МАТЕМАТИКА

УДК 519.234.3
MSC 62G10

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

В. Б. Мелас, Д. И. Сальников

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Мелас В. Б., Сальников Д. И.* Об асимптотической мощности «энергетического» теста для проверки гипотез о равенстве двух распределений // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2024. Т. 11 (69). Вып. 3. С. 477–488. https://doi.org/10.21638/spbu01.2024.304

В статье найдены асимптотическое распределение «энергетического» критерия проверки гипотез о равенстве двух распределений и формула для асимптотической мощности критерия в случае альтернативных распределений, отличающихся от нулевого величиной параметра сдвига и (или) параметра масштаба. Этот критерий является конкурентом широко известного критерия Манна — Уитни и в отличие от него позволяет сравнивать распределения, отличающиеся параметром масштаба. Эффективность полученных результатов продемонстрирована с помощью статистического моделирования на примере нормального распределения и распределения Коши.

Ключевые слова: проверка гипотез о равенстве распределений, «энергетический» критерий, асимптотическая мощность критерия, нормальное распределение, распределение Коши.

1. Постановка задачи. 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

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

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

Если два распределения отличаются только сдвигом, часто наиболее мощным является тест Вилкоксона — Манна — Уитни. Однако хорошо известно, что этот тест не позволяет дискриминировать распределения, различающиеся 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 сдвигом и (или) преобразованием масштаба. Задачи, в которых важно учитывать возможность различия в параметре масштаба, возникают во многих практических областях применения, включая физиологию и психологию (см., например, недавнюю работу [3]).

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. В дальнейшем мы уточним предположения относительно вида функции g, которые позволят получить аналитические результаты об асимптотической мощности теста. Этот тест был, по-видимому, впервые введен в работе [1] и назван «энергетическим». 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

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

С помощью статистического моделирования было показано, 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 [4] 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.

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

Пусть $f(x)$ обозначает плотность F_1,

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

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

$$J_3 = \int_R g(x - y) g(x - z) f(x) f(y) f(z) dxdydz.$$

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. Обозначим

$$J_1^*(h_1) = \frac{1}{2} h_1^2 \int_R g''(x - y) f(x) f(y) dxdy,$$

$$J_2^*(h_2) = \frac{1}{2} h_2^2 \int_R (y^2 - (x - y)^2/2) g''(x - y) f(x) f(y) dxdy,$$

$$b_1^2 = |J_1^*(h_1)|, \tag{6}$$

$$b_2^2 = |J_2^*(h_2)|. \tag{7}$$

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 мощности теста. Эта теорема является обобщением теоремы 3.1 из работы [4].

Теорема 1. *Рассмотрим задачу проверки гипотезы (1)–(2), где обе функции обладают свойством (3) и имеют плотности распределения, симметричные относительно некоторой точки. Пусть $g(0) = 0$, $g(x) = \psi(x^2)$, где ψ есть монотонная дважды непрерывно дифференцируемая функция. Тогда*

(i) при условии $n \to \infty$ функция распределения nT_n сходится при H_0 к функции распределения случайной величины

$$(aL)^2 + c, \tag{8}$$

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

$$c = J_1 - a^2,\ a^2 = \sqrt{J_2 + J_1^2 - 2J_3}. \tag{9}$$

(ii) Пусть $F_1(x) = F(x)$, $F_2(x) = F(x(1 + h_2/\sqrt{n}) + h_1/\sqrt{n})$, где F — произвольная функция распределения с плотностью $f(x)$, симметричной относительно некоторой точки, и обладающая свойством (3); h_1, h_2 — произвольные заданные числа.

Тогда функция распределения nT_n сходится при выполнении гипотезы H_1 к распределению случайной величины

$$(aL + \sqrt{b_1^2 + b_2^2})^2 + c, \tag{10}$$

где b_1 имеет вид (6), b_2 определено в (7), а и с заданы формулой (9).

Мощность критерия nT_n с уровнем значимости α асимптотически равна

$$Pr\{L \geq z_{1-\alpha/2} - b/a\} + Pr\{L \leq -z_{1-\alpha/2} - b/a\}, \tag{11}$$

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 ее 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 соответствие было установлено в статье [4].

Пример 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 1, получаем $a = 0.7303767$. Также b_2 вычисляем по формуле (7). В табл. 1 представлены теоретические значения мощности, вычисленные по формуле (11), и эмпирические мощности, полученные в результате численной обработки данных $N = 1000$ повторений статистического моделирования двух выборок размера $n = 100, 400, 900, 1600$. Критическое значение критерия T_n вычислялось с помощью 700 случайных перестановок.

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

$$J_1 = \ln 9.$$

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

	h_2	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Э	$n = 100$	5.3	6.7	8.5	12.5	22.1	30.2	40.2	50.3	61.9	71.1
Э	$n = 400$	5.5	6.8	9.4	13.6	22.7	33.8	46.4	58.6	71.4	82.5
Э	$n = 900$	5.9	7.7	10.4	14.8	23.5	34.2	47.9	62.8	74.9	85
Э	$n = 1600$	6.4	8.5	11.7	16.9	25.1	36.3	49.4	64.7	77	86.8
А		5.9	8.5	13	19.5	27.7	37.4	48.2	58.9	69.1	78

Численным интегрированием находим

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

По теореме 1 получаем $a = 0.8955417$. Теоретические и эмпирические значения мощности представлены в табл. 2. Эмпирические мощности вычислялись тем же способом, что и в примере 1.

Таблица 2. **Значение эмпирической (Э) и асимптотической (А) мощности для распределения Коши**

	h_2	1	2	3	4	5	6	7	8	9	10
Э	$n = 100$	6.6	10.5	17.9	26.9	37.3	49.5	59.8	68.1	77.2	83.8
Э	$n = 400$	6.3	10	19.3	29.7	41.4	54.4	67.9	78.5	86.1	91.5
Э	$n = 900$	6.2	10.9	21.6	32.1	45.4	61.6	75.3	85.8	91.9	95.5
Э	$n = 1600$	6.5	11	20.3	32.2	46.8	62.7	77.3	85.2	93.1	96.7
А		6.6	11.6	20.1	31.9	46.2	60.8	74	84.6	91.8	96.1

Доказательство теоремы 1. Сначала докажем, что теорема верна для $g(x) = \frac{1}{2}x^2$. В этом случае критерий может быть записан в более простом виде.

Лемма 1. *Рассмотрим тест* (4) *с функцией* $g(x) = g^*(x) = \frac{1}{2}x^2$. *Если условие* (3) *выполняется для* $g(x) = g^*(x)$, *то справедливо следующее тождество:*

$$\Phi_{nn} = \frac{1}{2}(\bar{x} - \bar{y})^2,$$

где

$$\bar{x} = (\sum_{i=1}^{n} X_i)/n, \bar{y} = (\sum_{i=1}^{n} Y_i)/n.$$

При $n \to \infty$ *распределение величины* $\sqrt{n\Phi_{nn}}$ *сходится к нормальному распределению с дисперсией* J_1 *и математическим ожиданием* 0 *при выполнении гипотезы* H_0.

Доказательство этой леммы можно найти в работе [4]. Для удобства читателя и в связи с тем, что этот результат имеет ключевое значение для доказательства теоремы, мы приводим полное доказательство этой леммы в приложении.

В силу леммы 1 при $g(x) = \frac{1}{2}x^2$ при выполнении гипотезы H_0 величина nT_n имеет вид (8) при $a^2 = J_1$. Таким образом, утверждение (i) теоремы в этом случае справедливо. Для доказательства части (ii) теоремы заметим, что в силу леммы 1 при $g(x) = \frac{1}{2}x^2$ при выполнении гипотезы H_1 величина nT_n имеет асимптотически

то же распределение, что величина $(aL+b)^2$, где $b=h_1$. Непосредственным вычислением можно проверить, что $b_2=0$ для любого h_2. Утверждение о мощности вытекает из того, что $(\sqrt{nT_n}-b)$ имеет нормальное распределение с дисперсией J_1.

Рассмотрим далее случай функции $g(x)$ общего вида. Имеет место следующая лемма.

Лемма 2. *(i) В условиях части (i) теоремы 1 распределение величины nT_n сходится при $n\to\infty$ к распределению величины*

$$a^2L^2+c, \tag{12}$$

где L имеет стандартное нормальное распределение; a и c — некоторые числа.

(ii) В условиях части (ii) теоремы 1 распределение величины nT_n сходится при $n\to\infty$ к распределению величины

$$(aL+b)^2+c,\ b=\sqrt{b_1^2+b_2^2}, \tag{13}$$

где L имеет стандартное нормальное распределение; a и c те же самые, что в части (i) леммы; b_1 определено равенством (6); b_2 определено в (7).

Доказательство леммы 2 дано в приложении.

Обозначим

$$R(L,a,c)=a^2L^2+c.$$

Согласно лемме 2, часть (i) есть предельное распределение nT_n при выполнении H_0. Прямое вычисление показывает, что

$$\mathbf{E}(nT_n)=J_1.$$

Заметим, что величину $(nT_n)^2$ можно представить как

$$\frac{1}{n^2}\{\sum_{i,j=1,\dots,n}(g(X_i-Y_j)-J_1)-\sum_{1\le i<j\le n}(g(X_i-X_j)-J_1)-\sum_{1\le i<j\le n}(g(Y_i-Y_j)-J_1)+nJ_1\}^2. \tag{14}$$

Используя формулу (14), представим предел $(nT_n)^2$ при $n\to\infty$ в виде

$$\lim_{n\to\infty}(nT_n)^2=J_1^2+\lim_{n\to\infty}(\hat{J}_1(n)-J_1)^2+\lim_{n\to\infty}\frac{1}{n^2}\{2n(\hat{J}_1(n)-J_1)V_n+V_n^2\}, \tag{15}$$

где

$$\hat{J}_1(n)=\sum_{i=1}^{n}\frac{1}{n}(g(X_i-Y_i),$$

$$V_n=\{2\sum_{1\le i<j\le n}(g(X_i-Y_j)-J_1)-\sum_{1\le i<j\le n}(g(X_i-X_j)-J_1)-\sum_{1\le i<j\le n}(g(Y_i-Y_j)-J_1)\}.$$

Прямое вычисление с использованием закона больших чисел для U-статистик показывает, что $\lim_{n\to\infty}\frac{1}{n}V_n=\mathbf{E}\frac{1}{n}V_n=0$. Таким образом,

$$\lim_{n\to\infty}(nT_n)^2=J_1^2+\lim_{n\to\infty}\frac{1}{n^2}V_n^2.$$

Заметим, что V_n^2 является алгебраической суммой произведений вида

$$R_{(i,j),(s,k)} = (g(Z_i - Z_j) - J_1)(g(Z_s - Z_k) - J_1),$$

где $Z = (Z_1, \ldots, Z_{2n}) = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$.

В силу независимости и одинаковой распределенности величин X_i, Y_j, $i, j = 1, \ldots, n$ получаем в силу закона больших чисел, что если все индексы различны, то

$$\mathbf{E}R_{(i,j),(s,k)} = 0,\ i,\ j,\ s,\ k \in \{1, \ldots, 2n\}.$$

Если пары индексов совпадают $(i, j) = (s, k)$, то получаем

$$\mathbf{E}R_{(i,j),(i,j)} = J_2 - J_1^2.$$

В случае, когда совпадает только один из индексов, получаем с помощью предельной теоремы для U-статистик (см. [5]), что имеет место равенство

$$\mathbf{E}R_{(i,j),(i,k)} = J_3 - J_1^2, \tag{16}$$

если $j \neq i, k \neq i, j$. Такое же соотношение выполняется для аналогичных случаев. Соответствующие произведения будем называть произведениями третьего вида.

Заметим, что число произведений вида $R_{(i,j),(i,j)}$ равно $n(2n-1)$. А число сумм третьего вида, входящих в V_n^2 со знаком минус, равно $4n(n-1)^2$, и число произведений со знаком плюс равно $4n(n-1)(n-2)$. Таким образом, математическое ожидание алгебраической суммы таких произведений равно $-4n(n-1)(J_3 - J_1^2)$.

Таким образом,

$$\mathbf{E} \lim_{n\to\infty} (nT_n)^2 = J_1^2 + \mathbf{E} \lim_{n\to\infty} (\frac{1}{n^2} V_n) = J_1^2 + 2(J_2 - J_1^2) - 4(J_3 - J_1^2).$$

Из соотношения

$$\mathbf{E} \lim_{n\to\infty} (nT_n)^2 = \mathbf{E}(R(L, a, c))^2 = 2a^4 + J_1^2$$

получаем, переходя к пределу при $n \to \infty$, что $a^2 = \sqrt{J_2 + J_1^2 - 2J_3}$.

Пусть теперь имеет место H_1, тогда по лемме 2 часть (ii) функции распределения величины nT_n сходится при выполнении H_1 при $n \to \infty$ к функции распределения величины

$$(aL + b)^2 + c,$$

где $b = \sqrt{b_1^2 + b_2^2}$; a и c те же самые, что в части (i). Из этого представления вытекает формула для асимптотической мощности, что завершает доказательство теоремы. □

3. Приложение. Доказательство леммы 1. Обозначим

$$Z = (X, Y) = (X_1, \ldots, X_n, Y_1, \ldots, Y_n),\ V(Z) = \frac{1}{2} \sum_{i=1}^{2n} \sum_{j=1}^{2n} (Z_i - Z_j)^2.$$

Доказательство следует из известной формулы (см., например, [5, с. 296])

$$\frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (X_i - X_j)^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{x})^2 \tag{17}$$

и очевидного тождества

$$\sum_{i=1}^{2n}\sum_{j=1}^{2n}(Z_i - Z_j)^2 = \sum_{i,j=1}^{n}(X_i - X_j)^2 + \sum_{i,j=1}^{n}(Y_i - Y_j)^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - Y_j)^2, \qquad (18)$$

путем прямых, но нетривиальных вычислений.

Действительно, давайте использовать стандартную форму записи:

$$S_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{x})^2.$$

S_y^2 и S_z^2 будем понимать аналогичным образом. Обозначим

$$S_{xy} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - Y_j)^2.$$

Используя формулы (17), получаем

$$V(Z) = 2n[\sum_{i=1}^{n}(X_i - (\bar{x}+\bar{y})/2)^2 + \sum_{j=1}^{n}(Y_i - (\bar{x}+\bar{y})/2)^2] = 2n(n-1)(S_x^2 + S_y^2) + n^2(\bar{x}-\bar{y})^2. \qquad (19)$$

Из (17) и (18) получаем

$$n^2 S_{xy} = V(Z) - n(n-1)(S_x^2 + S_y^2). \qquad (20)$$

Следовательно,

$$S_{xy} = \frac{1}{n}(n-1)(S_x^2 + S_y^2) + (\bar{x} - \bar{y})^2,$$

и мы получаем

$$\Phi_{nn} = [S_{xy} - \frac{1}{n}(n-1)(S_x^2 + S_y^2)]/2 = (\bar{x} - \bar{y})^2/2.$$

По классической центральной предельной теореме величина $\sqrt{n\Phi_{nn}}$ имеет распределение, сходящееся при выполнении H_0 для $n \to \infty$ к нормальному распределению с дисперсией J_1 и математическим ожиданием 0. Таким образом, лемма 1 доказана. □

Из этой леммы следует, что критерий Φ_{nn} в этом случае эквивалентен критерию $(\bar{x} - \bar{y})^2$.

Доказательство леммы 2. Начнем с доказательства части (i). В этой части предполагается выполненной гипотеза H_0. Мы проведем доказательство в форме, которая допускает естественное обобщение на случай, когда выполнена гипотеза H_1. Заметим, что при $g(x) = g^*(x) = x^2/2$ в силу леммы 1 и центральной предельной теоремы при $n \to \infty$ распределение nT_n сходится к распределению величины

$$(aL + b)^2, \qquad (21)$$

где $a = 1$; $b = h_2$ при выполнении гипотезы H_0 или гипотезы H_1, причем $b = 0$ для гипотезы H_0. Покажем, что этот факт остается верным для симметричной дважды

непрерывно дифференцируемой функции $g(z) = g(x-y)$, $x, y \in R^1$ общего вида, удовлетворяющей дополнительным условиям, сформулированным в теореме 1, для некоторых чисел a и b. Обозначим

$$U_n(u,g) = (n(n-1)/2)^{-1} \sum_{1 \le u_i < u_j \le n} g(u_i - u_j),\ u = (u_1, \dots, u_n). \tag{22}$$

Функция $U_n(u,g)$ есть по определению (см. [5]) U-статистика. Напомним, что мы приняли $m = n$ и

$$\Phi_{AB} = \Phi_{AB}(X,Y,g) = -\frac{1}{n^2} \sum_{i,j=1}^{n} g(X_i - Y_j),$$

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

Следовательно,

$$\Phi_A(X,g) = -\frac{1}{2}\frac{n-1}{n} U_n(X,g),\ \Phi_B(Y,g) = -\frac{1}{2}\frac{n-1}{n} U_n(Y,g),$$

$$\Phi_{AB}(X,Y,g) = \frac{n-1}{n} U_{2n}(Z,g) - \frac{1}{2}\frac{n-1}{n} U_n(X,g) - \frac{1}{2}\frac{n-1}{n} U_n(Y,g), \tag{23}$$

где $Z = (Z_1, \dots, Z_{2n}) = (X_1 \dots, X_n, Y_1, \dots, Y_n)$. Применим предельную теорему (см. теорему 7.1 [5]) к каждому из выражений $\Phi_A(X,g)$, $\Phi_B(Y,g)$ и $\Phi_{AB}(X,Y,g)$). По этой теореме величины $n^{\frac{1}{2}} U_n(X,g)$ и $n^{\frac{1}{2}} U_n(X,g^*)$ имеют в пределе нормальное распределение с нулевым средним. Заметим, что нормальные распределения полностью определяются параметрами сдвига и масштаба. Более того, в силу монотонности функции ψ имеет место равенство

$$2(n^{\frac{1}{2}}/n^2) \sum_{1 \le i < j \le n} g(X_i - X_j) = 2a^2 (n^{\frac{1}{2}}/n^2) \sum_{1 \le i < j \le n} g^*(X_i - X_j) + \tilde{\eta}_n, \tag{24}$$

где a — некоторое число, а величина $\tilde{\eta}_n$ есть случайная величина такая, что $\sqrt{n}\tilde{\eta}_n$ сходится по распределению к постоянной, равной нулю. И так как величины $X_i - X_j$ и $Y_i - Y_j$, $1 \le i < j \le n$ имеют одно и то же распределение, это равенство имеет место при замене $X_i - X_j$ на $Y_i - Y_j$ с той же самой постоянной a.

По той же причине и принимая во внимание (23), мы получаем формулу

$$2(n^{\frac{1}{2}}/n^2) \sum_{i,j=1}^{n} g(X_i - Y_j) = 2a^2 (n^{\frac{1}{2}}/n^2) \sum_{i,j=1}^{n} g^*(X_i - Y_j) + \bar{\eta}_n,$$

где постоянная a такая же, как в (24), но $\bar{\eta}_n \ne \tilde{\eta}_n$. И мы получаем

$$nT_n(X,Y,g) = a^2 nT_n(X,Y,g^*) + c + \eta_n,$$

где η_n сходится по вероятности к 0, и a есть то же самое, что в формуле (24),

$$c = \mathbf{E} nT_n(X,Y,g) - a^2 \mathbf{E} nT_n(X,Y,g^*).$$

По лемме 1 $nT_n(X, Y, g^*)$ сходится по распределению к L^2. Таким образом, предельное распределение $nT_n(X, Y, g)$ имеет вид $a^2L^2 + c$.

Рассмотрим теперь случай, когда условие (3) не выполняется для $g(x) = \frac{1}{2}x^2$.

Пусть K — произвольное положительное число,

$$\tilde{X} = (\tilde{X_1}, \dots, \tilde{X_n}),\ \tilde{Y} = (\tilde{Y_1}, \dots, \tilde{Y_n}),$$

где $\tilde{X_i} = X_i$, если $|X_i| \le K$ и $\tilde{X_i} = K$ и $X_i > 0$, $\tilde{X_i} = -K$, $X_i < 0$ в противном случае. Пусть $\tilde{Y_i}$ определены подобным образом. Через $\tilde{f}$ обозначим плотность распределения $\tilde{X_1}$ (которая есть также плотность распределения $\tilde{X_i}$, $i = 2, \dots, n$ и $\tilde{Y_i}$, $i = 1, \dots, n$). Теперь условие (3) очевидно выполняется для заданной функции g и для $g = g^*$ с функцией $f(x)$, замененной на $\tilde{f}(x)$.

Рассмотрим величину

$$n\{\frac{1}{n^2}\sum_{i,j=1}^{n} g(\tilde{X_i} - \tilde{Y_j}) - \frac{1}{n^2}\sum_{i<j} g(\tilde{X_i} - \tilde{X_j}) - \frac{1}{n^2}\sum_{i<j} g(\tilde{Y_i} - \tilde{Y_j})\}. \tag{25}$$

В силу представленных выше аргументов предельное распределение этой величины имеет вид $a^2L^2 + c$. При $K \to \infty$ предельное распределение существует (по теореме 7.1 [5]) и имеет такой же вид. Таким образом, часть (i) леммы 2 доказана.

Часть (ii) доказывается следующим образом. Обозначим

$$\tilde{Y} = (Y_1 - h_1/\sqrt{n})/(1 + Y_1h_2/\sqrt{n}), \dots, (Y_n - h_1/\sqrt{n})/(1 + Y_nh_2/\sqrt{n})).$$

При выполнении гипотезы H_1 величины $\tilde{Y_i}, i = 1, 2, \dots, n$ — независимые случайные величины с функцией распределения $F_1(x)$.

Рассмотрим величину

$$n[T_n(X, Y) - T_n(X, \tilde{Y})]. \tag{26}$$

Изучим предел этой величины при $n \to \infty$. Для упрощения обозначений рассмотрим случай $h_2 = 0$ (общий случай рассматривается аналогично). При $h_2 = 0$ величины $n\Phi_A(X, Y)$ и $n\Phi_B(X, Y)$ не зависят от h_1. Следовательно, величина (26) имеет вид

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i - Y_j) - g((X_i - \tilde{Y_i}) - h_1/\sqrt{n}). \tag{27}$$

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

$$g((x - y) - h_1/\sqrt{n}) = g(x - y) + \beta g'_\beta((x - y) - h_1\beta)|_{\beta=0} + \frac{1}{2}\beta^2 g''_\beta((x - y) - h_1\beta)|_{\beta=0} + o(\beta^2).$$

В силу закона больших чисел и центральной предельной теоремы получаем, что величина (26) сходится к пределу вида

$$r\tilde{L} + (b_1)^2,$$

где r — некоторое число, а $\tilde{L}$ — случайная величина со стандартным нормальным распределением. Так как в силу симметричности функции ψ величина T_n симметрична относительно нуля, то получаем

$$\lim_{n\to\infty} nT = (aL + b)^2 + c,$$

где a и c такие же, как в части (i), b определено в теореме 1. □

4. Заключение. Получены асимптотическое распределение «энергетического» критерия и формула для асимптотической мощности в случае альтернативных распределений, отличающихся от нулевого величиной параметра сдвига и (или) параметра масштаба. С помощью численных экспериментов установлено, что найденная формула позволяет получать теоретические значения мощности, которые для больших выборок отличаются не более чем на 10 % от эмпирических мощностей. 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), 109–119 (2005). https://doi.org/10.1080/00949650410001661440

2. Melas V., 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**, 204–214 (2021). https://doi.org/10.1007/978-3-030-83266-7_15

3. Rocha D. F. S., Bittencourt I. I., de Amorim Silva R., Ospina P. L. E. An assistive technology based on Peirce's semiotics for the inclusive education of deaf and hearing children. *Univ. Access Inf. Soc.* **22**, 1097–1116 (2023). https://doi.org/10.1007/s10209-022-00919-2

4. Мелас В. Б. Об асимптотической мощности одного метода проверки гипотез о равенстве распределений. *Вестник Санкт-Петербургского университета. Математика. Механика* **10** (2), вып. 2, 249–258 (2023). https://doi.org/10.21638/spbu01.2023.206

5. Hoeffding W. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics* **19**, 293–325 (1948). https://doi.org/10.1007/978-1-4612-0919-5_20

Статья поступила в редакцию 13 января 2024 г.;
доработана 31 января 2024 г.;
рекомендована к печати 22 февраля 2024 г.

Контактная информация:

Мелас Вячеслав Борисович — д-р физ.-мат. наук, проф.; vbmelas@yandex.ru
Сальников Дмитрий Игоревич — аспирант; mejibkop.ru@gmail.com

On the asymptotic power of the "energy" test for equality of two distributions

V. B. Melas, D. I. Salnikov

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Melas V. B., Salnikov D. I. On the asymptotic power of the "energy" test for equality of two distributions. *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy*, 2024, vol. 11 (69), issue 3, pp. 477–488. https://doi.org/10.21638/spbu01.2024.304 (In Russian)

In the paper the asymptotic distribution and the formula for the asymptotic power are found for the "energy" test in the case of alternative distributions differ from zero distribution by the parameter of shift and/or the parameter of scale. This criterion is an alternative for the well-known Mann — Whitney test but allows to compare distributions that differ by the scale parameter. The efficiency of the results obtained is demonstrated by the stochastic simulation for the normal distribution and the Cauchy distribution.

Keywords: testing hypothesis of equality of two distributions, "energy" test, the asymptotic power of criteria, Normal distribution, Cauchy distribution.

References

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), 109–119 (2005). https://doi.org/10.1080/00949650410001661440

2. Melas V., 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**, 204–214 (2021). https://doi.org/10.1007/978-3-030-83266-7_15

3. Rocha D. F. S., Bittencourt I. I., de Amorim Silva R., Ospina P. L. E. An assistive technology based on Peirce's semiotics for the inclusive education of deaf and hearing children. *Univ. Access Inf. Soc.* **22**, 1097–1116 (2023). https://doi.org/10.1007/s10209-022-00919-2

4. Melas V. B. On the Asymptotic Power of a Method for Testing Hypotheses on the Equality of Distributions. *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy* **10** (2), iss. 2, 249–258 (2023). https://doi.org/10.21638/spbu01.2023.206 (In Russian) [Eng. transl.: *Vestnik St. Petersburg University. Mathematics* **56**, iss. 2, 182–189 (2023). https://doi.org/10.1134/S1063454123020115].

5. Hoeffding W. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics* **19**, 293–325 (1948). https://doi.org/10.1007/978-1-4612-0919-5_20

Received: January 13, 2024
Revised: January 31, 2024
Accepted: February 22, 2024

Authors' information:

Vyacheslav B. Melas — vbmelas@yandex.ru
Dmitrii I. Salnikov — mejibkop.ru@gmail.com