Записки научных
семинаров ПОМИ
Том 278, 2001 г.

Я. Ю. Никитин, Н. А. Степанова

ПИТМЕНОВСКАЯ ЭФФЕКТИВНОСТЬ КРИТЕРИЕВ НЕЗАВИСИМОСТИ, ОСНОВАННЫХ НА ВЗВЕШЕННЫХ РАНГОВЫХ СТАТИСТИКАХ

1. Введение

Среди различных критериев независимости, предложенных за последние десятилетия и обобщающих классические критерии Кендалла и Спирмена, по идее построения и назначению интересны тесты, основанные на взвешенных ранговых коэффициентах корреляции. Некоторые идеи о применении взвешенных рангов для проверки гипотезы независимости высказывались еще Кендаллом [1]. Одна из наиболее интересных идей построения таких тестов принадлежит Иману и Коноверу [2]: для сравнения различных методов обработки данных в задачах компьютерного моделирования и определения точности прогнозов они предложили выборочный коэффициент корреляции с метками Сэвиджа [3].

Пусть (X_i, Y_i), $i = 1, \dots, n$, – выборка из двумерного распределения с абсолютно непрерывной функцией распределения (ф.р.) $F(x, y)$ и маргинальными ф.р. $G(x)$ и $H(y)$, и пусть R_i обозначает ранг X_i среди элементов $X_1, \dots, X_n$, а Q_i – ранг Y_i среди элементов $Y_1, \dots, Y_n$. Коэффициент Имана и Коновера [2] (так называемый коэффициент "top-down") определяется формулой

$$r_T = \frac{1}{n - S_1} \left(\sum_{i=1}^{n} S_{R_i} S_{Q_i} - n \right), \tag{1}$$

где $S_i = \sum\limits_{j=i}^{n} (1/j)$ – метки Сэвиджа. Цтатистика r_T акцентирует внимание на нескольких наименьших элементах в рядах наблюдений $X_1, \dots, X_n$ и $Y_1, \dots, Y_n$. В случае независимости X и Y

Работа выполнена при частичной поддержке грантов РФФИ 99-01-0724, 00-15-96019 и 01-01-0245.

имеет место сходимость [2]

$$\sqrt{n-1}\, r_T \xrightarrow{d} N(0,1).$$

Замена S_i в определении (1) метками $\hat{S}_i = \sum\limits_{j=i}^{n} 1/(n+1-j)$ приводит к коэффициенту $\hat{r}_T$, принимающему близкие к единице значения при согласованности максимальных рангов элементов X и Y.

В настоящей работе изучается более общий вариант взвешенной меры ранговой корреляции – линейная ранговая статистика с монотонной функцией меток, а именно

$$T_{n,\varphi} = \frac{\sum\limits_{i=1}^{n} \left(a_n\left(R_i/(n+1)\right) - \bar{a}\right)\left(a_n\left(Q_i/(n+1)\right) - \bar{a}\right)}{\sum\limits_{i=1}^{n} \left(a_n\left(i/(n+1)\right) - \bar{a}\right)^2}, \qquad (2)$$

где функции $a_n(x)$ определены на интервале $(0,1)$ и стремятся в $L^2(0,1)$ к некоторой монотонной функции меток $\varphi(x)$, при этом $\bar{a} = n^{-1}\sum\limits_{i=1}^{n} a_n\left(i/(n+1)\right)$. В частности, $a_n\left(i/(n+1)\right) = \varphi\left(i/(n+1)\right)$ – так называемые приближенные метки, порожденные функцией φ. Ниже будет показано, что требование монотонности функции φ позволяет рассматривать статистику $T_{n,\varphi}$ как ранговый коэффициент корреляции со значениями на отрезке $[-1,1]$. Если $\varphi(x) = \ln(1/x)$, то $T_{n,\varphi}$ совпадает с коэффициентом r_T.

С помощью $T_{n,\varphi}$ можно получать меры зависимости, принимающие большие значения при совпадении нескольких первых и последних рангов X-ов и Y-ов одновременно. Иман и Коновер [2] предложили использовать для этой цели метки Ван дер Вардена, порожденные функцией $\varphi(x) = \Phi^{-1}(x)$, где Φ – ф.р. стандартного нормального закона. Естественно также, по аналогии с метками Сэвиджа, рассмотреть функцию $(\alpha \in (0,1))$

$$\varphi_\alpha(x) = \begin{cases} \ln(x/\alpha), & 0 < x \le \alpha, \\ \ln[(1-\alpha)/(1-x)], & \alpha < x < 1. \end{cases}$$

Одно из многочисленных приложений взвешенных коэффициентов r_T и $T_{n,\varphi}$ – определение точности спортивных прогнозов. В футбольных, баскетбольных или шахматных соревнованиях наиболее точным считается прогноз с правильно предсказанными призерами и, иногда, участниками-аутсайдерами, выбывающими в низшую лигу. Коэффициенты r_T и $T_{n,\varphi}$ с соответствующим образом подобранной φ позволяют выделить наиболее точное из нескольких предсказаний относительно исхода соревнований. В отличие от этих мер взаимосвязи классические коэффициенты корреляции Кендалла τ_n и Спирмена ρ_n, не использующие взвешивания, оказываются неподходящими в подобного рода задачах. Напомним, что

$$\tau_n = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \operatorname{sign}(R_i - R_j)\operatorname{sign}(Q_i - Q_j), \tag{3}$$

$$\rho_n = \frac{12}{n^2-1}\left[\frac{1}{n}\sum_{i=1}^{n} R_i Q_i - \left(\frac{n+1}{2}\right)^2\right]. \tag{4}$$

За исключением крайних ситуаций, когда τ_n и ρ_n принимают значения ± 1, статистики (3) и (4), вообще говоря, не равны. Впрочем, их корреляция очень высока и при нулевой гипотезе они асимптотически эквивалентны [4, § 4.4].

При хорошей согласованности нескольких наименьших и/или наибольших элементов среди $X_1,\dots,X_n$ и $Y_1,\dots,Y_n$ и рассогласованности в оставшейся части выборки значения статистик τ_n и ρ_n могут быть очень близкими к нулю (см. [2]). Малые же значения τ_n и ρ_n свидетельствуют в пользу гипотезы независимости и не передают информации о хорошей согласованности наименьших и/или наибольших среди элементов X и Y.

Модификация классических мер взаимосвязи путем взвешивания позволяет решать более разнообразные задачи, связанные с проверкой гипотезы независимости.

Пусть w_i и w_{ij} – ограниченные функции, заданные на множествах N и N^2 соответственно со значениями в R. Кроме того, функция w_{ij} предполагается симметричной. Не ограничивая общности предположим, что $X_1 < X_1 < \ldots < X_n$, так что $R_i = i$, $i = 1,\dots,n$. В недавней работе [6] вводятся в рассмотрение три вида статистик: взвешенный коэффициент корреляции Кендалла τ_w

$$\tau_w = \frac{2}{\left(\sum\limits_{i,j} w_{ij} - \sum\limits_{i} w_{ii}\right)} \sum_{1\le i<j\le n} w_{ij}\operatorname{sign}(i-j)\operatorname{sign}(Q_i - Q_j), \tag{5}$$

взвешенный коэффициент корреляции Спирмена ρ_w

$$\rho_w = \sum_{i=1}^{n} w_i(i-(n+1)/2)(Q_i-(n+1)/2), \tag{6}$$

переходящий в ρ_n, когда $w_i = 12/[n(n^2-1)]$, $i = 1\dots,n$, и статистика

$$r_w = \sum_{i=1}^{n} w_i(S_{n+1-i}-1)(S_{n+1-Q_i}-1), \tag{7}$$

которую можно назвать взвешенным коэффициентом Имана и Коновера, поскольку она совпадает с r_T при выборе $w_i = 1/(n-S_1)$, $i = 1,\dots,n$. Несколько раньше коэффициент τ_w изучался в [5], а общие идеи относительно взвешенной корреляции были предложены и обсуждались также в [7] и [8].

Как правило, в практических приложениях функция w_{ij} имеет вид $w_{ij} = w_i w_j$, где w_i – тот или иной индикатор или сумма индикаторов. В частности, часто полагают $w_i = I\{i \le [(n+1)p]\}$ или $w_i = I\{i \ge [(n+1)p]\}$, $p \in (0,1)$.

Заметим, что коэффициент r_T есть частный случай линейной ранговой статистики (2). Наряду с τ_w последняя является основным объектом исследования в данной работе. Полученные нами результаты относятся к вычислению асимптотической относительной эффективности (АОЭ) по Питмену критериев независимости, основанных на этих статистиках. Меры зависимости ρ_w и r_w имеют несколько отличную от $T_{n,\varphi}$ взвешенную структуру. Относительно этих тестовых статистик будут сделаны соответствующие замечания.

Являясь одной из важных характеристик непараметрических критериев, АОЭ позволяет в зависимости от исходных распределений (и альтернатив) упорядочивать имеющиеся статистические критерии и выявлять среди них наилучший.

Впервые вопросом об эффективности критериев (5)–(7) заинтересовались авторы [5] и [6], однако вычисление ее ими не предпринималось. Результаты моделирования в [6] показали, что для ряда распределений критерий τ_w лишь незначительно уступает в мощности критериям r_w и ρ_w. Поэтому вполне естественным является вопрос, сформулированный в [6]: когда (при каких распределениях выборки) и насколько эффективны взвешенные критерии по отношению друг к другу с точки зрения теории асимптотической эффективности? В настоящей работе питменовская эффективность указанных критериев вычисляется для достаточно широкого класса близких к нулевой гипотезе альтернатив, описываемых так называемой функцией зависимости.

2. Постановка задачи. Свойства тестовых статистик

Пусть (X_i, Y_i), $i = 1, \dots, n$, – выборка из абсолютно непрерывного распределения с ф.р.

$$F_\theta(x, y) = G(x)H(y) + \theta\Omega(G(x), H(y)), \quad (x, y) \in R^2, \tag{8}$$

и маргинальными ф.р. G и H. Функция Ω, называемая обычно функцией зависимости (ф.з.), задана на единичном квадрате $I^2 = [0,1]^2$, неотрицательна и ограничена, а параметр θ – малое вещественное число. В дальнейшем при вычислении питменовской эффективности будем предполагать, что $\theta = \theta_n = \delta/\sqrt{n}$, $\delta \geq 0$.

Обозначим через $\mathcal{F}$ семейство ф.р. вида (8), удовлетворяющих условиям

(F1) Функция $\Omega(x, y)$ обладает на I^2 ограниченной смешанной производной $\omega(x, y)$, причем

$$\omega(x, y) = \Omega_{xy}(x, y) = \Omega_{yx}(x, y) \ \text{ для п.в. } (x, y) \in I^2.$$

(F2) $\Omega(\partial I^2) = 0$, где ∂I^2- граница квадрата I^2.

В этих предположениях наблюдения имеют совместную плотность $f_\theta(x, y) = g(x)h(y)(1 + \theta\omega(G(x), H(y)))$, где $g(x) = G'(x)$, $h(y) = H'(y)$.

С помощью тестовых статистик (1), (2), (5)–(7) проверяется гипотеза независимости

$$H_0 : \theta = 0$$

против односторонних альтернатив

$$H_1 : \theta > 0 \text{ или } H_1^{'} : \theta < 0.$$

В работе из двух по существу одинаковых предположений H_1 и $H_1^{'}$ рассматривается альтернатива H_1.

Сделаем несколько замечаний о модели зависимости (8). По-видимому, впервые функцию распределения такого вида предложил Фарли [9]. В качестве альтернативы к независимости она фигурирует у многих авторов [10–16] и др.

Гипотеза H_1 (H_1') является важным частным случаем альтернативы строгой положительной (отрицательной) квадрантной зависимости, согласно которой $F(x,y) \geq G(x)H(y)$ $(F(x,y) \leq G(x)H(y))$ при всех x и y со строгим неравенством хотя бы для одной пары (x,y). Если же

$$\Omega(x,y) = \begin{cases} yg(G^{-1}(x))/q(G^{-1}(x)) + xh(H^{-1}(y)), & q(x) \neq 0, \\ xh(H^{-1}(y)), & q(x) = 0, \end{cases}$$

то гипотеза H_1 по существу сводится к альтернативе

$$H_2 : X_i = X_i^* + \theta q(X_i^*)Z_i, \quad Y_i = Y_i^* + \theta Z_i,$$

где X_i^*, Y_i^*, Z_i – взаимно независимые случайные величины (с.в.), причем типы распределений X_i^* и Y_i^* известны, распределение Z_i произвольно, $\theta > 0$ – малый вещественный параметр. При $q(x) \equiv 1$ эта модель рассматривалась многими авторами, в том числе очень подробно Гаеком и Шидаком [17]. Альтернатива H_2 с произвольной монотонной функцией q впервые появляется в [5]. Наиболее важным является случай индикаторной функции q (см. [5, 6]), при этом тестовые статистики вычисляются по неполному набору данных. С практической точки зрения это означает сокращение затрат на эксперимент и построение критериев независимости по данным, часть из которых по каким-либо причинам была утеряна или не наблюдалась.

Частным случаем H_1 является также известная в робастной статистике модель Тьюки–Хьюбера [18, § 4.5]. В приложении к задаче о независимости двух признаков она имеет вид

$$H_3 : F_\theta(x,y) = (1-\theta)G(x)H(y) + \theta\tilde{F}(x,y), \quad x,y \in R,$$

где G и H – фиксированные ф.р., а $\tilde{F}$ – произвольная ф.р., $\theta > 0$. В этом случае ф.з. имеет вид $\Omega(x,y) = \tilde{F}(x,y) - G(x)H(y)$.

Остановимся на свойствах тестовых статистик (1), (2) и (5)–(7). Основные утверждения будут формулироваться для коэффициента корреляции τ_w и линейной ранговой статистики $T_{n,\varphi}$. Об остальных тестовых статистиках будут сделаны соответствующие замечания.

Чтобы доказать следующее ниже утверждение, воспользуемся неравенством Харди–Литтльвуда–Полиа [19, с. 315], согласно которому для любых двух совокупностей чисел $\{\alpha_i\}_{i=1}^n$ и $\{\beta_i\}_{i=1}^n$

$$\sum_{i=1}^n \alpha_{[i]}\beta_{[n+1-i]} \le \sum_{i=1}^n \alpha_i\beta_i \le \sum_{i=1}^n \alpha_{[i]}\beta_{[i]}, \tag{9}$$

где $\alpha_{[i]}$ и $\beta_{[i]}$ занимают i-ую позицию в упорядоченных наборах $\alpha_{[1]} \ge \ldots \ge \alpha_{[n]}$ и $\beta_{[1]} \ge \ldots \ge \beta_{[n]}$.

Утверждение 1. *Пусть φ – произвольная монотонная функция, заданная на интервале $(0,1)$. Тогда статистика $T_{n,\varphi}$ достигает своего максимального значения 1, когда $R_i = Q_i$, и минимального значения, удовлетворяющего соотношению*

$$\min_{\{(R_i,Q_i)\}_{i=1}^n} T_{n,\varphi} \to \frac{\int_0^1 \varphi(x)\varphi(1-x)dx - \left(\int_0^1 \varphi(x)dx\right)^2}{\int_0^1 \varphi^2(x)dx - \left(\int_0^1 \varphi(x)dx\right)^2}, \quad n\to\infty,$$

когда $R_i = Q_{n+1-i}$, $i = 1,\ldots,n$.

Доказательство. Выберем и зафиксируем произвольное $n \in N$. Применим к последовательностям $\hat\alpha_i = \hat\beta_i = a_n\left(i/(n+1)\right) - \bar a$ неравенство (9). Без ограничения общности полагаем, $R_i = i$, $i = 1,\ldots,n$. Заключение утверждения в силу монотонности φ следует из неравенства

$$\sum_{i=1}^n \hat\alpha_i\hat\beta_{n+1-i} \le \sum_{i=1}^n \hat\alpha_i\hat\beta_{Q_i} \le \sum_{i=1}^n \hat\alpha_i\hat\beta_i,$$

справедливого для всякого $n \in N$.

По доказанному утверждению и неравенству Коши–Буняковско-го–Шварца имеем $-1 \le T_{n,\varphi} \le 1$. В частности, $-1 \le r_T \le 1$, причем наименьшее значение r_T равно -1 лишь при $n = 2$. С ростом $n \to \infty$ минимальное значение возрастает и стремится к -0.645.

Асимптотическая нормальность коэффициента $T_{n,\varphi}$ как при нулевой гипотезе, так и при альтернативе легко следует из теоремы Реймхар-та–Шорака–ван Цвета [20]. Сформулируем соответствующий результат.

Предположим, что ф.р. (8) принадлежит классу $\mathcal{F}$ и, кроме того, что $G(x) = x$, $H(y) = y$, $x, y \in [0,1]$. Пусть, далее, имеется монотонная функция $\varphi \in L^2(0,1)$, дифференцируемая на $(0,1)$

всюду, за исключением, быть может, конечного числа точек, и удовлетворяющая условиям

$$|\varphi^{(i)}(x)| \leq D(r(x))^{\frac{1}{4}-\delta+i}, \quad i=0,1, \quad 0<\delta<1/4,$$

где $r(x)=(x(1-x))^{-1}$, D – положительная константа. Множество таких функций φ обозначим через $\mathcal{M}$.

Утверждение 2. *Пусть совместное распределение с.в. X и Y принадлежит классу $\mathcal{F}$, а функция φ – классу $\mathcal{M}$. Тогда при $n \to \infty$*

$$\sqrt{n}(T_{n,\varphi}-\mu_\varphi(\theta)) \xrightarrow{d} N(0,\sigma^2_\varphi(\theta)), \tag{10}$$

с $\mu_\varphi(\theta)$ и $\sigma_\varphi(\theta)$, определенными в (11) и (12).

Доказательство. Обозначим для краткости

$$c=\int_0^1 \varphi(x)dx,\ d=\int_0^1 (\varphi(x)-c)^2\,dx,$$

и заметим, что в интересных и содержательных случаях $d>0$.

По теореме 2.1 из [20] соотношение (10) имеет место с параметрами

$$\begin{aligned}\mu_\varphi(\theta)=&\frac{1}{d}\iint_{I^2}(\varphi(x)-c)(\varphi(y)-c)dxdy+\\ &+\frac{\theta}{d}\iint_{I^2}(\varphi(x)-c)(\varphi(y)-c)d\Omega(x,y)=\\ &=\frac{\theta}{d}\iint_{I^2}(\varphi(x)-c)\,(\varphi(y)-c)\,\omega(x,y)dxdy\end{aligned} \tag{11}$$

и

$$\begin{aligned}\sigma^2_\varphi(\theta)=&\frac{1}{d^2}D_\theta\left[(\varphi(X)-c)\right)(\varphi(Y)-c)+\\ &+\frac{1}{d^2}\iint_{I^2}\left(I_{[X,+\infty)}(x)-x\right)\varphi'(x)\,(\varphi(y)-c)\,dF_\theta(x,y)+\\ &+\frac{1}{d^2}\iint_{I^2}\left(I_{[Y,+\infty)}(y)-y\right)\,\varphi'(y)\,(\varphi(x)-c)\,dF_\theta(x,y)]=\\ &=\sigma^2_\varphi(0)+\theta C(\varphi,\Omega)+o(\theta),\quad \theta\to 0,\end{aligned} \tag{12}$$

где $C(\varphi, \Omega)$ – ограниченная величина, определяемая функциями φ и Ω, а

$$\sigma_\varphi^2(0) = \frac{1}{d^2} D_0[(\varphi(X) - c)(\varphi(Y) - c)] =$$

$$= \frac{1}{d^2} E_0[(\varphi(X) - c)^2 (\varphi(Y) - c)^2] = 1.$$

Определение класса $\mathcal{M}$ обеспечивает справедливость предположения 1 теоремы Реймхарта–Шорака–ван Цвета. Предположение 2 этой же теоремы, дающее условия сходимости меток a_n к порождающей их функции φ, имеет место в силу сходимости $a_n(x)$ к $\varphi(x) \in \mathcal{M}$ в $L^2(0,1)$ и неравенства Чебышева. Таким образом, справедливо заключение теоремы 2.1 из [20], а следовательно и соотношение (10).

Замечание 1. Если параметр θ лежит в малой окрестности нуля $[0, \theta_0], \theta_0 > 0$, то аргументы, используемые при доказательстве утверждения 2 и равенство $\sigma_\varphi^2(\theta) = 1 + \theta C(\varphi, \Omega) + o(\theta)$, $\theta \to 0$, в котором величина $C(\varphi, \Omega)$ ограничена, приводят ко второму утверждению теоремы Реймхарта–Шорака–ван Цвета. Таким образом, в классе функций распределения $\{F_\theta : F_\theta \in \mathcal{F}, \theta \in [0, \theta_0]\}$ сходимость (10) осуществляется равномерно по $\theta \in [0, \theta_0]$.

Рассмотрим взвешенный коэффициент корреляции Кендалла τ_w и введем следующие обозначения.

$$h_\theta(x, y) = E_\theta\{\mathrm{sign}(X_1 - X_2)\mathrm{sign}(Y_1 - Y_2) | (X_1, Y_1) = (x, y)\} -$$

$$-E_\theta\{\mathrm{sign}(X_1 - X_2)\mathrm{sign}(Y_1 - Y_2)\}, \quad m(\Omega) = 8 \iint\limits_{I^2} \Omega(x, y)\,dxdy,$$

$$U_{1,w} = \frac{2}{(\sum\limits_{i,j} w_{ij} - \sum\limits_{i} w_{ii})} \sum_{i,j} w_{ij} h_\theta(X_i, Y_i), \quad U_{2,w} = \tau_w - U_{1,w},$$

$$\sigma_w^2(\theta) = D_\theta U_{1,w}.$$

Взвешенная U-статистика τ_w обладает невырожденным ядром степени два: $\Phi((X_1, Y_1), (X_2, Y_2)) = \mathrm{sign}(X_1 - X_2)\mathrm{sign}(Y_1 - Y_2)$. Исследование предельного распределения взвешенных U-статистик в зависимости от ранга ядра и поведения весовых функций проводилось О'Нэйлом и Реднером [21], а несколько позже Майором [22]. Цформулированное ниже утверждение является следствием более раннего результата Шапиро и Хуберта [23].

Обозначим

$$W_1 = \lim_{n\to\infty} \frac{1}{n^2} \sum_{i,j=1}^{n} w_{ij}, \quad W_2 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{j=1}^{n} w_{ij} \right)^2 .$$

Утверждение 3. *Пусть при $n \to \infty$ выполнены условия*

$$\text{(I)}\ \frac{\sum_{i,j} w_{ij}^2}{\sum_i \left(\sum_j w_{ij}\right)^2} \to 0 \quad \text{и} \quad \text{(II)}\ \frac{\max\limits_i \left(\sum_j w_{ij}\right)^2}{\sum_i \left(\sum_j w_{ij}\right)^2} \to 0.$$

Тогда

$$\frac{\tau_w - \mu(\theta)}{\sqrt{\sigma_w^2(\theta)}} \xrightarrow{d} N(0,1), \qquad n \to \infty, \tag{13}$$

где $\mu(\theta) = \theta m(\Omega)$, а $m(\Omega)$ и $\sigma_w^2(\theta)$ определены выше. При этом $\mu(0) = 0$ и $\sigma_w^2(0) = (4W_2/9{W_1}^2)n^{-1}$.

Кроме того, существует производная $\mu'(\theta)$, строго положительная в точке нуль, причем $\lim\limits_{n\to\infty} \dfrac{\mu'(\theta_n)}{\mu'(0)} = 1$, $\lim\limits_{n\to\infty} \dfrac{\sigma_w'(\theta_n)}{\sigma_w'(0)} = 1$, когда $\theta = \theta_n = \delta n^{-1/2}$, $\delta > 0$.

Доказательство опирается на работу [23], в основе которой лежит метод проекций (см., например, [4, § 2.5]).

Так как

$$E_\theta\{\mathrm{sign}(X_1 - X_2)\mathrm{sign}(Y_1 - Y_2)\} = \theta m(\Omega)(1 + o(1)), \quad \theta \to 0,$$

то из определения $h_\theta(x,y)$ следует, что

$$h(x,y) = h_\theta(x,y) = (2x-1)(2y-1) + O(\theta), \quad \theta \to 0. \tag{14}$$

Из [23] следует, что сходимость (13) имеет место тогда, когда дисперсия $D_\theta h(X,Y)$ строго положительна. Справедливость этого требования вытекает из (8) и (14):

$$D_\theta h(X,Y) = E_\theta h^2(X,Y) + o(\theta) = \frac{1}{9} + O(\theta), \quad \theta \to 0. \tag{15}$$

Пусть далее $g_{ij} = \mathrm{sign}(X_i - X_j)\mathrm{sign}(Y_i - Y_j) - h(X_i, Y_i) - h(X_j, Y_j)$. Тогда

$$\tau_w = \frac{2}{(\sum\limits_{i,j} w_{ij} - \sum\limits_i w_{ii})} \left\{ \sum_{i<j} w_{ij}[h(X_i,Y_i) + h(X_j,Y_j)] + \sum_{i<j} w_{ij} g_{ij} \right\},$$

где $h(X_i, Y_i)$ – независимые одинаково распределенные с.в. Заметим, что $U_{1,w} = \frac{2}{(\sum\limits_{i,j} w_{ij} - \sum\limits_i w_{ii})} \sum\limits_i (\sum\limits_j w_{ij}) h(X_i, Y_i)$, поэтому при малых $\theta = \theta_n = \delta n^{-1/2}$, $\delta > 0$,

$$D_\theta U_{1,w} \sim \frac{4}{(\sum\limits_{i,j} w_{ij} - \sum\limits_i w_{ii})^2} \sum_i \left(\sum_j w_{ij} \right)^2 E_\theta h^2(X_i, Y_i),$$

$$D_\theta U_{2,w} \sim \frac{4}{(\sum\limits_{i,j} w_{ij} - \sum\limits_i w_{ii})^2} E_\theta g_{12}^2 \sum_{i,j} w_{ij}^2,$$

при не ограничивающем общности предположении, что $w_{ii} = 0$, $i = 1, \ldots, n$. Из (15) и условия (И) следует, что

$$\lim_{n\to\infty} n D_\theta \tau_w = \lim_{n\to\infty} \frac{4}{9n} \sum_i \left(\frac{1}{n} \sum_j w_{ij} \right)^2 \frac{n^4}{\left(\sum_{i,j} w_{ij} \right)^2} = \frac{4 W_2}{9 W_1^2}. \quad (16)$$

На основании представления $U_{1,w} = \frac{2}{(\sum\limits_{i,j} w_{ij} - \sum\limits_i w_{ii})} \sum\limits_{i,j} w_{ij} h(X_i, Y_i)$ и теоремы Линдеберга, которая справедлива в силу условия (II), вместе с (16) получаем

$$\frac{\tau_w - \mu(\theta)}{\sqrt{\sigma_w^2(\theta)}} \xrightarrow{d} N(0,1), \quad n \to \infty,$$

где $\mu(\theta) = \theta m(\Omega)$, $\sigma_w^2(\theta) = \frac{4\, W_2}{9\, W_1^2} \frac{1}{n}(1 + o(1))$, когда $\theta = \theta_n = \delta n^{-1/2}$.

Производная $\mu'(0) = m(\Omega)$ и поэтому строго положительна, а из определения $\sigma_w^2(\theta)$ и (15) следует, что

$$\lim_{n\to\infty} \frac{\mu'(\theta_n)}{\mu'(0)} = 1 \text{ и } \lim_{n\to\infty} \frac{\sigma_w'(\theta_n)}{\sigma_w'(0)} = 1.$$

Замечание 2. Условия (I) и (II) автоматически выполнены для любой функции w_{ij}, совпадающей во всех точках (i, j), $i, j = 1, \ldots, n$, ц некоторой функцией $w(i/(n+1), j/(n+1))$, где $w \in L^2((0,1)^2)$.

Замечание 3. Если весовая функция в определении τ_w имеет вид $w_{ij} = w_i w_j$, где w_i – индикатор или сумма индикаторов непересекающихся множеств, в частности, $w_i = I\{i \leq [(n+1)p]\}$ или $w_i = I\{i \geq [(n+1)p]\}$, $p \in (0,1)$, то в силу существования равномерной оценки в центральной предельной теореме для U-статистик ([24], цтр. 190) соотношение (13) имеет место равномерно по всем $\theta \in [0, \theta_0]$, $\theta_0 > 0$.

Отметим, что тестовая цтатистика τ_w принимает значения на отрезке $[-1, 1]$ и при гипотезе H_0 имеет нулевое математическое ожидание. Она дает состоятельный против альтернативы (8) критерий независимости, критическая область которого задается неравенством $\tau_w > \lambda_{n,w}$ с $\lambda_{n,w} = \sqrt{\frac{4W_2}{9W_1^2}\frac{1}{n}}z_\alpha$, где $\Phi(z_\alpha) = 1 - \alpha$.

3. Питменовская эффективность взвешенных критериев независимости

Питменовская эффективность является классическим и широко используемым средством для асимптотического сравнения критериев. Она обладает рядом достоинств, в числе которых ее высокая точность уже при небольших (порядка 30–50 для простейших критериев) размерах выборки [25]. Утверждения 2 и 3 позволяют воспользоваться условиями регулярности Питмена, предложенными им для вычисления АОЭ критериев (см. [4, с. 78–79] или [10, с. 26]) и определить, насколько эффективными являются тестовые статистики $T_{n,\varphi}$ и τ_w по отношению друг к другу и к классическим мерам зависимости.

Для линейной ранговой статистики $T_{n,\varphi}$ питменовская мера эффективности равняется

$$\mathrm{eff}(T_{n,\varphi}, \Omega) = \left(\frac{\mu'_\varphi(0)}{\sigma_\varphi(0)}\right)^2 = \frac{1}{d^2}\left(\iint\limits_{I^2} \varphi'(x)\varphi'(y)\Omega(x,y)dxdy\right)^2, \quad (17)$$

где $d = \int_0^1 (\varphi(x) - c)^2\,dx$, $c = \int_0^1 \varphi(x)dx$, а для статистики τ_w

$$\mathrm{eff}(\tau_w, \Omega) = \left(\frac{\mu'(0)}{\sigma_w(0)}\right)^2 = \frac{9\,W_1^2}{4\,W_2}\,m^2(\Omega) = 144\frac{W_1^2}{W_2}\left(\iint\limits_{I^2} \Omega(x,y)dxdy\right)^2, \quad (18)$$

причем в силу неравенства Коши–Буняковского–Шварца всегда $W_1^2/W_2 \le 1$. В частности, при $\varphi(x) = \ln(1/x)$ и $w(x) \equiv 1$ по формуле (17) имеем

$$\mathrm{eff}(r_T, \Omega) = \left(\iint\limits_{I^2} \frac{\Omega(x,y)}{xy} dxdy \right)^2 . \tag{19}$$

Следствием (17), когда $\varphi(x) = x$, и (18), когда $w_{ij} = 1$, $i,j = 1,\dots,n$, является известный результат Фарли [9], согласно которому

$$\mathrm{eff}(\rho_n, \Omega) = \mathrm{eff}(\tau_n, \Omega) = 144 \left(\iint\limits_{I^2} \Omega(x,y) dxdy \right)^2 .$$

Предположим теперь, что имеются две последовательности тестовых статистик $\{T_n\}$ и $\{V_n\}$ ц мерами эффективности $\mathrm{eff}(T)$ и $\mathrm{eff}(V)$ соответственно. Отношение

$$\mathrm{eff}(T|V) = \frac{\mathrm{eff}(T)}{\mathrm{eff}(V)} \tag{20}$$

называется АОЭ по Питмену критерия T_n по отношению к критерию V_n. При определенных условиях регулярности, наложенных на распределение исходных наблюдений, питменовская мера эффективности всегда не превосходит в силу [26] информации Фишера $I = I(0) = \iint\limits_{I^2} \left(\frac{\partial \ln f_\theta(x,y)}{\partial \theta} \right)^2 f_\theta(x,y) dxdy|_{\theta=0}$. Проверку этих условий для модели зависимости (8) можно найти в [27, § 2.4]. В отличие от меры эффективности $\mathrm{eff}(T)$ эффективность

$$e^P(T) = \mathrm{eff}(T)/I(0) \tag{21}$$

позволяет судить об абсолютном качестве критерия: основанный на статистике T_n критерий тем лучше, чем ближе к единице значение $e^P(T)$.

Возвращаясь к взвешенным ранговым критериям независимости, на основании (17)–(20) получаем, что АОЭ критериев $T_{n,\varphi}$ и τ_w по отношению к критерию r_T составляет

$$\mathrm{eff}(T_{n,\varphi}|r_T, \Omega) = \left(\frac{1}{d} \iint\limits_{I^2} \varphi'(x)\varphi'(y)\Omega(x,y) dxdy \Big/ \iint\limits_{I^2} \frac{\Omega(x,y)}{xy} dxdy \right)^2$$

и

$$\mathrm{eff}(\tau_w\,|r_T,\Omega) = 144\frac{W_1^2}{W_2}\left(\iint\limits_{I^2}\Omega(x,y)dxdy \Big/ \iint\limits_{I^2}\frac{\Omega(x,y)}{xy}dxdy\right)^2$$

соответственно.

Для того, чтобы определить насколько эффективны критерии ρ_w и r_w при произвольном выборе весов w_i, необходимо рассмотреть нормированную подходящим образом линейную ранговую статистику вида $T_n = n^{-1}\sum\limits_{i=1}^{n} a_1(R_i/(n+1))a_2(Q_i/(n+1))$. При естественных ограничениях на функции меток a_1 и a_2, рассуждения, аналогичные доказательству утверждения 2, приводят к формулам для питменовской эффективности взвешенных критериев ρ_w и r_w. Так, для статистики $R_w = n^{-1}\sum\limits_{i=1}^{n} w_i\,(i/(n+1)-1/2)\,(Q_i/(n+1)-1/2) = \rho_w \times n^{-1}(n+1)^{-2}$ ц весами $w_i = I\{i \le [(n+1)p]\}$, $p \in (0,1)$, согласно теореме 2.1 из [20] при $n \to \infty$

$$\frac{\sqrt{n}(R_w - \mu_p(\theta))}{\sigma_p(\theta)} \xrightarrow{d} N(0,1)$$

равномерно по всем $\theta \in [0,\theta_0]$ ц параметрами

$$\mu_p(\theta) = \theta\int\limits_0^p (x-1/2)\left(\int\limits_0^1 (y-1/2)\omega(x,y)dy\right)dx,$$

$$\sigma_p^2(\theta) = \sigma_p^2(0) + O(\theta),\ \theta \to 0,\ \sigma_p^2(0) = p[3(1-p)^3+p^2]/144.$$

Поскольку питменовская мера эффективности инвариантна относительно линейного преобразования тестовой статистики, то для коэффициентов R_w и ρ_w ее величина одна и та же и составляет

$$\mathrm{eff}(\rho_w,\Omega) = \left(\frac{\mu_p'(0)}{\sigma_p(0)}\right)^2 = \frac{144\left(\int\limits_0^p (x-1/2)\left(\int\limits_0^1 (y-1/2)\omega(x,y)dy\right)dx\right)^2}{p[3(1-p)^3+p^2]}. \tag{22}$$

В заключении рассмотрим несколько примеров.

Пример 1. Пусть $F_\theta(x,y)$ – ф.р. стандартного нормального закона с нулевыми средними, единичными дисперсиями компонент и коэффициентом корреляции θ. В этом случае при $\theta \to 0$ линейная часть $F_\theta(x,y)$ имеет вид (8) с ф.з. $\Omega(x,y) = \Omega_{\mathcal{N}}(x,y) = \varphi(\Phi^{-1}(x))\varphi(\Phi^{-1}(y))$, где φ – плотность распределения стандартного нормального закона. Несмотря на то, что данная ф.р. формально не принадлежит классу $\mathcal{F}$, утверждения 2 и 3, как несложно показать, остаются справедливыми. Пусть $w_{ij} = w_i w_j$, $w_i = I\{i \le [(n+1)p]\}$, $p \in (0,1)$. При таком выборе $W_1^2/W_2 = p$ и согласно (18)–(22)

$$\mathrm{eff}(\tau_w | r_T, \Omega_{\mathcal{N}}) = \frac{9}{\pi^2} p \Bigg/ \left(\int_0^1 \frac{\varphi(\Phi^{-1}(x))}{x} dx \right)^4,$$

$$\mathrm{eff}(\rho_w | r_T, \Omega_{\mathcal{N}}) = \frac{144}{p[3(1-p)^3 + p^2]} \left(\int_0^1 \frac{\varphi(\Phi^{-1}(x))}{x} dx \right)^{-4} \times$$

$$\times \left(\int_0^p (x - 1/2)\Phi^{-1}(x) dx \right)^2 \left(\int_0^1 (y - 1/2)\Phi^{-1}(y) dy \right)^2.$$

Так как для нормального закона информация Фишера в точке нуль $I(0) = \iint\limits_{I^2} \omega_{\mathcal{N}}^2(x,y)dxdy = \iint\limits_{I^2} (\Phi^{-1}(x)\Phi^{-1}(y))^2 dxdy = 1$, и $\int\limits_0^1 \frac{\varphi(\Phi^{-1}(x))}{x} dx \approx 0.9032$, то питменовская эффективность критерия Имана и Коновера составляет $e^P_{r_T}(\Omega_{\mathcal{N}}) \approx 0.6655$, а для критериев τ_w и ρ_w при $p = 0.3,\ 0.5,\ 0.7,\ 0.9$ она равняется

$$e^P_{\tau_w}(\Omega_{\mathcal{N}}) \approx 0.2736,\ 0.4559,\ 0.6383,\ 0.8207;$$

$$e^P_{\rho_w}(\Omega_{\mathcal{N}}) \approx 0.6147,\ 0.7295,\ 0.6271,\ 0.6395$$

соответственно.

Пример 2. Пусть $\Omega_{\sin}(x,y) = C \sin(\pi x)\sin(\pi y)$, $C > 0$. Данная ф.з. определяет распределение класса $\mathcal{F}$, на котором критерий независимости Блюма–Кифера–Розенблатта имеет максимальную эффективность по Бахадуру [11, 10, § 6.6]. В этом случае информация Фишера $I(0) = \iint\limits_{I^2} \omega_{\sin}^2(x,y)dxdy = \pi^4/4$, поэтому на основании (18), (19), (21) и (22) имеем $e^P_{r_T}(\Omega_{\sin}) \approx 0.4830$, а для критериев τ_w и ρ_w при $p = 0.3,\ 0.5,\ 0.7,\ 0.9$ получаем

$$e^P_{\tau_w}(\Omega_{\sin}) \approx 0.2914,\ 0.4856,\ 0.6799,\ 0.8742;$$

$$e^P_{\rho_w}(\Omega_{\sin}) \approx 0.6130,\, 0.7770,\, 0.7078,\, 0.8104.$$

Если $p = 1$, то есть ρ_w и τ_w – классические коэффициенты корреляции, то $e^P_{\tau_n}(\Omega_{\sin}) = e^P_{\rho_n}(\Omega_{\sin}) = 9216/\pi^8 \approx 0.9713$.

Пример 3. Рассмотрим распределение Фарли–Гумбеля–Моргенштерна (ФГМ) с ф.з. $\Omega_{\mathrm{F}}(x,y) = Cx(1-x)y(1-y)$, $C > 0$ [9, 28, 29]. Известно [10, § 6.6], что в этом случае критерии τ_n и ρ_n асимптотически оптимальны по Питмену. На ФГМ-распределении информация Фишера $I(0) = 1/9$, поэтому если $w_{ij} = w_i w_j$, $w_i = I\{i \le [(n+1)p]\}$, $p \in (0,1)$, то вновь согласно формулам (18), (19), (21) и (22) получаем, что $e^p_{r_T}(\Omega_{\mathrm{F}}) \approx 0.5625$ и

$$e^P_{\tau_w}(\Omega_{\mathrm{F}}) = 0.3000,\, 0.5000,\, 0.7000,\, 0.9000;$$

$$e^P_{\rho_w}(\Omega_{\mathrm{F}}) \approx 0.6524,\, 0.8000,\, 0.7081,\, 0.7811,$$

когда $p = 0.3,\, 0.5,\, 0.7,\, 0.9$ соответственно.

Более эффективным по сравнению с r_T в данном случае оказывается коэффициент $T_{n,\varphi}$, у которого функция $\varphi(x)$ убывает на $(0,1)$ медленее, чем $\ln(1/x)$. Статистики такого вида можно использовать для установления точности прогноза в ситуациях, когда важна не столько точность предсказания нескольких первых элементов, сколько правильное расположение более широкого набора данных.

Полученные выше значения питменовской эффективности указывают на различие между асимптотическими свойствами взвешенных статистик и их классических аналогов. Ярким примером являются взвешенные статистики Кендалла τ_w и Спирмена ρ_w: асимптотически эквивалентные при постоянных весах w_i и w_{ij}, они обладают различной питменовской эффективностью, когда $w_{ij} = w_i w_j$, $w_i = I\{i \le [(n+1)p]\}$, $p \in (0,1)$. Построенный по $100p$ процентам наблюдений взвешенный критерий τ_w теряет в эффективности пропорционально доле выбранных наблюдений. Зависимость же эффективности критерия ρ_w от p имеет более сложный характер (см. формулу (22)). Как следует из примеров 1–3, питменовская эффективность взвешенного критерия Спирмена даже при небольших p может быть достаточно высокой, сравнимой, а порой превосходящей эффективность критериев τ_w и r_T. Таким образом, идея взвешивания тестовых статистик, оказавшаяся столь плодотворной для рангового коэффициента ρ_n, заслуживает, по нашему мнению, дальнейшего изучения и развития. Можно предположить, что существуют взвешенные критерии, основанные на ранговых статистиках вида $T_n = n^{-1}\sum_{i=1}^{n} a_1(R_i/(n+1))a_2(Q_i/(n+1))$, которые на конкретных распределениях обладают высокой питменовской эффективностью. При этом выбор меток a_1 и a_2 будет зависеть от условий, целей и задач эксперимента.

Литература

1. М. Кендэл, *Ранговые корреляции.* Статистика, М. (1975).
2. R. L. Iman, W. J. Conover, *A measure of top–down correlation.* Technometrics, **29**, No. 3 (1987), 351–357.
3. I. R. Savage, *Ann. Math. Statist.,* **27** *(1965), 590–615.*
4. Т. Хеттманспергер, *Статистические выводы, основанные на рангах.* Финансы и статистика, М. (1987).
5. G. S. Shieh, *A weighted Kendall's tau statistic.* Statist. Probab. Letters., **39** (1998) 17–24.
6. G. S. Shieh, Z. Bai, W.-Y. Tsai, *Rank tests of independence – with a weighted contamination alternative.* Statistica Sinica **10**, No. 2 (2000), 577–594.
7. I. Salama, D. A. Quade, *A nonparametric comparison of two multiple regressions by means of a weighted measure of correlation.* Commun. Statist. Theory and Meth., **11** (1982), 1185–1195.
8. D. A. Quade, I. Salama, *A survey of weighted rank correlation.* In: P.K. Sen and I.A. Salama (Eds.), Order Statistics and Nonparametrics: Theory and Applications. Elsevier, Amsterdam (1992), 213–224.
9. Д. J. Г. Фарлие, *Тхе перформанце оф соме цорrелатион цоеффициентс фор а генерал бивариате дистрибутион.* Биометрика **47** (1960), 307–323.
10. Я. Ю. Никитин, *Асимптотическая эффективность непараметрических критериев.* Наука, М. (1995).
11. Я. Ю. Никитин, А. Г. Панкрашова, *Бахадуровская эффективность и локальная асимптотическая эффективность некоторых непараметрических критериев независимости.* Зап. научн. семинаров ЛОМИ, **166** (1988), 112–127.
12. P. Bajorski, *Local Bahadur optimality of some rank tests of independence.* Statist. Probab. Letters, **5**, No. 6 (1987), 255–262.
13. H. S. Konijn, *Positive and negative dependence of two random variables.* Sankhyā **24**, Pt. 2 (1959), 269–280.
14. R. L. Plackett, *A class of bivariate distributions.* Journ. Amer. Statist. Ass. **60** (1965), 516–522.
15. M. L. Puri, P. K Sen, *Nonparametric methods in multivariate analysis.* Wiley, New York (1971).
16. G. G. Woodworth, *On the asymptotic theory of tests of independence based on bivariate layer ranks.* Technical Report No. 75, Dept. of Statistics, Univ. of Minnesota (1966).

17. Я. Гаек, З. Шидак, *Теория ранговых критериев.* Наука, М. (1971).
18. П. Хьюбер, *Робастность в статистике.* Мир, М. (1984).
19. Г. Харди, Дж. Литтльвуд, Г. Полиа, *Неравенства.* ИЛ, М. (1948).
20. F. H. Ruymgaart, G. R. Shorack, W. R. van Zwet, *Asymptotic normality of nonparametric tests for independence.* Ann. Math. Statist., **43**, No. 4 (1972), 1122–1135.
21. K. A. O'Neil, R. A. Redner, *Asymptotic distribution of weighted U-statistic of degree 2.* Ann. Probab., **2** (1993), 1159–1169.
22. P. Major, *Asymptotic distributions for weighted U-statistics.* Ann. Probab., **22**, No. 3 (1994), 1514–1535.
23. C. P. Shapiro, L. Hubert, *Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions.* Ann. Statist., **7** (1979), 788–794.
24. В. С. Королюк, Ю. В. Боровских, *Теория* U-*статистик.* Наукова думка, Киев (1989).
25. P. Groeneboom, J. Oosterhoff, *Bahadur efficiency and small sample efficiency.* Intern. Statist. Review **49**, No. 2 (1981), 127–141.
26. C. R. Rao, *Criteria of estimation in large samples.* Sankhyā, **A25** (1963), 198–206.
27. Н. А. Цтепанова, *Асимптотическая эффективность ранговых критериев независимости.* Дис. ... канд. физ.-мат. наук. С.-Петербург (2000).
28. E. J. Gumbel, *Distributions à plusieurs variables dont les marges sont données.* C. R. Acad. Sci. **246** (1958), 2717–2719.
29. D. Morgenstern, *Einfache Beispiele zweidimensionaler Verleilungen.* Mitt. Math. Statist. **8** (1956), 234–235.

Nikitin Ya. Yu., Stepanova N. A. Pitman efficiency of tests for independence based on weighted rank statistics.

In this paper we consider some tests for independence based on weighted rank statistics and calculate their Pitman efficiency. It's shown that the efficiency of each of these tests depends on weights in its own peculiar way.

С.-Петербургский государственный университет

Поступило 28 мая 2001 г.