УДК 519.25:51-7

Сравнение подходов к оценке степени связи нечисловых факторов в четырехпольных таблицах

С.В. Дронов, С.А. Шепелев

Алтайский государственный университет (Барнаул, Россия)

A Comparison of Approaches to Non-numeric Factors Colligation Evaluation in Fourfold Tables

S.V.Dronov, S.A.Shepelev

Altai State University (Barnaul, Russia)

Рассмотрен набор статистических данных, оформленных в виде так называемой четырехпольной таблицы. В таком виде довольно часто бывают представлены данные наблюдений, связанные с двумя нечисловыми категорированными факторами, каждый из которых имеет по две категории. Особенно часто подобная форма данных используется в медицине, генетике, психологии. Специалисты-практики в этих нематематических областях знания для оценки силы (степени) связи между факторами такого типа используют так называемый коэффициент относительного риска, тогда как в математической статистике более привычно использование для решения подобных задач коэффициента корреляции Пирсона или статистики хи-квадрат. Изучаются соотношения между описанными практическим и теоретическим подходами к оценке силы связи нечисловых факторов упомянутого типа. Показано, что в наиболее распространенных, типичных случаях результаты, получаемые с помощью обоих подходов, совпадают или очень близки, особенно тогда, когда связь между факторами отсутствует. В некоторых необычных случаях (например, когда исходная таблица содержит нули) указан источник различия результатов подходов. Обсуждаются преимущества теоретического подхода в этих необычных ситуациях.

Ключевые слова: четырехпольные таблицы, коэффициент относительного риска, статистика хи-квадрат, медицинские нечисловые данные.

We consider non-numerical data to be organized in a form of a fourfold table. Such form is usual for two categorized factors, and every factor has two categories. In particular, such form of statistical data is widely used in medical, genetical, or psychological problems. Experts of those non-mathematical fields of science evaluate the factors colligation with the so-called relative risk coefficient. However, mathematical statistics operates with Pearson's correlation coefficient or chi-squared statistics. In the paper, we investigate relations between theoretical and practical approaches to non-numeric factors colligation evaluation. It is shown that these methods in the most common cases provide the same (or similar) results, especially, in situations when the presumed colligations do not exist. For some uncommon cases (for example, when the table contains zeroes) we demonstrate the source of the differences and discuss advantages of the theoretical approach in such cases.

Key words: fourfold tables, relative risk, chi-squared, medical data of non-numeric type.

DOI 10.14258/izvasu(2014)1.2-04

Во многих областях науки и практики, особенно в медицине, генетике и психологии [1–4] при изучении взаимодействия двух факторов принято результаты наблюдений представлять в виде таблицы сопряженности. В каждой клетке такой таблицы помещают количество объектов наблюдения, обладающих соответствующей этой клетке сочетанием категорий факторов. В простейшем

случае каждый из факторов имеет по две категории, и получающиеся здесь таблицы называют четырехпольными.

В работе изучаются распространенные среди практиков методы обработки данных, представленных такими таблицами, в частности, способы оценки степени связи между факторами, заданными этим способом. Целью работы является изучение соотношений между различными подходами к оценке степени связи между переменными, применяемыми в случае их задания в виде четырехпольной таблицы. Также нас будет интересовать возможность перевести методы четырехпольных таблиц на язык привычной практикам корреляционной зависимости.

Начнем с описания способа перевода данных наблюдений из одной формы в другую. Пусть сначала есть два ряда чисел $X = (x_1, x_2, ..., x_n)$ и $Y = (y_1, y_2, ..., y_n)$, представляющих собой результаты наблюдений за n объектами, при этом j-й объект обладает характеристиками (x_j, y_j). Построим четырехпольную таблицу, разбив объекты на четыре группы. Для этого возьмем два заданных граничных значения Γ_x и Γ_x по x, y соответственно. В первую группу, которую будем обозначать $\Delta_{1,1}$, войдут объекты, для которых $x_j < \Gamma_x,\ y_j < \Gamma_y$, во вторую ($\Delta_{1,2}$) те, для которых $x_j > \Gamma_x\ y_j < \Gamma_y$, в третью ($\Delta_{2,1}$) – с условиями $x_j < \Gamma_x\ y_j > \Gamma_y$ и в четвертую ($\Delta_{2,2}$) – $x_j > \Gamma_x,\ y_j > \Gamma_y$. Полученную таблицу, в клетках которой вписаны количества объектов, попавших в каждую из групп, запишем в виде

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad a + b + c + d = n, \tag{1}$$

где на (i, j)-м месте стоит число объектов группы с обозначением $\Delta_{i,j}$.

Наоборот, если данные изначально имели вид четырехпольной таблицы (1), то можно перевести их в числовую форму, построив ряды X и Y следующим образом: сначала поместим в эти ряды a пар (1,1), затем b пар (1,0), далее c пар (0,1) и, наконец, d пар (0,0). Конечно же, нули и единицы здесь можно поменять местами. Принятая система обозначений основана на привычном, например, в медицине, представлении, что первая строка и первый столбец таблицы представляют собой наборы данных о количествах объектов, у которых имеется определенный признак, а вторая строка и столбец – о тех, у которых эти признаки отсутствуют.

Рассмотрим два метода оценки зависимости признаков в четырехпольной таблице (см. [1]). Первый принято называть χ^2. Математически применение его основывается на том, что частость каждой из четырех клеток таблицы (1) тем более похожа на произведение частости соответствующих ей категорий столбца и строки, чем менее ярко выражена связь соответствующих признаков. Простые преобразования классической формулы критерия χ^2 (например, [5]) в нашем случае приводят к формуле

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(a + c)(d + c)(b + d)}. \tag{2}$$

В качестве второго метода будем применять коэффициент корреляции ρ, рассчитанный по двум рядам данных, сформированных по четырехпольной таблице описанным выше образом. Для этого способа образования рядов X и Y формула расчета коэффициента корреляции примет вид

$$\rho = \frac{\frac{a}{n} - \frac{a+c}{n} \cdot \frac{a+b}{n}}{\sqrt{\left(\frac{a+c}{n} - \left(\frac{a+c}{n}\right)^2\right) \cdot \left(\frac{a+b}{n} - \left(\frac{a+b}{n}\right)^2\right)}}. \tag{3}$$

Выбор именно этих двух способов для первоначального исследования объясняется тем, что с точки зрения математической статистики в ее классическом варианте именно они представляются наиболее естественными. При этом мы, конечно же, отчетливо понимаем, что способ с использованием χ^2 должен давать более достоверные результаты, если только мы не уверены в том, что изучаемая связь может быть только линейной. С другой стороны, при построении этого критерия существенно используется нормальный характер изучаемых переменных, а данные четырехпольной таблицы по сути сводят все к бинарным переменным, поэтому ценность χ^2 снижается. Попытки же использовать для оценки степени связи коэффициенты Спирмена, бисериальный коэффициент и точечно-бисериальную корреляцию, обычно применяемые практиками, вряд ли могут дать более адекватный результат, чем применение обычного коэффициента ρ, поскольку внимательный анализ формул всех перечисленных коэффициентов показывает их полную тождественность.

Нам потребуется следующее несложно проверяемое утверждение.

Лемма. *Пусть числа $p, q, \alpha \in [0, 1]$, тогда результат действия $\alpha p + (1-\alpha)q$ расположен между числами p, q, в частности, всегда содержится в $[0, 1]$. Если число α не равно ни 0, ни 1, то этот результат не совпадает ни с p, ни с q. В частности, при таком α в результате не могут быть получены ни 0, ни 1.*

Теорема 1. *В произвольной четырехпольной таблице вида (1), ни одна строка и ни один столбец которой не состоят целиком из нулей, максимальное значение статистики χ^2 равно n, причем оно достигается лишь в случае когда $a = d = 0$, либо когда $b = c = 0$.*

Доказательство. Выражение (2), разделив на n, преобразуем к виду

$$\frac{\chi^2}{n} = \frac{a}{a+b}\cdot\frac{a}{a+c} + \left(1-\frac{a}{a+b}\right)\cdot\frac{b}{b+d} + \\ + \frac{c}{c+d}\cdot\frac{c}{a+c} + \left(1-\frac{c}{c+d}\right)\cdot\frac{d}{b+d} - 1.$$

Далее обозначим $\frac{a}{a+b} = \alpha$, $\frac{a}{a+c} = p$, $\frac{b}{b+d} = q$. По лемме, $\alpha p + (1-\alpha)\, q \in [0,1]$.

Ясно, что аналогичное рассуждение можно провести, рассматривая $\beta = \frac{c}{c+d}$, $p = \frac{a}{a+c}$, $q = \frac{d}{d+b}$. Отсюда вытекает, что максимальное значение статистики χ^2/n не превышает 2. Заметим далее, что значение 2 достигается только в том случае, если

$$\begin{cases} \alpha\cdot\frac{a}{a+c} + (1-\alpha)\cdot\frac{b}{b+d} = 1; \\ \beta\cdot\frac{c}{a+c} + (1-\beta)\cdot\frac{d}{b+d} = 1. \end{cases}$$

Здесь отметим, что если $\alpha, \beta \neq 0$ и $\alpha, \beta \neq 1$, то выполнение обоих выписанных равенств возможно лишь в случае, когда $\frac{a}{a+c} = \frac{b}{b+d} = 1$. Получаем противоречие, ведь это в данном случае означает $c = d = 0$, что невозможно в силу условий теоремы. Также проверяется невозможность случая, когда $\frac{c}{a+c} = \frac{d}{b+d} = 1$. Пусть $\alpha \neq 0, \beta = 0$, тогда последняя система выполняется лишь если $\frac{b}{b+d} = 1$ и $\frac{d}{b+d} = 1$, что при сделанных предположениях вновь невозможно.

Следовательно, либо $\alpha = 0$, $\beta = 1$ и $\frac{b}{b+d} = \frac{c}{a+c} = 1$, откуда вытекает $a = d = 0$, либо $\alpha = 1$, $\beta = 0$, а следовательно, $\frac{a}{a+c} = \frac{d}{b+d} = 1$ и, окончательно, $b = c = 0$. Теорема доказана.

Заметим, что

$$an - (a+c)(a+b) = ad - bc,$$
$$n(a+c) - (a+c)^2 = (a+c)(b+d),$$
$$n(a+b) - (a+b)^2 = (a+b)(c+d),$$

следовательно, сравнивая формулы (2) и (3), мы приходим к справедливости следующей теоремы.

Теорема 2. $\chi^2 = n\rho^2$.

Следствие 1. $\chi^2 = 0 \;\Leftrightarrow\; \rho = 0$; $\quad \chi^2 = n \;\Leftrightarrow\; \rho = \pm 1$.

Вторая формула здесь дает независимое подтверждение теореме 1.

Таким образом, при изучении линейных связей есть основания считать два рассмотренных подхода практически эквивалентными, что позволяет объединить их в один, который мы назовем далее теоретическим.

В практических исследованиях, изучающих четырехпольные таблицы, чаще применяется другая характеристика, называемая относительным риском. Вероятно, это связано с простотой ее расчета и понятным смыслом. Формулу для его расчета возьмем из работы [1]:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}. \qquad (4)$$

Если считать, что первая строка четырехпольной таблицы содержит количества, например, пациентов с установленным диагнозом, а вторая — аналогичные данные о контрольной группе практически здоровых людей, то в числителе дроби (4) стоит доля пациентов, обладающих изучаемым фактором среди заболевших, а в знаменателе аналогичная доля в контрольной группе. Поэтому, если две этих доли практически равны (RR — почти единица), то наличие или отсутствие фактора не несет в себе информации о заболевании. Если же коэффициент риска значительно больше или значительно меньше единицы, то фактор либо повышает, либо, соответственно, понижает вероятность заболевания. Подход, связанный с расчетом относительного риска, условимся называть практическим.

Итак, в основе коэффициента относительного риска лежат совсем иные соображения, чем у рассмотренных выше способов. Поэтому прямая зависимость между ними отсутствует. Тем более что в отличие от ρ (ограниченного ± 1) и χ^2 (значения которой по теореме 1 всегда лежат между 0 и n), его значения ничем сверху не ограничены.

Но все же сравнивая определение RR с формулами, использованными в рамках теоретического подхода, немедленно получаем теорему.

Теорема 3.

$$RR - 1 \;=\; \rho\,\frac{\sqrt{(a+b)(a+c)(c+d)(b+d)}}{c(a+b)}.$$

Следствие 2. $RR = 1 \;\Leftrightarrow\; \rho = 0$.

Это означает, что в случае отсутствия связи факторов оба подхода дадут одинаковый результат. Рассмотрим крайние по величинам значения относительного риска. Если $RR = 0$, то $a = 0$, откуда

$$\rho = -\frac{1}{\sqrt{\left(1+\frac{d}{c}\right)\left(1+\frac{d}{b}\right)}}.$$

Если $d \neq 0$, то видно, что коэффициент корреляции не может равняться по модулю единице, что соответствовало бы одинаковым выводом обоих подходов в этом случае. Наибольшее различие выводов получается, если d значительно больше, чем оба числа b, c, – здесь коэффициент корреляции ρ оказывается практически нулевым.

Такое резкое различие выводов, очевидно, связано с тем, что в рассматриваемом сейчас случае во второй строке таблицы данных значительно больше, чем в первой, что указывает на недостаточность данных в ней и некоторый перекос в выборке, а следовательно, ставит под сомнение законность вывода, сделанного на основе RR.

Анализ случая очень большого RR может быть проведен полностью аналогично. Подводя итог,

видим, что в основной массе рядовых случаев и теоретический, и практический подходы дают фактически одинаковые результаты. В частности, отсутствие связи между факторами одинаково хорошо распознается обоими подходами. Для случаев же редких (например, если в таблице есть нули) более надежные результаты дает теоретический способ.

Библиографический список

1. Sasieni P.D. From Genotypes to Genes: Doubling the Sample Size // Biometrics. — 1997. — № 53.

2. Berkson J. Limitation of the Application of Fourfold Tables Analysis to Hospital Data. // Int.J. Epidemol Advance Access. — 2014. — Vol. 10.

3. Давыдов М.И., Шойхет Я.Н., Лазарев А.Ф., Алексеева И.В. Дронов С.В. Многофакторный анализ при дифференциальной диагностике узловой формы периферического рака легкого. Барнаул, 2011.

4. Петриков А.С., Шойхет Я.Н., Белых В.И., Дронов С.В. Многофакторный анализ в диагностике тромбозов глубоких вен нижних конечностей // Тромбоз, гемостаз и реология. — 2013. — №4 (56).

5. Дронов С.В. Многомерный статистический анализ. — Барнаул, 2006.