

ДИСКРЕТНЫЕ МОДЕЛИ РЕАЛЬНЫХ ПРОЦЕССОВ

УДК 519.168

МЕТОД ОЦЕНКИ СТЕПЕНИ СВЯЗИ БИНАРНОГО
И НОМИНАЛЬНОГО ПОКАЗАТЕЛЕЙ

С. В. Дронов, И. Ю. Бойко

Алтайский государственный университет, г. Барнаул, Россия

При обработке современных статистических данных всё чаще появляется необходимость исследовать связь между числовым (номинальным) и бинарным показателями. В работе рассмотрен новый коэффициент для оценивания силы связи между такими показателями, названный ледж-коэффициентом и ориентированный на выявление связей типа «ступеньки», означающей существование заранее неизвестных, но постоянных границ для числового показателя, при попадании в которые бинарный показатель в основном принимает только одно из своих значений. Предлагаются алгоритмы для вычисления ледж-коэффициента, а также для перебора всех случаев, в которых связь оказывается слабой и сильнейшей из возможных. Приведены сравнительные примеры оценивания силы связи по ледж-коэффициенту и коэффициентам корреляции Пирсона и Кендалла.

Ключевые слова: бинарные и номинальные переменные, нормативные интервалы, связь типа «ступенька», бинарные цепочки.

DOI 10.17223/20710410/30/11

METHOD FOR ESTIMATING CONNECTION POWER OF BINARY
AND NOMINAL VARIABLES

S. V. Dronov, I. Yu. Boyko

*Altai State University, Barnaul, Russia***E-mail:** dsv@math.asu.ru

Nowadays, one of the common connection types in statistical data processing is the connection between numerical (nominal) and binary variables. We propose a new coefficient for evaluation of this type connection. It is quite different from Pearson's correlation coefficient and focused at the "ledge" type of connection, which means that the binary variable is expected to take one of its values only within some unknown but fixed boundaries for the the numerical variable. We call it ledge-coefficient. Application scope for the proposed coefficient is discussed. Algorithms for calculating ledge-coefficient and for searching all cases when the ledge connection is extremely weak or strong are proposed. Comparative examples for estimating connection power by ledge-coefficient and by Pearson's correlation coefficient or by Kendall's tau are shown.

Keywords: binary and nominal variables, normative intervals, ledge connection, binary string.

Введение

Практически любая обработка данных эксперимента имеет одной из своих целей установления факта связи определённых показателей, а также оценку степени (силы) этой связи. Сегодня достаточно активно разрабатываются современные методы решения этой задачи для случая, когда показатели измеряются в шкалах разных типов (к примеру, аппараты качественной и логистической регрессий, модель Раша в тестологии и многие другие). Тем не менее в практических расчетах проблема все ещё часто решается с помощью коэффициента корреляции Пирсона между показателями, что фактически ограничивает исследование лишь линейными связями. Исследователь-практик объясняет такой выбор метода, как правило, тем, что он склонен считать распределения нужных ему показателей нормальными, тем самым оправдывая законность изучения только связей, имеющих существенную линейную составляющую. Однако в одном из случаев ясно, что предположения нормальности показателя заведомо не отвечают объективной структуре данных: это ситуация, когда показатель бинарен и может принимать только значения 0 или 1. Случай, когда один из показателей принимает числовые значения (номинальный), а второй бинарен, рассматривается в настоящей работе.

Описанная ситуация является довольно обычной, особенно в медицине, которая сегодня все чаще использует в своих моделях математику [1, 2]. Номинальными являются числовые медицинские показатели (кровенное давление, температура тела и пр.), бинарными — наличие или отсутствие определённых признаков. В медицинских работах используют либо уже упомянутый коэффициент Пирсона, либо коэффициенты бисериальной или точечно-бисериальной корреляции. Однако внимательное изучение этих коэффициентов показывает, что они являются лишь вариантами коэффициента Пирсона, а значит, применение их также не вполне законно.

Действительно, линейная связь, наличие которой обеспечивает значимая величина коэффициента корреляции, предполагает, что все точки поля корреляции выборочных данных довольно хорошо аппроксимируются некоторой прямой линией, что, как правило, невозможно, когда одна из координат принимает только два различных значения. Таким образом, даже в случае очень сильно выраженной связи аппроксимирующая линия должна быть иной.

Один из видов такой линии выявляется в медицинской практике. Так, большинство числовых показателей в медицине имеют нормативные интервалы, при выходе за которые возникает подозрение на заболевание. Если X — значение такого числового показателя, то при некоторых a, b признак здоровья пациента Y должен в случае наличия предельно сильной связи задаваться, таким образом, формулой

$$Y_{a,b}(X) = \begin{cases} 1, & a \leq X \leq b, \\ 0 & \text{иначе.} \end{cases} \quad (1)$$

Далее такую функцию будем называть ступенькой, а соответствующую связь (при её установлении) связью типа «ступенька». Как правило, интервал $[a, b]$ будем считать неизвестным. Этот тип связи впервые, по-видимому, изучался в работах [3, 4].

Исходными данными для оценки силы связи служат две связанные выборки объёма n каждая: $X = (x_1, \dots, x_n)$ и $Y = (y_1, \dots, y_n)$, где вторая состоит из 0 и 1. Требуется выяснить, насколько точно облако точек (x_i, y_i) , $i = 1, \dots, n$, при наилучшем выборе границ a, b аппроксимируется графиком (1), при этом оценив степень аппроксимации некоторым числовым коэффициентом.

Отметим, что в силу характера задачи сами значения первого показателя нам не нужны, а важно лишь расположение точек (x_i, y_i) , $i = 1, \dots, n$, относительно какой-нибудь ступеньки. Значит, без ограничения общности (меняя, если нужно, одновременно порядок следования элементов в обеих выборках) условимся далее считать, что $x_1, \dots, x_n$ расположены в выборке в порядке неубывания. Более того, можно заменить эти значения их последовательными номерами (рангами), тем самым сведя задачу лишь к рассмотрению нужным образом упорядоченной второй выборки. Теперь можно поставить задачу строго.

1. Постановка задачи

Будем рассматривать цепочки Y длины n (из нулей и единиц) с целью определить степень их соответствия зависимости (1) при наилучшем возможном выборе границ a, b , где в роли X выступает номер соответствующего элемента Y . Для этого в фиксированной цепочке Y будем минимизировать величину ошибок аппроксимации

$$S_{[a,b]}(Y) = \sum_{j=1}^n (y_j - Y_{a,b}(j))^2 \quad (2)$$

путём подбора границ a, b . Если a и b заданы, то $S_{[a,b]}(Y)$ равно сумме числа нулей внутри отрезка $[a, b]$ и числа единиц вне него. За критерий качества соответствия для цепочки Y примем

$$S(Y) = \min_{a,b} S_{[a,b]}(Y). \quad (3)$$

Тот отрезок $[a, b]$, на котором достигается минимум, назовём оптимальным. Оптимальных отрезков может быть несколько.

Выделим класс $A_{k,m}$ цепочек, в которых $k > 0$ единиц и $m > 0$ нулей ($k + m = n$). Минимальное и максимальное значения $S(Y)$ среди всех цепочек из $A_{k,m}$ были ранее найдены в [3], но непосредственное вычисление величины наименьшего возможного количества ошибок $S(Y)$ для заданной цепочки Y выделенного класса по всем возможным значениям границ отрезка $[a, b]$ внутри неё вызывает серьёзные затруднения. В связи с этим ставится задача создания относительно простых алгоритмов поиска наименьшего возможного количества ошибок $S(Y)$ для заданной цепочки $Y \in A_{k,m}$, а также перечисления всех оптимальных отрезков в ней.

При более подробном изучении ситуации оказывается, что цепочки из $A_{k,m}$, в которых $S(Y)$ максимально, устроены довольно сложно. Поэтому предлагается алгоритм, позволяющий последовательно конструировать все такие цепочки. Отметим, что в [3] сформулировано лишь достаточное условие того, что конкретная цепочка даёт максимально возможное количество ошибок и, как следствие, указаны только некоторые из таких цепочек.

После решения поставленных задач практическое оценивание силы связи показателей изучаемого типа не будет вызывать затруднений.

2. Наилучший и наихудший случаи

Предположим, что $Y \in A_{k,m}$, т. е. в бинарной цепочке длины n имеется k единиц и m нулей ($k + m = n$). Пусть $a, b \in \{1, \dots, n\}$, $b \geq a$, — границы отрезка внутри цепочки Y . Оптимальным является тот отрезок $[a, b]$, для которого $S_{[a,b]}(Y) = S(Y)$.

Лемма 1. Оптимальный отрезок начинается и заканчивается единицами.

Доказательство. Если это не так, то, исключая крайний 0 из отрезка, мы уменьшим число ошибок, что противоречит его оптимальности. ■

Далее оценим наибольшее возможное число ошибок в классе цепочек $A_{k,m}$:

$$\bar{S}(A_{k,m}) = \max_{Y \in A_{k,m}} S(Y).$$

Для любой цепочки $Y \in A_{k,m}$ справедливо $0 \leq S(Y) \leq \bar{S}(A_{k,m})$. Цепочки, для которых $S(Y) = 0$, будем называть наилучшими, а те, для которых $S(Y)$ максимально — наихудшими в классе $A_{k,m}$.

Очевидно, что наилучшими в $A_{k,m}$ являются те и только те цепочки, в которых все k единиц идут подряд. Поэтому число наилучших цепочек равно $m + 1$.

При определении количества наихудших цепочек важно различать случаи достаточно большого и достаточно малого количества единиц. Назовём цепочку насыщенную, если $k \geq m + 1$, и ненасыщенной иначе.

Лемма 2. В насыщенном случае $\bar{S}(A_{k,m}) = m$.

Доказательство. Нетрудно заметить, что $S_{[1,n]}(Y) = m$ для любой из рассматриваемых цепочек, откуда всегда $S(Y) \leq m$ и, следовательно, $\bar{S}(A_{k,m}) \leq m$. Рассмотрим цепочку, начинающуюся с 1, и такую, что следом за каждым нулём идёт единица (такие существуют в силу определения насыщенного случая). Начнём с отрезка $[1, 1]$ и будем расширять его вправо так, чтобы при включении в него очередного нуля вместе с ним добавлялась бы и идущая следом единица. Тогда число ошибок в расширенном отрезке не станет больше, чем до расширения. Поэтому минимальное число ошибок достигается на наиболее широком отрезке:

$$S(Y) = S_{[1,n]}(Y) = m.$$

Итак, верхняя граница числа ошибок достигается. ■

Лемма 3. Для ненасыщенного случая $\bar{S}(A_{k,m}) = k - 1$.

Доказательство. В произвольной цепочке Y выберем j так, чтобы $y_j = 1$. Получаем оценку $S(Y) \leq S_{[j,j]}(Y) = k - 1$. Рассмотрим далее такую цепочку Y , в которой имеется подцепочка длины $2k - 1$ вида 1010...01, где между любыми единицами размещается ровно один ноль. В этой подцепочке содержатся все единицы данной цепочки. Возьмём произвольный отрезок данной цепочки, начинающийся и заканчивающийся единицей (по лемме 1 оптимальный отрезок содержится среди таких). Пусть выбранный отрезок содержит s единиц и, следовательно, $(s - 1)$ ноль. Тогда для него $S_{[a,b]}(Y) = (k - s) + (s - 1) = k - 1$, а значит, таково это число и для оптимального отрезка. Оценка сверху для $S(Y)$ достигнута. ■

Таким образом, получено иное доказательство результата из [3]:

Теорема 1.

$$\bar{S}(A_{k,m}) = \begin{cases} k - 1, & k < m + 1, \\ m, & k \geq m + 1. \end{cases} \quad (4)$$

3. Строение и количество наихудших цепочек

Для наилучших цепочек вопрос об их количестве и устройстве оказался простым. Для наихудших разберём отдельно ненасыщенный и насыщенный случаи. Пусть сначала $k < m + 1$ (ненасыщенный случай). Обозначим через W_1 множество таких цепочек, в которых никакие две единицы не расположены рядом.

Теорема 2. Множество наихудших ненасыщенных цепочек совпадает с W_1 .

Доказательство. Пусть $Y \in W_1$, j — номер позиции любой единицы в ней, находящейся внутри оптимального отрезка. Заметим, что $S_{[j,j]}(Y) = k - 1$. При расширении отрезка $[j, j]$ в любую сторону вместе с каждой единицей в него будет добавляться хотя бы один нуль (добавление только одного нуля запретим, ориентируясь на лемму 1). Таким образом, при расширении отрезка число ошибок не уменьшается, а значит, $S(Y) \geq k - 1$. Из леммы 3 следует, что строгое неравенство здесь невозможно, и Y является наихудшей.

Обратно, пусть Y — наихудшая цепочка. Если она не принадлежит W_1 , то в ней есть две единицы, расположенные на соседних позициях. Пусть номера этих позиций j и $j + 1$. Тогда

$$S(Y) \leq S_{[j,j+1]}(Y) = k - 2 < \bar{S}(A_{k,m}),$$

а значит, эта цепочка не наихудшая. ■

Следствие 1. Имеется ровно C_{m+1}^k наихудших ненасыщенных цепочек.

Доказательство. Согласно теореме 2, это количество равно числу способов расставить в ряд k единиц и m нулей так, чтобы никакие две единицы не располагались рядом. Решение такой задачи широко известно и приведено, например, в [5]: расставим m нулей в ряд. Между ними и по краям образуется $m + 1$ место для возможного расположения k единиц. Выбирая положение для каждой из них, что можно сделать C_{m+1}^k способами, переберём все существующие возможности. ■

Перейдём к рассмотрению насыщенного случая: $k \geq m + 1$. Обозначим через W_2 множество цепочек такого типа, обладающих следующим свойством: для каждого нуля и слева, и справа от него число единиц строго больше, чем число расположенных с этой же стороны нулей.

Теорема 3. Множество наихудших насыщенных цепочек совпадает с W_2 .

Доказательство. Возьмём $Y \in W_2$. Выберем границы a, b произвольно. При расширении выбранного отрезка до $[1, b]$ в него добавится больше единиц, чем нулей, а значит, число ошибок может лишь уменьшиться. Те же соображения работают при расширении отрезка с произвольными границами вправо, поэтому

$$m = S_{[1,n]}(Y) \leq S_{[1,b]}(Y) \leq S_{[a,b]}(Y).$$

Таким образом, и для оптимального отрезка число ошибок не меньше m . Согласно лемме 2, строгое неравенство невозможно, значит, $S(Y) = m = \bar{S}(A_{k,m})$, т. е. рассматриваемая цепочка является наихудшей.

Обратно, пусть Y наихудшая. Тогда она начинается с единицы, так как иначе

$$S(Y) \leq S_{[2,n]}(Y) = m - 1 < \bar{S}(A_{k,m}).$$

Аналогично доказывается, что и заканчиваться такая цепочка должна единицей. Если $Y \notin W_2$, то в ней, например, найдётся некоторый нуль, справа от которого единиц не больше, чем нулей. Обозначим номер позиции этого нуля j , число единиц справа от него t , число нулей с той же стороны q . Ясно, что

$$S(Y) \leq S_{[1,j-1]}(Y) = m - (1 + q) + t \leq m - 1,$$

так как вне указанного отрезка $1 + q$ нулей, а значит, внутри него $m - (1 + q)$ нулей. Полученное неравенство в силу леммы 2 доказывает, что рассматриваемая цепочка

не может быть наихудшей. Случай нарушения соотношения нулей и единиц слева разбирается аналогично. ■

В насыщенном случае структура множества наихудших цепочек существенно сложнее. Для подсчёта их количества и последовательной генерации всех таких цепочек предлагается следующий способ.

Изобразим на координатной плоскости прямоугольник из k клеток по горизонтали и m клеток по вертикали. Построение каждой наихудшей цепочки насыщенного типа соответствует движению по точкам с целыми координатами в этом прямоугольнике, которое начинается в его левом нижнем углу (0 единиц и 0 нулей) и заканчивается в верхнем правом (k единиц и m нулей).

Каждый шаг ломаной вправо означает появление в цепочке единицы, а шаг вверх — нуля. Условия, которые обеспечивают попадание цепочки в множество W_2 , выясним следующим образом. Обозначим через (x, y) текущие координаты точки, движущейся по целочисленным узлам прямоугольника. Рассмотрим момент, когда появился очередной нуль (y увеличился на 1). Здесь имеется $y - 1$ нуль и x единиц слева от нового нуля, поэтому из теоремы 3 вытекает ограничение $y - 1 < x$. Условие на соотношение нулей и единиц справа имеет вид $m - y < k - x$. Выписанное строгое неравенство должно выполняться только при появлении на текущем шаге нуля, а значит, на предыдущем шаге вместо него могло быть выполнено равенство. Но поскольку после появления очередного нуля y увеличивается лишь на единицу, то противоположное неравенство ($m - y > k - x$) ни для одной точки ломаной недопустимо.

Окончательно, ограничения на движение точки, траектория которой соответствует цепочке из W_2 , записываются в виде системы неравенств

$$\begin{cases} y < x + 1, \\ y \geq x + m - k. \end{cases} \quad (5)$$

Следующий алгоритм позволяет вычислить количество наихудших насыщенных цепочек по заданным k и m .

Алгоритм 1. Расчёт числа наихудших цепочек в $A_{k,m}$

- 1: Проверка условий и подготовка к работе. Если $k \leq m$, то отказ. Иначе формируем массив H размера $(m + 1) \times (k + 1)$, нумерации строк и столбцов которого начинаются с 0. Заполняем его нулями. Для каждого $x \leq k - m$ полагаем $H(x, 0) = 1$.
 - 2: **Для** $y = 1, \dots, m$
 - 3: **Для** $x = y, y + 1, \dots, y + k - m$
 $H(x, y) := H(x - 1, y) + H(x, y - 1)$.
 - 4: **Вернуть** $H(k, m)$.
-

Графически работа алгоритма 1 при $k = 8, m = 5$ представлена на рис. 1.

Выделенная на рис. 1 линия соответствует одной из наихудших насыщенных цепочек (1011010010111). Светлая полоса на чертеже показывает границы, за которые траектории движения заходить не могут (правда, нижнего края полосы они могут касаться). На рисунке видно, что искомое число наихудших цепочек $H(8, 5) = 144$.

0	0	0	0	0	34	89	144	(144)
0	0	0	0	13	34	55	55	0
0	0	0	5	13	21	21	0	0
0	0	2	5	8	8	0	0	0
0	1	2	3	3	0	0	0	0
1	1	1	1	0	0	0	0	0

Рис. 1. Графическая иллюстрация алгоритма 1

4. Коэффициент L_E и алгоритм его расчёта

В работе [3] введён коэффициент, который по заданной цепочке нулей и единиц позволяет оценивать силу связи типа «ступенька»:

$$L_E(X, Y) = 1 - \frac{S(Y)}{\bar{S}(A_{k,m})}, \quad (6)$$

где $\bar{S}(A_{k,m})$ определяется формулой (4), а $S(Y)$ — формулой (3) для цепочки Y после её упорядочения по неубыванию соответствующих элементов X . Будем называть его ледж-коэффициентом (от ledge — ступенька).

В силу доказанных лемм $0 \leq L_E(X, Y) \leq 1$, и его возрастание возможно трактовать как увеличение степени связи соответствующих показателей.

Одним из недостатков, мешающих широкому применению ледж-коэффициента, является трудность вычисления $S(Y)$ для конкретной цепочки Y . Следующая очевидная лемма даёт возможность обосновать приводимый далее алгоритм этого вычисления. Введя обозначения $k(a, b)$, $m(a, b)$ соответственно для количеств единиц и нулей внутри отрезка $[a, b]$, рассмотрим

$$Q[a, b] = k(a, b) - m(a, b).$$

Лемма 4.

- 1) Если $a < b < c$, то $Q[a, c] = Q[a, b - 1] + Q[b, c]$.
- 2) $S_{[a,b]}(Y) = k - Q[a, b]$. В частности, отрезок $[a_0, b_0]$ оптимален тогда и только тогда, когда $Q[a_0, b_0] = \max Q[a, b]$.

Алгоритм 2. Вычисление ледж-коэффициента

- 1: Под каждым из элементов цепочки нулей и единиц подпишем число $q(j)$ по следующему правилу. Пусть номер очередного элемента цепочки j . Если этот элемент равен 0 и единиц до него не встречалось, то $q(j) = 0$. Для первой из единиц в цепочке $q(j) = 1$, для всех последующих $q(j) = q(j - 1) + 1$. Если элемент равен 0 и перед ним были единицы, то полагаем

$$q(j) = \begin{cases} q(j - 1) - 1, & q(j - 1) \neq 0, \\ 0, & q(j - 1) = 0. \end{cases}$$

- 2: Находим $q = \max_{j=1, \dots, n} q(j)$. Тогда $S(Y) = k - q$. Вычислим $\bar{S}(A_{k,m})$ по формуле (4) и $L_E(X, Y)$ согласно (6).
-

Число $q(j)$ представляет собой величину $Q[i, j]$, где i — ближайшая слева от j граница, такая, что отрезок $[i, j]$ начинается с 1 и не содержит внутри себя ни одного отрезка $[a, b]$ с условием $Q[a, b] = 0$. Ясно, что если отрезок $[i, j]$ оптимален, то $q(i) = 1$, $q(j) = q$, а если $q(j) = 0$, то обязательно $y_j = 0$.

Например, для цепочки на рис. 1 значения чисел $q(j)$ получаются равными 1; 0; 1; 2; 1; 2; 1; 0; 1; 0; 1; 2; 3. Максимальное $q = 3$, откуда $S(Y) = 8 - 3 = 5$. При этом цепочка является насыщенной, следовательно, $\bar{S}(A_{k,m})$ равно числу нулей в соответствии с (4) и $L_E(X, Y) = 0$, что подтверждает тот факт, что цепочка наихудшая. Коэффициент корреляции ρ между значениями Y и их рангами для такой цепочки равен 0,127. Это, как и при использовании нашего метода, может быть интерпретировано как отсутствие связи. Те же численные значения и выводы получаются и при расчёте коэффициента бисериальной корреляции. Если, учитывая непараметрический характер изучаемой связи, привлечь к исследованию τ -коэффициент Кендалла, то его значение (0,56) показывает здесь, напротив, наличие статистически значимой ранговой связи. В то же время для наилучшей цепочки 0011111000 ($L_E = 1$) значение ρ равно $-0,17$, а $\tau = 0,45$. Следовательно, оба эти коэффициента указывают на отсутствие значимой связи между показателями. Это даёт повод констатировать некорректность выводов классических методик в рассматриваемой ситуации.

5. Поиск оптимальных отрезков

Оптимальных отрезков внутри изучаемой цепочки может быть несколько. Непересекающиеся оптимальные отрезки следует, вероятно, рассматривать, как до определённой степени парадоксальную ситуацию с точки зрения медицинских приложений, хотя чисто математически исключить такую возможность нельзя. Тем не менее случай, когда имеются два вложенных друг в друга оптимальных отрезка, является довольно обычным.

При этом возникает вопрос, какой из двух отрезков выбрать — больший или меньший? Конечно, всё зависит от решаемой задачи. В медицинских задачах, где значения $Y = 1$ интерпретируются как признак здоровья человека, интересен наиболее короткий отрезок, если нужно выбрать тех пациентов, которые наверняка здоровы. Если же значение 1 означает наличие некоторого опасного заболевания, то следует предпочесть более длинный отрезок, чтобы оставить за его пределами как можно меньше больных. Понятно, что приведённые рассуждения близки к дискуссии о том, какую из двух возможных ошибок, первого или второго рода, следует считать более серьёзной при проверке статистических гипотез. Отметив это, решим далее обе поставленные задачи — приведём алгоритмы поиска самого короткого и самого длинного из оптимальных отрезков. Сначала докажем несколько утверждений о строении набора всех оптимальных отрезков в данной бинарной цепочке.

Лемма 5. Если пересечение двух оптимальных отрезков не пусто, то оно само является оптимальным отрезком.

Доказательство. Пусть $a < c < b < d$ и отрезки $[a, b]$ и $[c, d]$ оба оптимальны. Если $Q[a, c - 1] < 0$, то из утверждения 1 леммы 4 получим $Q[c, b] > Q[a, b]$. Это противоречит оптимальности $[a, b]$. Более того, случай $Q[a, c - 1] > 0$ также невозможен, иначе $Q[a, d] = Q[a, c - 1] + Q[c, d] > Q[c, d]$, а значит, не оптимален отрезок $[c, d]$. При помощи этих рассуждений убеждаемся, что $Q[a, c - 1] = 0$, откуда по лемме 4 получаем $Q[c, b] = Q[a, b] - Q[a, c - 1] = Q[a, b]$, что означает оптимальность пересечения $[c, b]$. ■

Из леммы 5 следует, что можно построить систему попарно непересекающихся минимальных по длине оптимальных отрезков Ψ , такую, что каждый оптимальный отрезок изучаемой цепочки содержит внутри себя хотя бы один из отрезков системы Ψ . Поскольку ни один из отрезков системы Ψ не содержит внутри себя других оптимальных отрезков, будем называть их атомарными. Отрезок $[a, b]$ назовём нейтральным, если $Q[a, b] = 0$.

Утверждение 1. Произвольный неатомарный оптимальный отрезок получается из некоторого атомарного добавлением к нему одного или двух (с разных сторон) нейтральных отрезков.

Доказательство. Пусть $[c, d]$ — оптимальный неатомарный отрезок. Внутри него расположен $[a, b] \in \Psi$. Пусть, например, $c < a$ (добавлен левый отрезок). Тогда нейтральность добавленного отрезка немедленно вытекает из соотношения

$$Q[a, b] = Q[c, b] = Q[c, a - 1] + Q[a, b],$$

справедливого в силу лемм 4 и 5. Поскольку рассуждения для добавляемого правого отрезка полностью аналогичны, утверждение доказано. ■

Следовательно, укоротить оптимальный отрезок можно, лишь отыскав нейтральный, расположенный внутри него и примыкающий к одному из его краев, что соответствует появлению 0 в наборе соответствующих чисел $q(j)$.

Алгоритм 3. Поиск самого короткого оптимального отрезка

- 1: Выполняем первый шаг алгоритма 2.
 - 2: Пусть $q = \max_{j=1, \dots, n} q(j)$. Проходим набор чисел $q(j)$ слева направо, находя все отрезки, начинающиеся с 1 и заканчивающиеся q , такие, что внутри них нет нулей.
 - 3: Если на шаге 2 найдено несколько отрезков, то находим из них тот, что имеет наименьшую длину.
-

Проиллюстрируем работу алгоритма 3 следующей таблицей ($k = 12, m = 10$). По третьей строке таблицы видим, что $q = 4$. По второму утверждению леммы 4 число ошибок для оптимального отрезка $S(Y) = k - q = 8$. Интервалы, обнаруживаемые на втором шаге алгоритма, — $[5, 10]$, $[5, 14]$ и $[19, 22]$. Наименьшую длину имеет последний из них. Атомарными являются отрезки $[5, 10]$ и $[19, 22]$.

X	01	02	03	04	05	06	07	08	09	10	11
Y	0	0	1	0	1	1	1	0	1	1	0
$q(j)$	0	0	1	0	1	2	3	2	3	4	3
X	12	13	14	15	16	17	18	19	20	21	22
Y	0	1	1	0	0	0	0	1	1	1	1
$q(j)$	2	3	4	3	2	1	0	1	2	3	4

Утверждение 2. Если для некоторых $j_0 < j_1$ верно $q(j) = 0$, $j_0 \leq j \leq j_1$, то отрезок $[j_0, j_1]$ не содержится ни в каком оптимальном.

Доказательство. Предположим противное. Пусть нашлись такие индексы i, j , что $[j_0, j_1] \subset [i, j]$, и отрезок $[i, j]$ оптимален. Заметим, что $y_{j_0} = y_{j_1} = 0$. Поэтому включение отрезков является строгим, так как по лемме 1 $y_i = y_j = 1$. Без ограничения общности можно считать, что $[j_0, j_1]$ — максимальный по включению отрезок, внутри

которого все числа $q(j)$ равны 0, что по определению этих чисел означает $q(j_0 - 1) = q(j_1 + 1) = 1$. Ясно, что тогда отрезок $[i, j_0]$ нейтрален, $Q[j_0 + 1, j_1] < 0$. В силу оптимальности $[i, j]$ выполнено $q(j) = q$, а значит, $Q[i, j] = Q[i, j_0] + Q[j_0 + 1, j_1] + Q[j_1 + 1, j] < q$, что противоречит оптимальности $[i, j]$. ■

Утверждение 2 означает, что в наборе чисел $q(j)$, когда j пробегает оптимальный отрезок, два нуля не могут следовать подряд, что обосновывает алгоритм 4.

Алгоритм 4. Поиск самого длинного оптимального отрезка

- 1: Выполняем первый шаг алгоритма 2.
 - 2: Пусть $q = \max_{j=1, \dots, n} q(j)$. Проходим набор чисел $q(j)$ слева направо, находя все отрезки, начинающиеся с 1 и заканчивающиеся q , такие, что внутри них содержится не более одного нуля подряд.
 - 3: Среди всех обнаруженных на шаге 2 отрезков выбираем тот, который имеет наибольшую длину.
-

В примере (см. таблицу) на втором шаге будут обнаружены отрезки $[3, 10]$, $[3, 14]$, $[3, 22]$, $[5, 10]$, $[5, 14]$, $[5, 22]$, $[17, 22]$ и $[19, 22]$. Ответ — $[3, 22]$.

Заключение

Рассмотрен новый вид связи («ступенька») между номинальным и бинарным показателями. Он является естественным и часто встречается в приложениях, особенно медицинских, но описание и анализ связей такого вида на основе классических методов не являются вполне удовлетворительными и корректными по причине отсутствия информации о распределении выборок.

С помощью предлагаемых алгоритмов может быть оценена сила связи типа «ступенька» и найдены границы, в которых должен изменяться числовой показатель, чтобы оказаться наиболее сильно связанным с бинарным. Эти алгоритмы делают расчёт ледж-коэффициента L_E , а также поиск нормативных интервалов изучаемых показателей в предположении, что они неизвестны, доступными даже при работе без привлечения компьютера. Ранее (в работе [3]) предлагалось производить эти вычисления методом полного перебора, что предполагало написание специальной компьютерной программы (особенно при объёмах выборок больше 20).

Можно рекомендовать изучаемый в работе ледж-коэффициент для использования оценки силы связи показателей описанных типов как альтернативу традиционно используемым разного рода коэффициентам корреляции. Он также может быть полезен, когда вместо номинального показателя используется порядковая переменная.

Результаты работы можно также использовать для оценки качества разбиения объектов на два кластера [6].

ЛИТЕРАТУРА

1. *Straus Sh. E., Richardson W. S., Glasziou P., and Haynes R. B.* Evidence Based Medicine. Elsevier Churchill Livingstone, 2011. 295 p.
2. *Дударева Ю. А., Гурьева В. А., Дронов С. В., Шойхет Я. Н.* Сравнительная оценка состояния здоровья потомков лиц, проживавших на территории с неблагоприятной экологической обстановкой // Клиническая медицина. 2014. Т. 92. № 5. С. 66–70.
3. *Дронов С. В., Петухова Р. В.* Один вид связи между номинальной и бинарной переменными // Известия АлтГУ. 2010. Т. 65. № 1/2. С. 34–36.

4. *Мирмоминов Р. М.* Один вид связи между номинальной и бинарной переменными // Сб. научных статей Междунар. конф. «Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования». Барнаул, 11–14 ноября 2014. Барнаул: Изд-во АлтГУ, 2014. С. 171–173.
5. *Виленкин Н. Я., Виленкин А. Н., Виленкин П. А.* Комбинаторика. М.: ФИМА, МЦНМО, 2006. 400 с.
6. *Dronov S. V. and Dementjeva E. A.* A new approach to post-hoc problem in cluster analysis // Model Assisted Statistics and Applications. 2012. V. 7. No. 1. P. 49–55.

REFERENCES

1. *Straus Sh. E., Richardson W. S., Glasziou P., and Haynes R. B.* Evidence Based Medicine. Elsevier Churchill Livingstone, 2011. 295 p.
2. *Dudareva Yu. A., Gur'eva V. A., Dronov S. V., Shoykhet Ya. N.* Sravnitel'naya otsenka sostoyaniya zdorov'ya potomkov lits, prozhivavshikh na territorii s neblagopriyatnoy ekologicheskoy obstanovkoy [Comparative evaluation of the health status of descendants of people residing in environmentally unfriendly regions by means of mathematical modeling]. Klinicheskaya meditsina, 2014, vol. 92, no 5, pp. 66–70. (in Russian)
3. *Dronov S. V., Petukhova R. V.* Odin vid svyazi mezhd nominal'noy i binarnoy peremennymi [One type of connection between nominal and binary variables]. Izvestiya AltSU, 2010, vol. 65, no. 1/2, pp. 34–36. (in Russian)
4. *Mirmominov R. M.* Odin vid svyazi mezhd nominal'noy i binarnoy peremennymi [One type of connection between nominal and binary variables]. Proc. conf. «Lomonosovskie chteniya na Altae: fundamental'nye problemy nauki i obrazovaniya». Barnaul, 11–14 November 2014. Barnaul, AltSU Publ., 2014, pp. 171–173. (in Russian)
5. *Vilenkin N. Ya., Vilenkin A. N., Vilenkin P. A.* Kombinatorika [Combinatorics]. Moscow, MCCME Publ., 2006. 400 p. (in Russian)
6. *Dronov S. V. and Dementjeva E. A.* A new approach to post-hoc problem in cluster analysis // Model Assisted Statistics and Applications, 2012, vol. 7, no. 1, pp. 49–55.