УДК 519.235 + 519.237.5

С.В. Дронов, Р.В. Петухова

Один вид связи между номинальной и бинарной переменными

S.V. Dronov, R.V. Petukhova

One Type of Connection between Nominal and Binary Variables

Мы полагаем, что естественный вид связи между переменными x и y упомянутых типов имеет вид индикатора некоторого отрезка, т.е. $y = 1$ тогда и только тогда, когда x лежит между какими-то границами a и b. На основе высказанного предположения в работе определен новый коэффициент, характеризующий силу этой связи.

Ключевые слова: дихотомическая переменная, коэффициент корреляции, стохастическая связь.

We propose that natural type of connection between variables x and y is a dependence having form of some interval indicator, i.e. $y = 1$ if x lies between some boundaries a and b. From this point of view we define a new dependence coefficient, characterizing a strength of this connection.

Key words: Dychotomous variable, correlation coefficient, stochastic dependence.

В последние десятилетия проблема установления и характеризации силы связи бинарной (дихотомической) и номинальной (числовой) переменных приобрела большую актуальность. Это связано прежде всего со все большей востребованностью статистических методов со стороны медицины, в которой многие исследуемые показатели имеют дихотомический характер (человек здоров или болен, анализ проводился или нет, симптом имеется или отсутствует (см., например: [1]). С другой стороны, в решении этой же задачи заинтересовано такое бурно развивающееся направление педагогической науки, как тестология. Наиболее ярко это проявляется при выявлении взаимной зависимости общего балла, набранного испытуемым при прохождении теста и факта верного выполнения им конкретного задания (см.: [2, 3]).

В цитированных работах для исследования рассматриваемого типа зависимости применялись различные виды коэффициентов корреляции. Это коэффициенты Пирсона, бисериальный коэффициент, а также коэффициент φ. Между тем любому специалисту, знакомому с теорией вероятностей, известно, что коэффициент корреляции Пирсона адекватно работает лишь для установления линейных связей, а другие два упомянутых вида коэффициентов представляют собой тот же коэффициент Пирсона, формула которого переписана на случай, когда одна или обе исследуемые переменные имеют бинарный тип.

Общепринятый способ оценки адекватности предлагаемого описания статистической связи между двумя переменными x и y в виде $y = f(x)$ состоит в нанесении на поле корреляции этих переменных графика $f(x)$ с последующей оценкой отклонения выборочных точек от этого графика. Но в случае, когда ординаты всех выборочных точек могут принимать значения лишь 0 или 1, а значения абсцисс достаточно разнообразны, эти точки не могут удовлетворительно группироваться вокруг какой-либо прямой линии – графика линейной зависимости, – а следовательно, ни один из упомянутых выше коэффициентов корреляции не может служить адекватной мерой степени связи между числовой и бинарной переменными. Наша цель – предложить новый коэффициент, который с большей достоверностью позволяет оценивать степень такой связи.

Анализ рассмотренных выше бинарных переменных в медицинской практике наводит на мысль о том, что правильная форма изучаемой зависимости имеет вид индикатора некоторого отрезка – пока показатель x находится в определенных границах, пациент здоров ($y = 1$), иначе – болен. Почти такой же вид имеет зависимость и в тестологии. Действительно, если балл, набранный испытуемым по тесту максимален, то он обязан был правильно выполнить изучаемое задание, и если связь между баллом и заданием существует, то чем больше набранный им балл,

тем вероятнее появление признака верного решения напротив номера этого задания. Тем самым и в этом случае мы имеем зависимость в виде индикатора отрезка, но только правая его граница здесь совпадает с максимально возможным числом набранных баллов (правая ступенька).

Перейдем к формулам. Наблюдаются две связанных выборки объема n. В одной из них собраны числовые значения $x_1, ..., x_n$ – наблюдения над номинальной переменной x, другая (Y) содержит в своем составе только числа 0 или 1. Ставится задача подобрать числа a, b из набора значений переменной x так, чтобы функция

$$y(x, a, b) = \begin{cases} 1, & x \in [a, b]; \\ 0 & \text{иначе} \end{cases} \qquad (1)$$

наилучшим образом описывала связь между x и y среди всех функций такого вида. Критерием качества зависимости (1) будет служить величина

$$S(a, b) = \sum_{j=1}^{n} (y_i - y(x_i, a, b))^2, \qquad (2)$$

равная числу ошибок формулы предлагаемой связи среди выборочных данных. Пару (a^*, b^*), на которой достигается минимум рассматриваемого критерия, назовем оптимальной, а значение $S(a^*, b^*)$ – наименьшим числом ошибок (или числом ошибок наилучшей аппроксимации) для выборки Y.

Найти наименьшее число ошибок при небольших объемах выборок можно, например, полным перебором. Для ускорения процесса рекомендуется начать с самой длинной цепочки идущих подряд единиц, выбрать a, b равными границам этого участка и раздвигать его границы до достижения искомого максимума.

Далее предположим, что выборки согласованным образом упорядочены по возрастанию элементов X. Отметим также, что настоящие числовые значения X для нас не имеют значения, поэтому присвоим им порядковые номера по возрастанию и будем далее без ограничения общности считать, что $x_i = i, i = 1, ..., n$. После того как наименьшее число ошибок для рассматриваемой выборки Y найдено, сравним его с наихудшей возможной ситуацией при фиксированных количествах нулей и единиц. Идея состоит в том, что наилучшее с точки зрения (1) расположение нулей и единиц в выборке Y состоит в расположении всех единиц подряд, – тогда число ошибок наилучшей аппроксимации равно нулю. Насколько велико может быть число ошибок при максимально неблагоприятном расположении такого же, как в Y, количества единиц?

Пусть при наблюдении дихотомической переменной мы получили некоторое расположение k единиц и m нулей. Назовем это расположение минимаксным, если самая длинная цепочка единиц, идущих подряд в нем, будет наиболее короткой среди всех возможных расположений. Длину наибольшей цепочки единиц в минимаксном расположении обозначим $M(k, m)$ и будем называть фатально неизбежной длиной. Таким образом, цепочка единиц не меньше, чем фатально неизбежной длины, встретится при любом расположении k единиц и m нулей, и если $z > M(k, m)$, то среди всех возможных расположений обязательно найдется такое, что все цепочки единиц в ней имеют длину менее z.

Теорема 1. *Число $M(k, m)$ равно наименьшему натуральному числу, которое больше либо равно дроби $t = k/(m + 1)$, т.е.*

$$M(k, m) \;=\; \begin{cases} t, & t \text{ целое}, \\ [t] + 1 & \text{иначе}. \end{cases}$$

Теорема 2. *Минимаксное расположение k единиц и m нулей в цепочку имеет вид* 1...101...10...01...10...0, *где более одного нуля подряд может встретиться лишь в конце всей цепочки. При этом длина первой цепочки единиц равна $M(k, m)$, а длины всех последующих либо такие же, либо на единицу меньше. Если $M(k, m) \neq 1$, то минимаксное расположение оканчивается на* 1.

Эти теоремы нетрудно доказать методом математической индукции по числу единиц. Займемся теперь поиском такого значения $S_{k,m}$, которое является максимальным количеством ошибок наилучшей аппроксимации по всем возможным цепочкам из k единиц и m нулей. При этом сохраним обозначение (2), подразумевая под y_i элементы рассматриваемой цепочки, а x_i полагая равными i.

Если $k < m + 1$, то цепочку назовем ненасыщенной, иначе – насыщенной. Ненасыщенную цепочку назовем наименее благоприятной, если между любыми двумя единицами расположен хотя бы один ноль. Заметим, что цепочка является ненасыщенной в том и только том случае, если $M(k, m) = 1$.

Лемма 1. *Если цепочка ненасыщена, то $S_{k,m} = k - 1$ и достигается на любой из наименее благоприятных цепочек, в частности, на минимаксной.*

Доказательство. Возьмем $a = b$ равными любому из номеров мест, на котором в цепочке находится 1. Тогда, очевидно, $S(a, b) = k - 1$. Отсюда на любой такой цепочке $S(a^*, b^*) \leq k-1$,

а значит $S_{k,m} \leq k-1$. Покажем, что для наименее благоприятной цепочки $S(a^*, b^*) = k-1$. Действительно, выберем a, b как в уже проведенной части рассуждения и попытаемся расширить этот отрезок. Поскольку цепочка наименее благоприятная, то прежде, чем добавить в $[a, b]$ единицу и уменьшить тем самым S, мы будем вынуждены добавить хотя бы один ноль. Это означает, что число ошибок может лишь увеличиться. Поскольку интервалам, вообще не содержащим единиц, очевидно, соответствуют еще большие S, то утверждение доказано.

Лемма 2. *Если цепочка насыщена, то $S_{k,m} = m$ и достигается на минимаксной цепочке.*

Доказательство. Так же, как при доказательстве леммы 1, убедимся, что $S_{k,m} \leq m$, поскольку в рассматриваемом случае $S(1, n) = m$. Теперь проверим, что для минимаксной цепочки $S(a^*, b^*) = m$. Возьмем какие-то a, b. Если слева или справа от интервала $[a, b]$ граница цепочки еще не достигнута и следующим символом является 1, то можно расширить интервал в эту сторону, уменьшив S. Итак, без ограничения общности, можно считать, что интервал $[a, b]$ ограничен нулями или границами цепочки. В силу структуры минимаксной цепочки за каждым из нулей находится по меньшей мере одна единица. Отсюда следует, что, расширив отрезок на два шага в сторону имеющегося нуля, мы не изменим величину S. Ясно, что этот процесс можно продолжить, не увеличивая S(но, возможно, уменьшая ее), до момента достижения границ цепочки. Итак, доказано, что на минимаксной цепочке наименьшее из возможных значений $S(a, b)$ достигается при выборе $a = 1, b = n$, а значит, равно m. Лемма доказана.

Из доказанных двух лемм следует, что

$$S_{k,m} = \begin{cases} k-1, & k \leq m+1; \\ m, & k > m+1. \end{cases}$$

Теперь все готово для того, чтобы ввести тот коэффициент, который будет характеризовать степень связи (1). Этот коэффициент мы назвали коэффициентом эминентильности, поскольку eminentia в переводе с латинского означает выступ, что соответствует характеру изучаемой связи. Пусть X, Y – выборки объема $n = k + m$, причем Y состоит из k единиц и m нулей. Тогда коэффициентом эминентильности между X и Y назовем число

$$E(X, Y) = 1 - \frac{S(a^*, b^*)}{S_{k,m}}.$$

Утверждение следующей теоремы вытекает из уже проведенных рассуждений.

Теорема 3. $0 \leq E(X, Y) \leq 1$, *причем ноль достигается на минимаксной цепочке, а единица лишь в том случае, если при некоторых a, b справедливо $y_i = y(x_i, a, b)$ при всех $i = 1, ..., n$.*

Из определения коэффициента эминентильности следует, что величина этого коэффициента тем больше, чем лучше зависимость (1) описывает имеющиеся данные. Возможность же применения введенного коэффициента на практике была нами проверена на реальных данных рентгенографического обследования пациентов Алтайского пульмонологического центра. В качестве числовой переменной рассматривалось количество лейкоцитов в анализе крови пациентов, в качестве бинарной – наличие или отсутствие заболевания легких. Предложенный коэффициент показал наличие умеренной связи между этими величинами ($E \approx 0,39$), тогда как обычные коэффициенты корреляции дают результат около 0,21, что, согласно общепринятым соглашениям, трактуется как отсутствие связи.

Библиографический список

1. Лазарев А.Ф., Шойхет Я.Н., Алексеева И.В., Дронов С.В. Многофакторный анализ при дифференциальной диагностике узловой формы периферического рака легкого // Российский биотерапевтический журнал. – 2009.– №4.
2. Челышкова М.Б. Теория и практика конструирования педагогических тестов: учеб. пособие. – М., 2002.
3. Аванесов В.С. Композиция тестовых заданий. – М., 2002.