Об асимптотической мощности "энергетического" теста для проверки гипотез о равенстве двух распределений

В.Б. Мелас, Д.И. Сальников [1]

1. Введение

Найдены асимптотическое распределение и асимптотическая мощность критерия проверки гипотез о равенстве двух распределений, предложенного в работах [1], [2], для случая, когда альтернативное распределение отличается сдвигом и(или) параметром масштаба. Эти результаты являются обобщением и развитием результатов статьи [3].

2. Постановка проблемы

Задача проверки гипотезы о равенстве двух распределений является классической задачей математической статистики и для её решения предложено значительное число различных методов. Однако, ряд популярных методов, например , t–критерий, требуют предположения о нормальности распределений, а метод отношения правдоподобия предполагает распределения заданными, хотя бы с точностью до параметров. Универсальный критерий Колмогорова–Смирнова часто имеет низкую мощность, а непараметрический критерий Манна- Уитни непригоден в случае, когда распределения различаются только масштабом.

Рассмотрим задачу проверки гипотез о равенстве двух распределений

$$H_0 \,:\, F_1 = F_2 \tag{1}$$

против альтернативы

$$H_1 \,:\, F_1 \neq F_2 \tag{2}$$

в случае двух независимых выборок $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_m)$ с функциями распределения F_1 и F_2 соответственно.

Предположим, что функции распределения F_1 и F_2 принадлежат классу функций распределений случайных величин ξ, таких, что

$$E[g(\xi)^2] < \infty, \tag{3}$$

где g - некоторая заданная функция. Многие распределения, в том числе нормальное распределение и распределение Коши, обладают этим свойством при $g(x) = \ln(1 + x^2)$.

В случае, когда два распределения отличаются только сдвигом, во многих случаях наиболее мощным является тест Манна-Уитни- Вилкоксона. Однако хорошо известно, что этот тест не позволяет дискриминировать распределения, разлиичающиеся параметром масштаба (см. [1], [2]). Мы хотели бы иметь тест, подходящий для ситуаций, когда нулевое распределение относится к классу распределений, обладающих свойством (3) для функции g общего вида, а альтернативное распределение отличается только

[1]St.-Petersburg State University, Russia, e-mail: vbmelas@yandex.ru

сдвигом и/или преобразованем масштаба. Задачи, в которых важно учитывать возможность различия в параметре масштаба, возникают во многих практических областях применения, влючая физиологию и психологию (см. например недавнюю работу [4]). Рассмотрим следующий тест

$$\Phi_{nm} = \Phi_{nm}(X, Y) = \Phi_{AB} - \Phi_A - \Phi_B, \tag{4}$$

$$\Phi_A = \frac{1}{n^2} \sum_{1 \le i < j \le n} g(X_i - X_j), \Phi_B = \frac{1}{m^2} \sum_{1 \le i < j \le m} g(Y_i - Y_j),$$

$$\Phi_{AB} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} g(X_i - Y_j),$$

где $g(x)$ есть некоторая заданная функция. Будем предполагать, что эта функция неотрицательна, симметрична относительно начала координат и дважды непрерывно дифференцируема. Этот тест был, по-видимому, впервые введен в работе [1], и назван "энергетическим"тестом, но его мощность исследовалась только с помощью статистического моделирования и лишь для случая $g(x) = \ln(|x|)$.

В работе [2] критерий (4) изучался для случая

$$g(x) = \ln(1 + x^2).$$

Было показано с помощью статистического моделирования, что критерий (4) с такой функцией g имеет для многих распределений примерно такую же мощность как лучший из критериев Вилкоксона, Андесона- Дарлинга и Колмогорова-Смирнова при альтернативе, которая отличается только величиной параметра сдвига, но значительно превосходит эти критерии, если есть различие в параметре масштаба. .

Рассмотрим класс распределений, задаваемых свойством (3). Асимптотическая мощность теста для функций распределения, удовлетворяющих свойству (3), была изучена в работе [3] в случае функций g общего вида для распределений, отличающихся только сдвигом.

В настоящей работе мы изучаем асимптотическую мощность теста (4) для альтернативных распределений, отличающихся от нулевого величиной параметра сдвига и/или параметра масштаба.

3. Асимптотическая мощность

Рассмотрим случай двух распределений, обладающих свойством (3) и отличающихся сдвигом и(или) параметром масштаба.

Пусть $f(x)$ обозначает плотность F_1, $f_2(x)$ обозначает плотность F_2

$$J(h_1, n) = \int_R g(x - y - h_1/\sqrt{n}) f(x) f(y) dx dy,$$

$$J_1 = J(0, n), J_2 = \int_R g^2(x - y) f(x) f(y) dx dy,$$

$$J_3 = \int_R g(x - y) g(x - z) f(x) f(y) f(z) dx dy dz$$

$$J1(h_2, n) = \int_R g(x - y(1 + h_2/\sqrt{n})) f(x) f(y) dx dy,$$

Заметим, что

$$\int_R g'(x-y)f(x)f(y)dxdy = 0, \tag{5}$$

так как функция g(x), по предположению, дифференцируема и симметрична относительно нуля.

В силу того, что функция $g(x)$ предполагается дважды непрерывно дифференцируемой, она может быть представлена в виде

$$g(x) = \psi(x^2), \tag{6}$$

где $\psi(x)$- дважды непрерывно дифференцируемая функция.

Обозначим

$$J^*(h_1) = \frac{1}{2}h_1^2 \int_R g''(x-y)f(x)f(y)dxdy,$$

$$J1^*(h_2) = \frac{1}{2}h_2^2 \int_R (y^2 - (x-y)^2/2)g''(x-y)f(x)f(y)dxdy$$

Обозначим

$$b_1^2 = |J^*(h_1)|, \tag{7}$$

$$b_2^2 = |J1^*(h_2)| \tag{8}$$

Для упрощения обозначений будем рассматривать случай $n = m$. Общий случай рассматривается аналогичным образом. Обозначим

$$T_n = T_n(X,Y) = \Phi_{nn}(X,Y).$$

Основным результатом настоящей работы является следующая теорема, которая устанавливает вид предельного распределения величины nT_n и представление для асимптотической эффективности теста. Эта теорема является обобщением Теоремы 3.1 из работы ([3]).

Теорема 3.1 *Рассмотрим задачу проверки гипотезы (1)-(2), где обе функции обладают свойством (3) и имеют плотности распределения симметричные относительно некоторой точки. Тогда*
(i) при условии $n \to \infty$ функция распределения nT_n сходится при H_0 к функции распределения случайной величины

$$(aL)^2 + c, \tag{9}$$

где L - случайная величина, которая имеет стандартное нормальное распределение,

$$c = J_1 - a^2, a^2 = \sqrt{J_2 + J_1^2 - 2J_3}, \tag{10}$$

(ii) Пусть $F_1(x) = F(x)$, где F — произвольная функция распределения с плотностью $f(x)$ симметричной относительно некоторой точки и обладающая свойством (3), $f_2(x) = f(x(1 + h_2/\sqrt{n}) + h_1/\sqrt{n})$, h_1, h_2 - произвольные заданные числа .
Тогда функция распределения nT_n сходится при выполнении гипотезы H_1 к распределению случайной величины

$$(aL)^2 + rL + b^2 + c, \tag{11}$$

где r и b имеют вид (24), *Мощность критерия nT_n с уровнем значимости α асимптотически равна*

$$Pr\{(aL)^2 + rL + b^2 - (az_{1-\alpha/2})^2 \geq 0\}, \tag{12}$$

где $z_{1-\alpha/2}$ является таким, что

$$Pr\{L \geq z_{1-\alpha/2}\} = \alpha/2.$$

Замечание 3.1 *Величина (11) может быть записана в виде*

$$(aL + b)^2 + \rho(h_1, h_2)L + c, \tag{13}$$

где $b = \sqrt{b_1^2 + b_2^2}$, b_1 имеет вид (7), *b_2 определено в (8), а и с заданы формулой* (10), *$\rho(h_1, h_2)L$ - остаточный член, который предположительно мало влияет на мощность теста. Для $g(x) = x^2/2$ этот член (как показывает непосредственное вычисление)равен нулю.*

Мощность критерия nT_n с уровнем значимости α при отбрасывании остаточного члена $r(h_1, h_2)L$ асимптотически эквивалентна

$$Pr\{L \geq z_{1-\alpha/2} - b/a\} + Pr\{L \leq -z_{1-\alpha/2} - b/a\}, \tag{14}$$

где $b = \sqrt{b_1^2 + b_2^2}$, $z_{1-\alpha/2}$ является таким, что

$$Pr\{L \geq z_{1-\alpha/2}\} = \alpha/2.$$

Прежде, чем доказывать теорему, проиллюстрируем её применение на двух примерах для случая, когда два распределения различаются только параметром масштаба. Рассмотрим случай $g(x) = \ln(1+x^2)$, который уже изучался в работе [2]. Непосредственная проверка показывает, что условие (3) выполняется для нормального распределения и распределения Коши среди многих других. На двух примерах мы демонстрируем, что асимптотические формулы дают хорошее соответствие эмпирическим оценкам мощности для случая, когда распределения различаются (только) параметром масштаба. Для этих же примеров, но с распределениями, различающимися только сдвигом, подобное соответсвие было установлено в статье [3].

Пример 3.1 *Нормальное распределение. Пусть $f(x)$ - функция плотности стандартного нормального распределения, $h_1 = 0$, $\alpha = 0.05$. Численное интегрирование дает следующие результаты*

$$J_1 = 0.810113, J_2 = 1.155022, J_3 = 0.763368.$$

Вычисляя коэффициент a по формуле из теоремы 3.1, получаем a = 0.7303767. Также b_2 вычисляем по формуле (8). В таблице 3.1 представлены теоретические значения мощности, вычисленные по формуле (14), и эмпирические мощности, полученные в результате численной обработки данных N= 1000 повторений статистического моделирования двух выборок размера n=100. Критическое значение критерия T_n вычислялось с помощью 700 случайных перестановок.

Пример 3.2 *Распределение Коши. Пусть $f(x)$ - плотность стандартного распределения Коши, $h_1 = 0$, $\alpha = 0.05$, n=100. В этом случае в работе [2] с помощью таблиц интегралов показано, что*

$$J_1 = \ln 9, \bar{b} = \frac{1}{3}.$$

Таблица 1. Значение эмпирической (Э) и асимптотической (А) мощности для нормального распределения

h_2	1.0	2.0	3.0	4.0	5.0	6.0	7.0	9.0
Э.Мощность								
А.Мощность								

Численным интегрированием находим

$$J_2 = 9.577512,\ J_3 = 6.881056.$$

По теореме 3.1 получаем, что a=0.8955417. Теоретические и эмпирические значения мощности представлены в таблице 3.2. Эмипирические мощности вычислялись тем же способом, что в Примере 3.1. Положим $v = h_2/\sqrt{n}$.

Таблица 2. Мощности для распределения Коши

v	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.9
Э.Мощность	0.074	0.108	0.172	0.272	0.376	0.482	0.587	0.760
А.Мощность	0.061	0.095	0,160	0.247	0.349	0.476	0.592	0.804

Доказательство теоремы. Часть (i) теоремы 3.1 совпадает с частью (i) теоремы 3.1 из статьи [3].

Докажем часть (ii). Пусть гипотеза H_1 имеет место. Положим сначала $h_2 = 0, h_1 = h$. Тогда при выполении гипотезы H_1 имеем $f_1(x) = f(), f_2(x) = f(x + h/\sqrt{n})$. Введем новые случайные величины

$$\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n), \tilde{Y}_i = Y_i + h/\sqrt{n}), i = 1, 2, \ldots, n.$$

При выполнении гипотезы H_1 величины $\tilde{Y}_i, i = 1, 2, \ldots, n$ - независимые случайные величины с плотностью распределения $f(y)$). Заметим, что

$$nT_n(X, Y) - nT_n(X, \tilde{Y}) = [\frac{1}{n^2} \sum_{i=1.j=1}^{n} [ng((X_i - \tilde{Y}_j) + h/\sqrt{n}) - g((X_i - \tilde{Y}_j)]]. \qquad (15)$$

Найдём предел

$$\lim_{n\to\infty} n[T_n(X, \tilde{Y}) - T_n(X, Y)] \qquad (16)$$

Положим $\beta = \sqrt{n}$. Так как функция $g(x)$, по предположению, симметрична и дважды непрерывно дифференцируема, то

$$g((x - y) - h/\sqrt{n}) - g(x - y) = \beta g'_\beta((x - y) + h\beta)|_{\beta=0}$$
$$+\beta^2 \frac{1}{2} g''_\beta((x - y) + h\beta|_{\beta=0} + o(\beta^2).$$

Положим $x = X_i, y = \tilde{Y}_j$ и просуммируем это соотношение по i и j $i = 1, \ldots, n, j = 1, \ldots, n$. Применяя центральную предельную теорему и закон больших чисел для U-статистик (см. [5]) получаем,что искомый предел имеет вид

$$r_1 L + b_1^2,$$

где L - случайная величина со стандартным нормальным распределением,

$$r_1 = h_1\bar{r_1}, b_1^2 = h_1^2\bar{b_1}^2,$$

$\bar{r}_1, \bar{b}_1$ записываются как интегралы, которые нужно найти численно.

А именно, для коэффициета r (он равен корню дисперсии предельной случайной величины) с учетом равенств

$$E(g'(X_i - \tilde{Y}_j)(g'(X_l - \tilde{Y}_k) = 0 \tag{17}$$

если $i \neq l$ и/или $j \neq k$ (в силу симметричности функции g(x)) получаем

$$\bar{r}_1^2 = \lim_{n\to\infty} E[\sum_{i,j=1}^{n} g'(X_i - \tilde{Y}_j)]^2/n^2 = \int_R (g'(x-y))^2 f(x)f(y)ddy. \tag{18}$$

А для коэффициента b_1 по закону больших чисел для U-статистик получаем $b_1^2 = J^*(h_1)$, где

$$J^*(h_1) = \frac{1}{2}h_1^2 \int_R g''(x-y)f(x)f(y)dxdy,$$

Рассмотрим теперь случай $h_1 = 0$. Введем новые случайные величины

$$\tilde{Y}_i = Y_i/(1 + h_2/\sqrt{n}), i = 1, 2, \ldots, n.$$

При выполнении гипотезы H_1 величины $\tilde{Y}_i, i = 1, 2, \ldots, n$ - независимые случайные величины с функцией распределения $F_1(x)$. Заметим, что в рассматриваемом случае

$$nT_n(X,Y) - nT_n(X,\tilde{Y}) = [\frac{1}{n^2}\sum_{i=1.j=1}^{n} n[g((X_i - \tilde{Y}_j) + h_2\tilde{Y}_j/\sqrt{n}) - g((X_i - \tilde{Y}_j)]]-$$

$$[\frac{1}{n(n-1)/2}\sum_{i=1.j=2,j>i}^{n} n[g((\tilde{Y}_i - \tilde{Y}_j) + h/(\tilde{Y}_i - \tilde{Y}_j\sqrt{n}) - g((\tilde{Y}_i - \tilde{Y}_j)]] \tag{19}$$

Используя закон больших чисел, центральную предельную теорему для U-статистик и разложение T_n в ряд по степеням $h/\sqrt{n}$, найдём предел

$$\lim_{n\to\infty} n[T_n(X,\tilde{Y}) - T_n(X,Y)]. \tag{20}$$

Искомый предел имеет вид

$$r_2 L + b_2^2$$

где L - случайная величина со стандартным нормальным распределением,

$$r_2 = h_2\hat{r}_2, b_2^2 = h_2^2\hat{b}_2^2.$$

Используя ЦПТ получаем формулу

$$\hat{r}_2^2 = \int_R (yg'(x-y))^2 f(x)f(y)dxdy +$$
$$\frac{1}{4}\int_R ((x-y)g'(x-y))^2 f(x)f(y)dxdy+$$

$$\int_R y(z-y)g'(x-y))g'(z-y)f(x)f(y)f(z)dxdydz, \tag{21}$$

Вид интеграла для $\hat{b}^2$ определяется второй производной для функции g:

$$\hat{b}^2 = \frac{1}{2}\int_R (y^2 - (x-y)^2/2)g''(x-y)f(x)f(y)dxdy$$

Для случая, когда $f(x)$ - плотность стандартного распределения Коши, получаем

$$\hat{b} = 1/3, \hat{r} = \hat{r}^2 = (4+15-10)/9 = 1.$$

Таким образом, получаем переходом к пределу при $n \to \infty$, что функция распределение nT_n сходится при выполнении гипотезы H_1 к распределению случайной величины

$$(aL)^2 + r_2 L + b_2 + c. \tag{22}$$

Пусть $h_1 h_2 \neq 0$. Положим

$$\tilde{Y}_i = Y_i/(1 + h_2/\sqrt{n}) - h_1/\sqrt{n}, i = 1, 2, \ldots, n.$$

Повторяя проведенные выше вычисления для этих величин, получим, что распределение nT_n сходится при выполнении гипотезы H_1 к распределению случайной величины

$$(aL)^2 + r_2 L + b_2 + c, \tag{23}$$

где

$$b_2^2 = b_1^2 + b_2^2, r^2 = r_1^2 + r_2^2 + h_1 h_2 r_{12}, r_{12} = \int_R (g'(x-y))^2 (y - (x-y)/2))f(x)f(y)ddy. \tag{24}$$

Поскольку этв величина является квадратичной функцией величины L, асимптотическая мощность имеет вид, указанный в формулировек теоремы 3.1.

4. Заключение

В данной работе получено асимптотическое распределение рассматриваемого критерия и найдена формула для асимптотической мощности в случае функций g общего вида для альтернативных распределений, отличающихся от нулевого величиной параметра сдвига и/или параметра масштаба. С помощью статистического моделирования установлено, что найденная формула позволяет получать теоретические значения мощности, которые статистически незначимо отличаются от эмпирических мощностей, найденных с помощью моделирования. Полученные результаты могут быть использованы для определения рационального размера выборки.

Литература

1. Zech, G., Aslan, B.: New test for the multivariate two-sample problem based on the concept of minimum energy. Journal of Statistical Computation and Simulation 75(2), pp. 109-119 (2005).

2. Melas V. and Salnikov D.: On Asymptotic Power of the New Test for Equality of Two Distributions. In: A. N. Shiryaev et al. (eds.), Recent Developments in Stochastic Methods and Applications, Springer Proceedings in Mathematics and Statistics 371, pp. 204-214 (2021).

3. Мелас В.Б. Об асимптотической мощности одного метода проверки гипотез о равенстве распределений. Вестник СПбГУ, Математика. Механика. Астрономия, 10(2), 249-258(2023).

4. Rocha, D.F.S., Bittencourt, I.I., de Amorim Silva, R. et al. An assistive technology based on Peirce's semiotics for the inclusive education of deaf and hearing children. Univ Access Inf Soc (2022).

5. Hoeffding W.: A class of statistics with asymptotically normal distribution. Ann. Math. Statistics 19, pp. 293-325 (1948).