	–0,199
5	0,707	–0,413	–0,603	0,437	–0,084	0,166	0,145	0,212	–0,417
6	0,505	0,228	0,264	–0,346	–0,190	0,249	0,333	–0,043	–0,740
7	0,419	0,118	0,085	–0,121	0,127	–0,303	–0,364	0,821	–0,222
8	0,270	-0,178	–0,004	–0,109	–0,032	–0,704	0,671	0,070	0,064

Математически задача факторного анализа практически совпадает с задачей приведения квадратичной формы к каноническому виду: решение полной проблемы собственных чисел и нахождение собственных направлений в пространстве признаков. Направляющим вектором этого пространства и будет соответствующий собственный вектор (или фактор).

По известному свойству собственных чисел их сумма должна быть равна следу матрицы (сумме диагональных элементов). В случае корреляционной матрицы – след совпадает с числом строк (столбцов). В нашем случае число строк 8, а сумма собственных чисел 7,959 (т. е. методы приближенного вычисления собственных чисел обеспечивают адекватную их оценку).

Степень обусловленности матрицы COND есть отношение абсолютных величин большего собственного числа к меньшему. В нашем случае степень обусловленности корреляционной матрицы не слишком велика (чуть более 10).

Смысл процедуры факторного анализа – выявление устойчивых комбинаций исходных признаков (факторов), относительная значимость которых определяется величиной собственного числа.

Например, наиболее "мощным" фактором, соответствующим первому собственному числу 2,885 будет $F_1 \approx 0{,}296t + 0{,}491s + 0{,}493w + +0{,}467ps + 0{,}435as$ (исключили слагаемые, вносящие малый вклад). Интерпретация этого фактора – наличие устойчивой комбинации, когда повышение температуры и всех показателей солнечной активности взаимно усиливают друг друга. Вторым по значимости фактором будет $F_2 \approx -0{,}570p - 0{,}723vcp$ (высокие перепады давления и повышение

силы ветра действуют совместно, усиливая друг друга). Третьим фактором будет $F_3 \approx -0,486t + 0,414p + 0,719ak$ (интерпретируется как взаимно усиливающее действие низкой температуры, больших перепадов давления и высокой интенсивности магнитного поля Земли).

5. РЕГРЕССИОННЫЙ АНАЛИЗ

Идея данного подхода – связать уровень смертности n (как зависимую переменную) с метеорологическими данными и характеристиками магнитной и солнечной активности (как независимыми переменными). Как правило, строят линейные уравнения вида

$$n = \alpha_1 t + \alpha_2 p + \alpha_3 vcp + \alpha_4 ak + \alpha_5 s + \alpha_6 w + \alpha_7 ps + \alpha_8 as.$$

5.1. Уравнения линейной регрессии

Поиск коэффициентов α_i осуществляется на основе процедуры метода наименьших порядков. Решение соответствующей системы линейных уравнений решается любыми традиционными методами (Крамера, Гаусса и т. д.). Возможно построение более сложных регрессионных уравнений, прежде всего, создавая новые параметры, например вида $t^{\alpha} p^{\beta} vcp^{\gamma}$ и т. п. Истолковать физический смысл новых искусственно созданных параметров будет затруднительно. Тем не менее можно будет строить новые уравнения линейной регрессии с большим числом переменных.

Примеры некоторых уравнений линейной регрессии приведены ниже:

$n = 0{,}454795 + 0{,}03994 p$,

$n = 0{,}454795 + 0{,}105515 as$,

$n = 0{,}454795 - 0{,}0000346 t$,

$n = 0{,}1026 + 0{,}0049 t + 0{,}04113 p$,

$n = 0{,}40142 - 0{,}00321 t - 0{,}01568 p + 0{,}03174 vcp + 0{,}00095 ak -$

$-0{,}00066 s + 0{,}00394 w + 0{,}01133 ps.$

По ряду причин прогностическая ценность этих уравнений линейной регрессии мала. Первая причина заключается в отсутствии дискриминации классов по большинству признаков (t, ak, vcp, s, w, ps). В этом случае практически любым значениям указанных независимых переменных будут соответствовать практически любые значения функции n. Для независимой переменной p дискриминация классов установлена с высоким уровнем значимости. Однако механизм дискриминации носит пороговый характер, а это в большинстве случаев не позволяет построить уравнение регрессии.

5.2. Доверительные интервалы для уравнений регрессии

Поскольку все признаки, входящие в уравнения, являются случайными величинами, то случайны и коэффициенты регрессионных уравнений, вариации которых связаны с вариациями исходных данных.

Примеры поиска доверительных областей уравнений линейной регрессии приведем ниже [1].

Ограничимся построением доверительных областей только применительно к первому из уравнений линейной регрессии $n = 0{,}454795 + 0{,}03994p$ или, округлив для наглядности коэффициенты уравнения, $n \approx 0{,}455 + 0{,}040\,p$, $p \in [0;50]$. Заметим, что при построении доверительных областей линейной регрессии не будет задаваться вопросом о физическом смысле функции n. Иными словами, если в доверительную область будут входить значения $n<0$, они не будут исключаться из рассмотрения. При этом с физической точки зрения очевидна бессмысленность отрицательного уровня смертности!

В данной работе не воспроизводим полное и достаточно подробное описание методики поиска линейных и гиперболических границ доверительной области регрессии, данное в [1].

Приведем необходимые промежуточные расчетные параметры и конечные результаты. При этом используем обозначения, идентичные предложенным в книге [1]. Так, параметр $\lambda \approx 0{,}90$. Выбираем доверительную вероятность $P = 90\%$. Случайные величины $u_{n-2} \approx 2{,}126$; $v_{n-2} \approx 1{,}894$ (так как $N = 365 >> 100$).

Границы доверительных областей и само регрессионное уравнение показаны на рис. 1. Квадратами обозначается сама прямая регрессии, треугольниками – гиперболические, кругами – линейные границы.

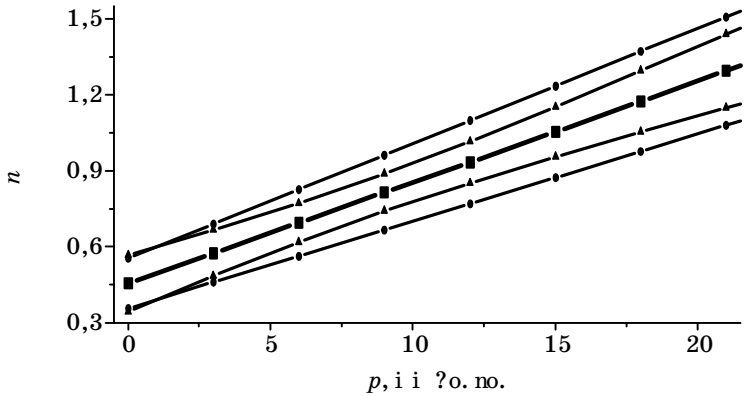

Рис. 1. Доверительные области для уравнения линейной регрессии с доверительной вероятностью 90%

Доверительные интервалы для определенных зон регрессии традиционно базируются на предположении о том, что распределение случайных величин (независимых переменных) удовлетворяет нормальному (гауссову) закону распределения. На практике достаточно качественного (примерного) соответствия.

5.3. Распределение значений признака *p*-перепада давления в течение суток

Разобьем диапазон изменения величины p на 7 интервалов: от 0 до 5, от 5 до 10 и т.д. Все измерения $p \geq 30$ объединены в один интервал. Результаты группирования представлены в табл. 8. Оптимальный выбор числа интервалов группирования m осуществляется в соответствии с правилом [6] $m \approx n^{0,4}$.

Таблица 8

Распределение значений признака *p*

Число измерений	Диапазон измерения p						
	[0;5)	[5;10)	[10;15)	[15;20)	[20;25)	[25;30)	≥ 30
n_i	143	112	53	33	11	5	8
np_i^0	71,2	95,3	82,5	46,4	17,2	4,0	0,7

Примечание: n_i – число измерений, попавших в интервал, p_i^0 – вероятность попадания в данный интервал, рассчитанная на основе выбранного закона распределения. В нашем случае предполагаем нормальный закон распределения с математическим ожиданием 8,36 и дисперсией 55,46.

Введя нормированную величину $y = \dfrac{p - M_p}{\sqrt{D_p}}$, можно рассчитать теоретическую вероятность $p_i^0 = \Phi(y_{\max}) - \Phi(y_{\min})$, где интеграл вероятности Ф – одна из базовых величин теории вероятностей, имеющаяся практически в любом учебнике или справочнике, например, в [1,7,8]. Принять или отвергнуть гипотезу о нормальном распределении случайной величины p можно в соответствии с каким-либо критерием согласия. Наиболее распространенным из таких критериев является так называемый χ^2-критерий (хи-квадрат критерий). Соответствующие пороговые значения, зависящие от уровня значимости и числа интервалов приведены в большинстве учебников и справочников по теории вероятностей; χ^2-критерий согласия требует вычисления величины $\eta = \sum_{i=1}^{m} \dfrac{(n_i - np_i^0)^2}{np_i^0}$ и сравнение ее с парамет-

рами соответствующего χ^2-распределения с $(m-1)$-степенью свободы (m – число интервалов разбиения). Гипотеза о выбранном характере распределения принимается с каким-либо уровнем значимости (доверительной вероятностью), если η не превзойдет порога, здесь $m = 7$, следовательно, число степеней свободы – 6. Величина η получилась более 168,3. В нашем случае, поскольку величина критерия равна 168,3, а пороговое значение есть 8,558 при доверительной вероятности 20% (10,645 при уровне значимости 10% , либо 16,812 при уровне значимости 1%), гипотеза о нормальном распределении измерений признака p должна быть отвергнута.

Тем самым метод вычисления границ зон регрессии как линейных, так и гиперболических не может быть применим к нашей конкретной задаче. Однако в том случае, когда распределение независимых переменных не столь сильно отличается от нормального, указанная выше методика полностью применима.

Построение как самого уравнения линейной регрессии, так и соответствующих границ зон регрессии также способствует выявлению статистической (и, возможно, причинно-следственной связи между переменными). В частности, если построенные границы области регрессии допускают попадание в эту область регрессионной прямой с нулевым тангенсом угла наклона, это может означать отсутствие статистической связи между независимой и зависимой переменной. В этом случае можно допустить отсутствие и причинно-следственной связи.

6. ПРИМЕР ПРИМЕНЕНИЯ МЕТОДОВ ОБРАБОТКИ ИНФОРМАЦИИ

Рассмотрим модельный пример статистического анализа данных, представленных в табл. 9.

Таблица 9

Исходные данные для модельного примера

№ п/п	Цвет	Масса, г	Длина хвоста, см	ЧСС, мин$^{-1}$	Реакция на лекарство
1	б	130	3,7	64	0
2	То же	170	4,1	53	2
3	–"–	210	4,3	56	2
4	–"–	140	3,8	70	3
5	–"–	165	3,9	73	1
6	–"–	180	4,0	58	1
7	ч	120	3,7	66	0
8	То же	145	3,9	75	1
9	–"–	170	4,1	71	1
10	–"–	190	4,4	80	3
11	–"–	180	4,4	74	3
12	–"–	175	4,3	68	1
13	–"–	155	3,9	83	2
14	–"–	230	4,7	72	2
15	–"–	150	3,6	81	1

Примечание: данные таблицы представляют собой только модельный пример; использовать данные в качестве медико-биологической информации не следует!

Проведем первичный анализ исходных данных (см. разд. 1). В табл. 9 представлены медико-биологические характеристики двух видов мышей: белые (б) и черные (ч). Остальные характеристики: масса (в граммах), длина хвоста (в см), частота сердечных сокращений (ЧСС) (ударов в мин) и степень реакции на лекарство, определяемая как (0 – отсутствие реакции, 1 – слабая, 2 – средняя, 3 – сильная, 4 – смертельная). Величины – масса, длина хвоста (далее –

просто длина) и ЧСС являются метрическими величинами (тип **А**), реакция на лекарства является кластерной величиной типа **В**. При этом номер Класса (от 0 до 4) возрастает с возрастанием степени чувствительности к лекарству. Однако поскольку эта степень может быть описана только на качественном уровне (т. е. соответствующим образом количественно измерить эту величину невозможно), то она не может быть отнесена к балльным величинам (или классу **Б**). Наконец, признак "цвет" формально следует отнести к типу **В**. Вместе с тем, так как он принимает только два значения: б и ч, его удобнее относить к индикаторному типу **Г** по схеме (0 – отсутствие черного цвета или белый цвет и 1 – наличие черного цвета). В дальнейшем будем полагать именно так: 0 – белый цвет, 1 – черный цвет.

Опишем возможные постановки задач и схемы их решения:

1. Анализ данных (каждого из признаков) с точки зрения исключения выбросов и выявления закона распределения измерений.

2. Дискриминантный анализ с целью разделения Классов 0 и 1 (т. е. белых и черных мышей, соответственно) на основании измерений признаков: масса, длина, ЧСС и реакция на лекарства.

3. Корреляционный анализ с целью выявления статистических и, возможно, причинно-следственных связей между признаками.

4. Факторный анализ также позволяет получить данные о взаимосвязи между признаками.

5. Регрессионный анализ позволяет связать одну из величин (в нашем случае – цвет или номер Классов 0–1) с измерениями признаков масса, длина, ЧСС и реакция на лекарства.

Этап 1. Анализ измерений признаков дан в разд. 1.

Выполним расчет выборочных средних и стандартных отклонений. Для каждого измерения оценим величину отклонения от среднего, разделенную на величину стандартного отклонения

$$t_i = \frac{X_i - < X >}{\sigma} .$$

Далее эта величина должна сравниваться с табличной величиной [1] в соответствии с критерием Стьюдента. Число степеней свободы критерия есть $(n-1)$, где n – объем выборки. Доверительная вероятность должна выбираться не менее 90%. В принципе, если для всех измерений не очень длинной выборки (менее 120) указанная величина не превосходит трех, то можно утверждать, что выбросы (резко отстоящие значения) отсутствуют. В нашем случае никакое из измерений таблицы выбросом не является.

Другая группа оценок требует анализа упорядоченной (ранжированной) выборки. Получаемые оценки (медиана и отклонение Хемпеля) также характеризуют центр распределения (среднее в некотором смысле значение) и разброс. В целом близость выборочного среднего и медианы свидетельствует об отсутствии значимого вклада крайних (больших или меньших) значений признаков. Наоборот, существенное расхождение выборочного среднего и медианы может свидетельствовать о наличии относительно малого числа больших или меньших измерений, влияющих на величину выборочного среднего. При этом данные измерения по критерию Стьюдента могут не быть выбросами.

Рассмотрим выборку {64, 53, 56, 70, 73, 58}. Это измерения частоты сердечных сокращений (ЧСС) для Класса 0 (белых мышей). Рассчитаем параметры для этой выборки: $<X> = 62{,}33$; $\sigma = 8{,}02$; наиболее отстоящее от среднего измерение 73 дает величину $t = 1{,}33$, т. е. не является выбросом. Ранжирование измерений (упорядочивание по возрастанию) позволяет получить выборку {53, 56, 58, 64, 70, 73}. Медиана – как середина выборки, в нашем случае будет полусуммой третьего и четвертого измерения $\mathrm{Med} = (58+64)/2 = 61$. Составим новую величину – модуль отклонения измерения от медианы, сосчитаем ее медиану. По отношению к исходному измерению "медиана от медианы" и будет отклонением Хемпеля. Получаем $H = 6{,}5$. Дополнительный вывод об относительной однородности распределения измерений можно сделать, исходя из близости выборочного среднего и медианы.

Выявление закона распределения измерений требует разбиения выборки на несколько интервалов (диапазонов), подсчет числа измерений, попадающих в каждый диапазон и построение гистограммы (см. разд. 5). Требуется разбиение, по крайней мере, на 5–6 интервалов, для чего необходимы объемы выборки не менее 30 измерений (см. разд. 5). В нашей ситуации выборки 6, 9 (Классы 0 и 1) и 15 (объединенная выборка) недостаточны.

Этап 2. Решение задачи 2 дискриминации Классов 0 и 1 на основе измерения признаков дано в разд. 2. Основная идея – оценивание по критерию Стьюдента степени расхождения между выборочными средними измерений, принадлежащих разным классам. Отличие от отбраковки выбросов состоит: в определении числа степеней свободы и в оценке величины σ (см. разд. 2). В табл. 10 приведены выборочные средние (первые 2 строки), параметры σ и критерий Стьюдента t для признаков масса, длина, ЧСС и реакция на лекарства.

Таблица 10

Расчет величины критерия Стьюдента t для дискриминации классов

Величина	Масса	Длина	ЧСС	Реакция
Класс 0	165,83	3,967	62,33	1,500
Класс 1	168,33	4,111	74,44	1,556
σ	10,90	0,110	2,549	0,378
t	0,229	1,312	4,751	0,148

Данные свидетельствуют о том, что только измерения ЧСС по Классам 0 и 1 значимо различаются (с доверительной вероятностью не менее 0,999 или $p < 0{,}001$). По всем остальным признакам не отмечено статистически значимых различий. Таким образом, дискриминация Классов 0 (белые мыши) и 1 (черные мыши) возможна только на основании измерений ЧСС.

Алгоритмы дискриминации могут быть различны. Наиболее простой – пороговый алгоритм: вводим порог P и следующее решающее правило – "если $X \le P$, соответствующее измерение относится к Классу 0, если $X > P$ – к классу 1". При этом в качестве порога выбирают значение, лежащее между выборочными средними. Формальные критерии выбора приведены в разд. 2.

В нашем случае в качестве порога можно выбрать $P = 70$. Тогда измерения {64, 53, 56, 70, 58} достоверно отнесены к Классу 0 (а измерение 73 ошибочно – ошибка 1-го рода отнесена к Классу 1). Соответственно, измерения {75, 71, 80, 74, 83, 72, 81} достоверно отнесены к Классу 1, а {66, 68} ошибочно отнесены к Классу 0 (ошибка 2-го рода), т. е. суммарное число ошибок – 3 составили $3/15 = 20\%$ выборки. Достоверность дискриминации Классов 0 и 1 на основании измерений ЧСС составила 80%.

Этап 3. Корреляционный анализ (см. разд. 3) позволяет выявить статистическую связь между признаками. Различные методы вычисления коэффициентов корреляции описаны в разд. 3 (модельный пример). Матрица выборочных коэффициентов корреляции приведена ниже. Здесь столбцы: 1 имеет признак "масса", 2 – длина, 3 – ЧСС и 4 – реакция на лекарства

$$\begin{pmatrix} 1 & 0{,}896 & -0{,}125 & 0{,}484 \\ 0{,}896 & 1 & -0{,}065 & 0{,}546 \\ -0{,}125 & -0{,}065 & 1 & 0{,}203 \\ 0{,}484 & 0{,}546 & 0{,}203 & 1 \end{pmatrix}.$$

Интерпретация этих результатов: высокая положительная корреляция между массой и длиной хвоста свидетельствует о том, что мыши большей массы, как правило, имеют более длинные хвосты; достаточно сильная корреляционная связь между реакцией на лекарства и массой (длиной хвоста) – более сильная реакция на лекарства характерна для более крупных мышей; наблюдается не очень сильно выраженная положительная корреляция между ЧСС и степенью реакции на лекарства.

Этап 4. Метод факторного анализа также позволяет сделать некоторые выводы о связи между признаками. Как говорилось выше, математически требуется решить полную проблему собственных чисел. Сумма собственных чисел равна следу матрицы (сумме диагональных элементов), и в случае корреляционной матрицы это будет число признаков. Собственные числа (в случае числа признаков большем четырех) должны находиться численными методами, так как никакое уравнение общего вида степени 5 и выше аналитически не решается. В табл. 11 представлены величины собственных чисел и соответствующие собственные векторы.

Таблица 11

Структура собственных векторов корреляционной матрицы (факторов)

Число	Вектор
2,3035	$0{,}611X_1 + 0{,}624X_2 - 0{,}014X_3 + 0{,}486X_4$
1,1160	$0{,}176X_1 + 0{,}100X_2 - 0{,}904X_3 - 0{,}376X_4$
0,4804	$-0{,}346X_1 - 0{,}284X_2 - 0{,}426X_3 + 0{,}787X_4$
0,1000	$-0{,}690X_1 + 0{,}721X_2 - 0{,}030X_3 - 0{,}059X_4$

В нашем случае интерпретация результатов факторного анализа достаточно сложна. В целом его можно применять для выявления некоторых устойчивых комбинаций признаков, “работающих” кооперативно или в противодействии друг другу. В частности, применительно к первому (наиболее мощному, исходя из величины собственного числа) фактору, можно допустить совместную работу признаков 1, 2 и 4 (положительные большие коэффициенты) при практическом отсутствии влияния 3-го признака. Возможная интерпретация: если у особи большая масса (признак 1), то, как правило, и большая длина хвоста (признак 2) и более сильная реакция на лекарства (признак 4).

Обращаем внимание на то, что сумма собственных чисел действительно равна 4, собственные векторы (факторы) ортогональны.

Этап 5. Регрессионный анализ позволяет выявить и сформулировать (в форме уравнения) связь между различными признаками. Важным классом таких уравнений являются уравнения линейной регрес-

сии, в которых один из признаков представляется в форме линейной комбинации каких-либо других признаков. Коэффициенты регрессионного уравнения вычисляются на основе известного метода наименьших квадратов (МНК). Часто, если проводится хорошая (достоверная) дискриминация классов по пороговому алгоритму, уравнение линейной регрессии имеет низкую достоверность.

Поскольку сами измерения интерпретируются как случайные величины, то коэффициенты регрессии также являются случайными величинами со своими характеристиками положения и разброса. Последнее приводит к необходимости оценивания так называемых "доверительных зон" уравнения регрессии. Коротко эта процедура и результаты ее применения приведены в разд. 5. Подробное описание изложено в работе [1].

Методы статистической обработки информации достаточно традиционны и имеют весьма широкое распространение, в том числе в экологии и медицине. Из описанных выше методов, лишь факторный анализ применяется относительно редко. По-видимому, это связано с необходимостью решения достаточно сложных проблем (полная проблема собственных чисел) и с необходимостью содержательной интерпретации факторов.

Заметим, что самостоятельную ценность имеет медицинская информация, представленная в Приложении данного пособия. В частности, эти данные можно использовать в качестве контрольных выборок для отладки различных алгоритмов (заполнения пропущенных значений, разбиения выборки на обучающую и контрольную и т. д.)

ВОПРОСЫ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

1. Какие типы измеряемых величин Вы знаете ? К какому типу величин относится принадлежность человека к европеоидной (*1*), монголоидной (*2*) или негроидной (*3*) расе в качестве измеряемого признака ?

2. Если в качестве другого измеряемого признака человека рассмотреть его финансовый статус, определяемый, как 0 – отсутствие доходов; 1 – заработок или пенсия ниже прожиточного минимума; 2 – заработок не выше среднего по стране; 3 – заработок, от среднего по стране до превышающего его в 10 раз; 4 – заработок, превосходящий средний по стране более, чем в 10 раз, то, что общего и в чем различия между данными признаками?

3. В чем основные различия между выборочным средним и медианой как характеристиками среднего значения измеряемой величины ? Для какой из них требуется ранжирование измерений ? Какая из них чувствительна к выбросам ?

4. С каким критерием связано решение задачи об выявлении выбросов ? Требуется ли при решении оценивать стандартное отклонение измерений ?

5. При дискриминации одних и тех же классов: в первом случае величина критерия $t = 2{,}75$, а во втором – $t = 3{,}25$. Когда дискриминация более достоверна ?

6. В чем разница между ошибками дискриминации первого и второго рода ?

7. Для нахождения порога в дискриминантном анализе требуется минимизировать или максимизировать суммарную долю ошибок первого и второго рода ? Возможны ли какие-либо другие критерии выбора порога и при каких условиях ?

8. Что характеризует коэффициент корреляции ? Какие алгоритмы вычисления корреляционной матрицы Вы знаете ? В каких случаях медианный коэффициент корреляции не может быть вычислен ? Приведите пример соответствующих выборок измерений.

9. Если корреляционная матрица имеет размерность 6x6, чему должна быть равна сумма собственных чисел этой матрицы ? Можно ли при произвольных элементах этой матрицы найти собственные числа и собственные векторы (факторы) аналитически ? Если нет, то почему?

10. Какой метод применяется для определения коэффициентов уравнения линейной регрессии ? Требуется ли для использования этого метода ранжирование измерений ? Является ли в этом случае оценка коэффициентов уравнения линейной регрессии, устойчивой к выбросам ?

Библиографический список

1. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М.: Наука, 1983. 416 с.

2. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973. 900 с.

3. *Рыжик И. М.* Таблицы интегралов, сумм, рядов и произведений. М.,–Л.: Гостехиздат, 1943. 400 с.

4. *Лоусон Ч., Хенсон Р.* Численное решение задач метода наименьших квадратов: Пер. с англ. М.: Наука, 1986. 232 с.

5. *Анго А.* Математика для электро- и радиоинженеров / Пер. с фр. Под ред. *К. С. Шифрина* М.: Наука, 1964. 772 с.

6. *Новицкий П. В., Зограф И. А.* Оценка погрешностей результатов измерений. Л.: Наука, 1991. 248 с.

7. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. М.: Наука, 1968. 720 с.

8. *Гнеденко Б. В.* Курс теории вероятностей. М.: Наука, 1988. 448 с.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
01/01	0	–21,6	2,8	3,5	7	650	121	8	1
02/01	1	–10,7	3,7	3,8	5	500	116	9	1
03/01	1	–21,8	5,8	5	7	610	54	1	1
04/01	0	–21,1	3,9	3,5	5	430	88	2	1
05/01	0	–14,1	5	3,9	4	710	132	3	1
06/01	3	–13,6	36,1	5,5	3	1040	179	11	1
07/01	2	–2,6	37,6	5,3	3	1220	202	12	1
08/01	0	0,7	3,4	5	7	1170	177	19	1
09/01	0	–0,6	8,8	4,6	55	970	159	11	1
10/01	0	–2,9	8,7	3,5	148	830	146	12	1
11/01	0	–4,5	7,1	1,6	9	490	109	2	1
12/01	0	–6,9	19,1	3,1	18	640	150	12	1
13/01	1	–2,1	22,5	3,9	8	970	154	15	1
14/01	0	–4,3	4,2	4,9	9	750	145	3	1
15/01	0	–1,5	17,8	5,4	30	860	149	7	1
16/01	0	–3,3	16,4	5,8	32	990	134	7	1
17/01	1	–3,1	19,2	2,9	56	1190	155	13	1
18/01	2	–5,7	11,8	2,5	78	1190	140	16	1
19/01	0	–8,5	4,2	1,3	17	1060	135	3	1
20/01	0	–8,5	1,9	2,5	11	950	119	0	1
21/01	0	–9,4	3,4	1,8	10	750	82	1	1
22/01	0	–12,8	3,4	3,3	9	660	100	2	1
23/01	0	–19,7	15,2	1,8	7	620	80	1	1
24/01	1	–19,3	25,8	1,6	25	720	103	8	1
25/01	1	–14,3	37,2	3,1	15	530	120	9	1
26/01	0	–3,2	2,4	3,5	10	820	120	11	1
27/01	0	–4,8	1,3	2,6	11	830	100	10	1
28/01	0	–7,7	2,9	1	16	810	131	13	1
29/01	1	–0,5	4,1	1,5	35	970	171	14	1
30/01	0	–10	0,8	3,5	18	1140	153	3	1
31/01	0	–9,8	4,4	3,4	16	1370	182	20	1
01/02	0	–8,9	2,4	3,1	11	1730	169	10	1

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
02/02	0	−7,5	2,7	2,9	5	1570	139	13	1
03/02	0	−5,3	3,3	2,9	8	1120	149	9	3
04/02	0	−7	19	3,8	134	1280	156	14	1
05/02	0	−16,9	8,1	2	308	1510	150	22	2
06/02	0	−18,8	4,4	1,1	72	1350	117	14	1
07/02	1	−22,7	10,2	1,6	74	1180	84	5	1
08/02	1	−20	5,2	1,5	22	1250	111	5	1
09/02	0	−20,3	7	1,1	67	850	107	4	1
10/02	1	−16,5	9,2	2,6	21	510	34	0	1
11/02	3	−8,5	1,5	2,8	110	380	32	0	1
12/02	0	−9	4	3,3	50	70	25	0	1
13/02	0	−12,8	8,5	1,6	49	20	16	0	1
14/02	1	−15,7	13,3	1,5	49	10	15	0	1
15/02	1	−14	14,5	1,8	45	20	34	0	1
16/02	1	−3,2	22,4	5,3	42	50	48	1	1
17/02	0	−7,3	0,2	3,1	23	210	38	0	1
18/02	1	−4,8	3,6	1,6	18	300	46	7	1
19/02	2	−0,6	8,4	4	16	240	54	0	1
20/02	0	−1,6	5,4	3,4	104	200	50	3	1
21/02	1	−7,6	8,8	2,6	54	250	62	0	1
22/02	0	−15	1,7	1,4	27	230	61	1	2
23/02	1	−8,9	4,9	3,4	28	400	79	1	1
24/02	0	−6,9	9,2	2,6	19	500	78	3	1
25/02	0	−12,2	6,7	3	7	660	98	4	1
26/02	0	−6,3	22,2	4	5	800	110	2	1
27/02	0	−20,9	3,1	2,5	9	940	148	14	1
28/02	1	−22,6	19,2	1,8	14	750	65	2	1
01/03	1	−15	13,7	1,3	43	1120	151	15	1
02/03	0	−21,3	4,5	1	163	1130	154	9	1
03/03	0	−9	10,3	2,1	70	1100	136	6	1
04/03	0	−3,6	11,8	3,1	99	1130	182	7	1
05/03	0	−5,3	13,5	4,6	54	1140	152	6	1

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
06/03	0	–9,5	5,4	5,3	19	870	118	8	1
07/03	0	–12,3	5,9	3,3	5	860	99	1	1
08/03	1	–18,3	9,9	1,3	6	730	85	25	1
09/03	1	–0,7	4,5	1,1	7	750	119	13	1
10/03	0	–15,8	0,4	1,3	5	610	87	13	2
11/03	2	–11,4	12,6	2	76	440	78	2	1
12/03	0	–16	1,7	1,8	152	230	55	0	1
13/03	0	–8,9	6,1	2,6	68	10	11	0	1
14/03	0	–4,6	13,9	1,6	33	190	46	12	1
15/03	0	–3,1	11,6	4,1	20	360	68	4	1
16/03	0	–12,6	8,2	1,3	8	580	90	14	1
17/03	0	–11,6	3,7	1,3	11	800	96	14	2
18/03	0	–1,8	6,3	1,5	24	810	120	13	2
19/03	1	–2,5	8,2	2,3	41	750	129	6	1
20/03	3	–0,1	8,3	1,3	55	920	122	3	1
21/03	0	–6,2	6,5	1,4	30	850	128	4	1
22/03	0	–9,1	0,1	1,1	12	810	105	4	1
23/03	0	–2,5	7,5	4,1	11	710	90	7	1
24/03	0	1,1	10	3,1	13	650	105	3	1
25/03	0	0,3	0,1	2	78	350	108	4	1
26/03	0	1,2	5,6	2,4	11	550	121	9	1
27/03	1	2,2	9,3	1,8	6	360	106	1	1
28/03	0	2	3,8	1,9	85	170	79	3	1
29/03	0	3,8	2,4	2,8	96	150	76	2	1
30/03	0	1,7	1,3	3,6	50	190	79	2	1
31/03	0	1,9	2,5	3,8	61	210	87	1	1
01/04	0	0,9	2,5	2,4	36	260	94	3	1
02/04	0	2,3	3,5	1,5	24	320	126	3	1
03/04	0	2,6	3	3,4	17	260	101	3	1
04/04	1	4,8	9	2	59	10	73	6	1
05/04	1	6,1	7,8	1,4	22	280	79	9	1
06/04	3	5,6	5,6	0,9	67	230	103	4	1

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
07/04	0	5,2	2,7	3	35	270	88	6	1
08/04	1	4,7	14,7	2,4	23	300	89	11	1
09/04	0	3,2	7,6	2,9	24	340	90	7	1
10/04	2	2,3	4,9	2,9	30	390	100	3	1
11/04	0	0,2	11,3	2	10	370	85	5	1
12/04	0	–6,1	10,4	5,3	14	280	89	1	1
13/04	0	–3,4	3,1	4	52	430	90	8	1
14/04	0	–2,3	2	2,6	72	220	112	10	1
15/04	0	–5,5	5,8	2,1	164	250	125	7	1
16/04	0	–2,5	2,5	2,4	65	200	123	2	1
17/04	0	–1,3	5,7	1,9	17	450	115	3	1
18/04	1	5,5	12,3	3,6	17	790	110	0	2
19/04	0	7,7	4	3,4	14	940	125	16	1
20/04	4	9,9	7,2	2,8	17	1050	123	7	1
21/04	3	8,6	4,3	1,5	22	1090	122	8	1
22/04	1	7,8	2,6	1,4	18	960	106	8	1
23/04	0	8,6	4,9	1,9	39	1090	132	4	1
24/04	0	12	3	2,4	131	1230	158	2	1
25/04	1	11,3	5,9	1,5	79	1240	195	0	1
26/04	0	9,3	2,5	1	61	930	144	19	1
27/04	0	6,9	1,8	2,1	20	830	166	21	2
28/04	0	2,4	1	3,1	15	970	171	11	2
29/04	0	0,4	3,5	3	94	1100	177	14	1
30/04	0	–0,6	5,5	2,5	46	1110	131	7	2
01/05	4	0,3	7,4	1,4	60	970	137	20	2
02/05	2	2,3	9	3,8	25	1070	128	19	1
03/05	0	6,3	2,4	2,8	14	1010	130	9	1
04/05	0	6,5	7,3	0,9	60	850	125	5	1
05/05	1	6,3	16,7	2,8	24	660	137	14	1
06/05	0	2,9	0,1	2,9	19	550	99	10	1
07/05	0	3,7	1,8	2,3	13	840	120	19	2
08/05	1	7	5	1,3	17	880	140	20	3
09/05	0	12,4	2,9	1,9	9	1080	144	28	2

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
10/05	2	12,8	5,5	1	12	1260	168	27	3
11/05	0	13,3	2,2	2,1	89	2030	171	13	2
12/05	0	11,6	2,4	1,3	106	2410	175	41	3
13/05	1	11,2	5,9	2,5	84	2310	163	16	2
14/05	0	4,2	1,9	3,1	35	2170	170	21	2
15/05	1	5,5	7,1	2,8	42	1550	176	16	3
16/05	0	6,1	7,4	2,9	20	780	124	13	2
17/05	1	6,9	12,4	2,1	128	1220	153	9	0
18/05	0	8,7	5,6	3,4	19	1260	139	19	1
19/05	0	4,9	0,8	3,3	10	1310	157	15	1
20/05	0	6,2	19,2	3,3	13	1270	157	14	1
21/05	1	9,8	38,5	3,4	28	1210	131	12	1
22/05	0	3,9	11	4,9	123	1090	152	17	1
23/05	1	3,5	19,6	2,1	75	840	174	17	1
24/05	1	7,3	14,3	2,5	175	820	153	5	1
25/05	0	7,5	1,2	2,4	13	870	170	14	2
26/05	0	8,9	6,3	2,6	14	640	175	9	1
27/05	0	7	5,7	2,1	19	550	166	8	1
28/05	0	10,3	1,3	1,9	7	550	114	7	1
29/05	0	9,7	8	3,9	9	710	155	10	1
30/05	1	11,8	8,3	2,3	14	990	114	9	1
31/05	0	12,3	0,4	2,3	13	1230	86	16	2
01/06	0	12	4,2	2,3	14	940	104	1	2
02/06	0	10,2	1,2	3,1	15	710	106	8	2
03/06	0	12,3	10	3,4	12	760	111	23	2
04/06	0	14,7	6,9	5,1	7	1260	111	17	2
05/06	1	5,7	11,6	3,8	12	1970	112	30	2
06/06	0	4,2	0,3	2,9	14	2670	201	23	3
07/06	0	5,6	2,2	2,3	10	2780	178	18	2
08/06	0	8,1	1,6	1,4	17	2260	151	8	2
09/06	0	11	9,1	2,3	34	2240	149	13	2
10/06	0	8,5	7,8	2,3	49	1900	133	11	2

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
11/06	1	12,2	17,6	2,6	18	1360	101	0	1
12/06	2	14,2	16	2,9	16	640	105	7	1
13/06	0	15,4	3,6	2,1	86	700	126	19	1
14/06	1	17,4	8,1	1,4	15	630	136	14	1
15/06	0	19,5	1,2	1,3	27	560	121	5	1
16/06	0	22,3	1,9	2	14	610	123	4	2
17/06	1	22,5	5,3	4	20	510	121	10	2
18/06	0	15,7	2	3,6	41	490	122	5	1
19/06	0	9,3	5,8	1,5	27	590	153	28	2
20/06	2	8,9	10	4,3	23	720	16	15	2
21/06	0	5,7	0,6	3,8	18	770	182	15	1
22/06	0	4,5	3,5	4,4	32	850	202	16	1
23/06	0	3,6	4,2	5,8	23	930	215	8	1
24/06	1	6,6	8,1	4,6	9	990	230	11	2
25/06	1	7,6	8,2	2,8	5	850	217	9	2
26/06	1	7,2	17,8	5,1	18	970	162	7	2
27/06	0	11,8	1,3	2,5	10	990	136	11	2
28/06	1	11,6	8	3,1	18	0	121	0	2
29/06	0	19,6	7,1	3	14	710	114	20	2
30/06	0	21,6	0,4	2,4	11	670	113	8	2
01/07	0	16,3	1,8	3,3	8	500	112	9	2
02/07	0	19,3	5,6	2,5	9	450	89	25	1
03/07	0	22,5	3,8	1,6	10	710	90	14	1
04/07	0	20,9	3,5	2,1	10	690	107	16	1
05/07	0	20,6	1,8	2	12	680	101	11	1
06/07	0	19,1	6,5	2,3	17	940	132	7	1
07/07	1	13,2	8	3,6	20	860	113	3	1
08/07	1	14,5	1	2,3	12	990	109	1	1
09/07	0	18,1	0,9	2,9	13	760	105	1	1
10/07	0	16,4	0,4	2,8	7	740	89	13	2
11/07	0	17,3	3,8	3,6	9	970	132	12	1
12/07	0	14,6	5,4	2,4	22	920	134	6	1
13/07	0	15	11	4,6	28	800	95	11	1

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
14/07	1	15,1	21,7	3,9	12	700	129	0	1
15/07	0	12,3	1,6	2,3	9	590	141	4	1
16/07	0	15,9	5,5	4,6	35	550	122	7	1
17/07	0	16,9	13,7	4,8	39	560	159	8	1
18/07	1	10,1	19,2	3,5	24	570	159	13	1
19/07	1	11,6	13,2	2,1	13	330	140	0	1
20/07	1	15,3	12,2	3,4	7	350	155	4	1
21/07	1	18,9	2,7	2,8	15	510	167	8	1
22/07	0	21,8	1,7	2,6	13	860	176	12	1
23/07	0	22,5	3,6	2,4	43	990	155	4	2
24/07	1	25,2	4,4	2,1	54	670	150	11	1
25/07	0	17,7	4,7	3,1	19	880	132	9	2
26/07	2	11,1	2,9	3,9	12	780	73	5	1
27/07	0	12,4	1,2	1,9	10	520	77	14	2
28/07	0	12,2	2,9	2,1	20	910	107	1	1
29/07	0	10,4	6,6	2,3	19	1150	131	4	1
30/07	0	12,8	5,8	4,1	24	1330	138	17	2
31/07	0	15,4	1,4	3,9	13	1420	179	15	1
01/08	0	13,3	10,4	3	9	1270	164	14	2
02/08	1	12,2	19,9	3,5	41	1050	158	8	1
03/08	0	13,4	1,3	0,9	26	1080	174	9	1
04/08	0	18	2,7	1,9	6	880	180	14	2
05/08	0	19,1	5,9	1,6	8	710	138	4	2
06/08	2	17,5	12,3	3,9	10	750	118	10	1
07/08	1	12,6	6,8	3,9	63	830	106	3	2
08/08	1	14,3	3,4	3,6	134	940	76	7	2
09/08	1	16,6	11,5	2,6	13	1400	142	8	1
10/08	1	16,1	15,8	5,4	10	950	118	13	1
11/08	0	15,1	7,6	3,3	9	970	127	7	1
12/08	1	14,3	17,5	3,4	38	970	165	27	1
13/08	1	8,5	21,1	3,5	43	970	147	13	2
14/08	1	13,7	18,8	2,9	15	530	126	9	2
15/08	0	8,4	5,2	4,1	16	540	143	18	2

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
16/08	1	9,6	14,5	2,4	10	380	120	3	1
17/08	0	11,7	11,1	3,9	8	380	110	3	2
18/08	0	10,9	2,3	4,9	4	270	94	11	1
19/08	0	8	5,3	1,4	11	210	84	1	1
20/08	0	10,9	1,5	2,5	17	150	81	1	1
21/08	0	8,7	8,2	3	32	100	78	2	2
22/08	0	8,6	5,4	2,3	22	120	82	6	1
23/08	0	14,3	1,2	2,9	51	120	78	0	1
24/08	0	10,7	0,6	2,8	31	240	92	1	2
25/08	0	6,6	1,5	3,5	42	170	70	1	1
26/08	0	6,7	1,8	4,1	25	210	80	1	1
27/08	0	6,7	2,5	4,4	9	260	84	4	1
28/08	0	5,6	4,8	1,8	11	270	94	2	1
29/08	0	6,5	8,8	2	40	210	87	2	1
30/08	1	8,3	17,2	2,4	21	230	100	4	1
31/08	0	10,4	7,6	3,9	26	220	83	7	1
01/09	0	8,4	20,7	5,4	21	360	63	10	1
02/09	0	7,6	2,3	3,6	7	470	80	5	2
03/09	0	9,9	7,5	2,4	7	240	81	4	1
04/09	1	10,1	8,8	1,5	6	480	98	10	1
05/09	0	7,4	4,7	4,6	7	480	101	10	1
06/09	1	10,5	7,2	3,9	7	420	98	1	1
07/09	0	9,8	4,2	3,1	26	360	102	2	1
08/09	1	10,3	6	3,9	13	420	121	3	1
09/09	0	11,2	0,1	4,8	14	410	101	2	1
10/09	1	10,3	7,7	2,9	13	260	116	2	1
11/09	0	9	4,4	2,3	11	120	97	5	1
12/09	1	10,8	11,5	0,5	9	120	101	4	2
13/09	1	9,2	8,5	2	10	130	82	5	1
14/09	0	8,6	6,3	1	7	17	60	2	1
15/09	2	9,6	8,7	2,4	42	220	57	9	1
16/09	0	9,7	7,5	1,8	28	250	75	4	1

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
17/09	0	10,8	1,8	1,3	23	320	56	8	1
18/09	0	13,4	7	3,4	21	380	69	6	1
19/09	0	12,4	1,2	3	13	190	60	5	1
20/09	0	13,3	0	4,9	25	150	54	6	1
21/09	0	12,3	1,5	4,1	13	160	74	4	1
22/09	0	11,4	0,6	4	10	390	54	10	1
23/09	1	9,3	11,9	2,9	6	630	48	3	1
24/09	0	10,9	2,8	4,1	8	740	64	5	1
25/09	0	7,2	9,7	2	39	760	66	10	1
26/09	1	6,7	17,8	2,9	28	950	83	3	1
27/09	2	0,3	12,2	4,5	16	910	62	4	1
28/09	1	0,2	19,2	3,5	14	750	65	2	1
29/09	1	2,4	10,6	5	9	760	66	10	1
30/09	1	1,4	12,9	3,6	5	660	48	4	1
01/10	0	2,3	10,8	2,5	16	1080	58	5	1
02/10	0	2	1,1	2,6	24	990	75	3	1
03/10	0	2,6	7,5	2,5	25	1930	95	4	1
04/10	0	0,9	7,8	2,3	63	1490	92	2	1
05/10	0	-0,2	15,2	3,4	10	1120	94	3	2
06/10	1	0,4	28,7	7	26	103	115	9	1
07/10	1	3,5	10,2	5	11	990	115	17	1
08/10	0	4,7	2,7	2,3	17	1090	178	1	1
09/10	0	3,1	0,3	2,8	5	930	144	12	1
10/10	0	3,6	7,7	1,6	9	960	140	17	1
11/10	1	4	11,8	2,4	5	1020	154	22	1
12/10	2	2,8	2,9	2,5	4	1070	153	11	1
13/10	0	1,9	4,3	3,1	45	890	140	11	1
14/10	0	1,3	4,5	2,4	30	1120	97	19	2
15/10	0	3,9	2,3	3,8	25	960	100	18	1
16/10	1	6,1	9,9	3,3	12	830	95	17	2
17/10	1	5,3	3,9	3,3	98	710	107	17	1
18/10	1	8,4	7,8	4,9	105	510	80	5	1

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
19/10	0	7,8	0,3	4,4	9	270	66	4	2
20/10	0	6,3	17,6	3,8	9	30	36	2	1
21/10	1	5,2	12,6	3	17	20	23	0	1
22/10	1	2,3	31	3,9	19	20	33	0	1
23/10	0	1,9	5,2	3,3	23	30	23	0	0
24/10	0	3	9,7	4	19	80	24	0	0
25/10	0	1,3	21,7	5,9	6	70	63	0	0
26/10	0	−2	11,1	3,6	3	50	36	0	0
27/10	1	0,1	50	6,4	5	40	28	0	1
28/10	0	−2,4	4,7	3,8	14	10	37	0	0
29/10	0	−5,1	1	2,4	51	30	22	0	0
30/10	0	−0,2	8,9	5,4	22	30	29	1	0
31/10	1	2,4	15,5	5,1	8	160	17	1	1
01/11	3	2,5	22,4	4,4	32	210	39	1	2
02/11	2	0,6	36,7	3,5	34	270	51	0	1
03/11	0	0	17,3	3,8	22	370	62	3	0
04/11	0	−2,6	4,6	3,8	9	410	66	3	1
05/11	0	−5,8	7,6	3,5	6	420	92	0	1
06/11	2	−3	21	5,5	2	430	32	0	1
07/11	0	1,4	5,9	6,1	23	330	122	6	1
08/11	1	0,5	19,4	3	43	240	107	12	2
09/11	0	2,9	9,5	8	84	10	75	0	1
10/11	2	3,3	15,4	2,6	37	10	103	0	1
11/11	1	−7,6	16,4	1	43	180	67	0	1
12/11	1	−9,2	19,5	1,4	31	80	80	2	1
13/11	0	−12	16,9	2	60	100	62	0	1
14/11	0	−14,1	12	1,6	41	60	85	0	0
15/11	2	−11,8	14,9	5	60	180	63	1	1
16/11	2	−9,1	15,4	4,4	69	100	60	1	1
17/11	0	−14,4	8,1	1,4	71	120	37	0	0
18/11	0	−11,6	8,2	4,5	23	80	40	0	0
19/11	0	−6,3	6	6,3	14	10	33	0	0

Продолжение прил.

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
20/11	1	−9,8	27,8	4,9	27	0	0	0	1
21/11	0	−8	11,9	4,4	13	0	0	0	0
22/11	0	−9,5	2,1	1,9	6	0	0	0	0
23/11	0	−13,6	1,6	0,6	2	10	0	0	0
24/11	1	−11,2	0,8	3,5	8	0	1	0	1
25/11	0	−13,8	11,6	4,5	14	10	0	0	0
26/11	1	−15,3	25	2,6	22	10	1	0	1
27/11	1	−9,4	14,3	4,4	4	100	12	2	1
28/11	0	−11,3	7,5	4	32	70	30	1	0
29/11	0	−15,4	7,2	1,1	18	50	33	2	0
30/11	0	−11,8	6,9	1,6	20	60	35	1	0
01/12	2	−10,4	9	3,3	11	40	62	0	1
02/12	1	−9	22,9	3,5	6	30	47	0	1
03/12	0	−7,3	11,2	3,4	7	70	30	0	0
04/12	0	−5,9	10,2	3,5	7	70	53	0	0
05/12	0	−26,4	5	1,8	19	180	65	3	1
06/12	0	−19,6	5,2	3,6	28	290	62	15	2
07/12	1	−10,8	11,5	2,6	47	260	71	12	2
08/12	0	−15,4	15,1	2,8	9	420	119	8	1
09/12	1	−6,6	22,2	3,4	5	390	122	1	1
10/12	1	−7,9	12,5	2,1	35	330	143	7	1
11/12	0	−21,9	4	1,9	50	250	117	1	1
12/12	0	−15,5	4	2	32	330	90	3	1
13/12	1	−11,6	25	2,8	37	240	99	6	1
14/12	0	−8,9	5,7	4,9	42	190	76	4	0
15/12	0	−3	9,2	4,4	21	160	73	14	0
16/12	0	−0,8	2,6	3,9	4	300	63	7	0
17/12	0	−0,3	1,3	4,6	7	250	61	12	1
18/12	0	−3,6	5	2,9	12	190	38	2	1
19/12	1	−6,7	15,2	2,6	11	160	44	5	1
20/12	0	−7	7,5	2,1	5	190	28	3	0
21/12	1	−8,2	12,7	3,4	4	200	29	1	1

Дата	*n*	*t*	*p*	*vcp*	*ak*	*s*	*w*	*ps*	*as*
22/12	0	−5,7	3,4	2	7	130	27	0	0
23/12	0	−0,3	2,6	2,9	8	90	28	0	0
24/12	0	1,1	7,8	3,8	26	130	27	0	0
25/12	1	0,7	11,1	3,3	11	30	37	1	1
26/12	0	−3,5	9,8	4,1	23	10	51	1	0
27/12	1	−7,9	37,8	3,1	21	10	13	0	1
28/12	0	−8,8	2,5	4,4	9	30	12	0	0
29/12	0	−8,2	12,6	2	8	40	26	0	0
30/12	0	−9,8	1,2	1,1	35	0	28	0	0
31/12	0	−13	1	2	32	20	11	0	0

Условные обозначения: *n* – число случаев смерти в течение суток (0 – 5); *t* – среднесуточная температура (°С); *p* – перепад давления в течение суток (мм рт.ст.); *vcp* – средняя скорость ветра (м/с); *ak* – индекс возмущенности магнитного поля Земли (усл. ед.); *s* – площадь солнечных пятен (миллионные доли площади полусферы); *w* – число Вольфа; *ps* – число солнечных вспышек в течение суток; *as* – интегральный показатель солнечной активности (балл от 0 до 3).

Оглавление

Учебное издание

Буляница Антон Леонидович
Курочкин Владимир Ефимович
Кноп Инга Сергеевна

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКОЛОГИЧЕСКОЙ ИНФОРМАЦИИ: ДИСКРИМИНАНТНЫЙ, КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Учебное пособие

Редактор *А. В. Семенчук*
Компьютерная верстка *А. Н. Колешко*

Сдано в набор 07.12.04. Подписано к печати 18.02.05. Формат 60×84 1/16. Бумага офсетная. Печать офсетная. Усл. печ. л. 2,79. Усл. кр.-отт. 2,9. Уч. -изд. л. 3,45. Тираж 100 экз. Заказ №

Редакционно-издательский отдел
Отдел электронных публикаций и библиографии библиотеки
Отдел оперативной полиграфии
СПбГУАП
190000, Санкт-Петербург, ул. Б. Морская, 67