ISSN 2074-1863 Уфимский математический журнал. Том 1. № 4 (2009). С. 3-23.

УДК 519.234.3

ПРОВЕРКА ГИПОТЕЗ ОБ ОДНОРОДНОСТИ И СИММЕТРИЧНОСТИ РАСПРЕДЕЛЕНИЙ ДЛЯ МНОГОМЕРНЫХ ДАННЫХ

Н.К. БАКИРОВ

Аннотация. Рассматриваются задачи проверки непараметрических гипотез для многомерных данных.

Ключевые слова: проверка непараметрических гипотез для многомерных данных.

1. Введение

В настоящей работе изучаются новые критерии проверки непараметрических гипотез для многомерных данных: об однородности распределений двух случайных векторов и симметричности многомерного распределения. Построенные тестовые статистики имеют простую структуру и инвариантны к линейным преобразованиям данных, после надлежащей нормировки их распределения слабо сходятся при нулевой гипотезе к распределениям типа омега-квадрат. Построенные критерии состоятельны против широкого класса альтернатив при минимальных моментных ограничениях.

2. Проверка симметричности

Пусть $X_1, X_2, \dots, X_n$ — повторная выборка с общей функцией распределения (ф.р.) $F(x)$. Для проверки гипотезы о симметричности:

$$H_0 : 1 - F(x+0) - F(-x) = 0 \quad \text{для всех } x \in R^1 \tag{1}$$

в одномерном случае используют известные критерии: w^2, Колмогоровского типа, Уотсона-Дарлинга, Хилла-Рао, знаковых статистик и др. [1]. Асимптотические уровни значимости могут быть найдены в предположении, что ф.р. $F(x)$ непрерывна. К сожалению, асимптотическое распределение многомерных аналогов этих статистик зависит от ф.р. F, так что они не являются подобными. Задачи проверки симметричности многомерных распределений рассматривались, в частности, в работах [2]–[10]. В настоящем параграфе мы строим тесты для проверки многомерной симметричности ф.р. с заданным асимптотическим уровнем значимости, инвариантные к линейным преобразованиям данных.

Итак, пусть $X_1, X_2, \dots, X_n$, — повторная выборка $X_k \in R^d, X_1 \not\equiv 0, a.s., E|X_1| < \infty$. Мы рассматриваем гипотезы о симметричности распределения выборки с центром в нуле:

H_{01} : *диагональная симметричность* или $X_1 \stackrel{D}{=} -X_1$;

H_{02} : *сферическая симметричность* или $X_1 \stackrel{D}{=} CX_1$ для всех ортогональных матриц C;

H_{03} : *эллиптическая симметричность*, то есть случайный вектор MX_1 сферически симметричен для некоторой положительно определенной матрицы M, а также варианты этих гипотез с неизвестными центрами симметрии.

N.K. Bakirov, Testing homogeneity and symmetry for multivariate data.

Поступила 01 июня 2009 г.

Все наши построения базируются на использовании эмпирических характеристических функций:

$$f_n(t) = \frac{1}{n}\sum_{i=1}^{n}\exp\{i(t, X_i)\}, \tag{2}$$

где (.,.) — скалярное произведение в R^d. В работах [11]-[12] они были применены для проверки гипотезы о независимости двух многомерных распределений.

§ 1. Диагональная симметричность

Пусть $f(t)$ — характеристическая функция случайного вектора X_1. Легко видеть, что

$$H_{01} \Longleftrightarrow f(t) = f(-t), \forall t \in R^d \Longleftrightarrow J(f,\varphi) \stackrel{def}{=} \int_{R^d} |f(t) - f(-t)|^2 \varphi(t)\, dt = 0$$

для любой весовой функции $\varphi(t)$ такой, что интеграл существует и $\varphi(t) > 0$ почти всюду (п.в.). Мы будем предполагать, что $E|X_1| < \infty$.

Рассмотрим статистику

$$J(f_n, \varphi_0) = \int_{R^d} |f_n(t) - f_n(-t)|^2 \varphi_0(t)\, dt, \qquad \varphi_0(t) = |t|^{-1-d}.$$

Заметим, что $|f(t) - f(-t)| \le E|X_1||t|$, поэтому $J(f, \varphi_0)$ имеет интегрируемую особенность в нуле. По аналогичным причинам интеграл $J(f_n, \varphi_0)$ существует. Учитывая тождество

$$\int_{R^d} (1 - \exp\{i(t, X)\})|t|^{-d-\gamma}\, dt = C(d,\gamma)|X|^{\gamma}, \quad \forall \gamma \in (0,2), \tag{3}$$

где $C(d,\gamma)$ — положительная константа, мы можем легко вычислить

$$J(f_n, \varphi_0) = 2C(d,1)n^{-2}\left(\sum_{i,j=1}^{n}|X_i + X_j| - \sum_{i,j=1}^{n}|X_i - X_j|\right)$$

(интеграл (3) сходится в нулевой точке в смысле главного значения, при необходимости, ввиду симметричности функции $\sin(t, X)$, мы можем в (3) заменить $\exp\{i(t, X)\}$ на $\cos(t, X)$ и считать несобственный интеграл (3) сходящимся в обычном смысле). Используя формулу 3.241.4 [13] и рассматривая интеграл (3) как повторный, можно показать, что $C(d,1) = \pi^{\frac{d+1}{2}}/\Gamma\left(\frac{d+1}{2}\right)$. Далее, по закону больших чисел для статистик Мизеса [1]

$$J(f_n, \varphi_0) \xrightarrow[n\to\infty]{P} 2C(d,1)\left(E|X_1 + X_2| - E|X_1 - X_2|\right) =$$

$$= 2C(d,1)\int_{R^d} |f(t) - f(-t)|^2 \varphi_0(t)\, dt. \tag{4}$$

Так что, для всех альтернатив к гипотезе H_{01} мы имеем $J(f_n, \varphi_0) \xrightarrow[n\to\infty]{P} \infty$.

Аналогично

$$G(f_n) \stackrel{def}{=} 2\int_{R^d}(1 - f_n(2t))\varphi_0(t)\, dt = 4C(d,1)n^{-1}\sum_{i=1}^{n}|X_i| \xrightarrow[n\to\infty]{P}$$

$$\xrightarrow[n\to\infty]{P} 4C(d,1)E|X_1| = 2\int_{R^d}(1 - f(2t))\varphi_0(t)\, dt = G(f).$$

Заметим, что $G(f) \neq 0$, если $X_1 \not\equiv 0$ п.в. Рассмотрим теперь асимптотическое поведение статистики $J(f_n, \varphi_0)$ при нулевой гипотезе H_{01}. Обозначим

$$\xi_n(t) = \sqrt{n}\left(f_n(t) - f_n(-t)\right).$$

Нетрудно вычислить, что при $H_{01}, \forall t, s$

$$E\xi_n(t) = 0, \quad E\xi_n(t)\overline{\xi_n(s)} = 2(f(t-s) - f(t+s)). \tag{5}$$

Конечномерные распределения случайного процесса (сл.пр.) $\xi_n(t)$ сходятся при $n \to \infty$ к конечномерным распределениям гауссовского (сл.пр.) $\xi(t)$ с моментами (5) в силу многомерной центральной предельной теоремеы (ЦПТ). Из (5) следует, что

$$E|\xi_n(t) - \xi_n(s)|^2 = E|\xi(t) - \xi(s)|^2 = 4(1 - f(t-s)) +$$
$$+2(2f(t+s) - f(2t) - f(2s)) \le 8|t-s|E|X_1|. \tag{6}$$

Поэтому [14], гауссовский (сл.пр.) $\xi(t)$ имеет модификацию с непрерывными траекториями. С другой стороны,

$$E|\xi_n(t)|^2 = E|\xi(t)|^2 = 2(1 - f(2t)) \le 4\min(1, |t|E|X_1|). \tag{7}$$

Следовательно, $\int_{R^d} E|\xi(t)|^2\varphi_0(t)\,dt < \infty$, поэтому по теореме Фубини мы можем определить $Q \stackrel{def}{=} \int_{R^d} |\xi(t)|^2\varphi_0(t)\,dt$ (для непрерывной модификации $\xi(t)$) как римановский интеграл.

Обозначим $A_1 = \{t : |t| < \varepsilon\}, A_2 = \{t : |t| > 1/\varepsilon\}, A_3 = \{t : \varepsilon \le |t| \le 1/\varepsilon\}$. Применяя (7), мы можем записать

$$E\int_{A_2} (|\xi_n(t)|^2 + |\xi(t)|^2)\varphi_0(t)\,dt \le 8 \int\limits_{|t|>1/\varepsilon} \varphi_0(t)\,dt = 8w_d\varepsilon, \tag{8}$$

где w_d — площадь единичной сферы в R^d. С другой стороны, используя (7), запишем

$$E\int_{A_1} (|\xi_n(t)|^2 + |\xi(t)|^2)\varphi_0(t)\,dt = 4\int\limits_{|t|<\varepsilon} (1 - f(2t))\varphi_0(t)\,dt =$$

$$= 4E|X_1|\delta, \quad \delta = \int\limits_{|t|<\varepsilon|X_1|} (1 - \cos 2t_1)\varphi_0(t)\,dt,$$

здесь $t = (t_1, t_2, \ldots, t_d)$ и мы использовали замену переменных. Очевидно, $0 \le \delta \le 2C(d,1)$ и $\delta \xrightarrow{\text{п.в.}} 0$, если $\varepsilon \to 0$. Итак, по теореме Лебега об ограниченной сходимости

$$4E|X_1|\delta \underset{\varepsilon\to 0}{\longrightarrow} 0. \tag{9}$$

Далее, для любого разбиения $\Delta_k, k = 1, 2, \ldots, N$ множества A_3 ($\max\limits_k \operatorname{diam}\Delta_k = \tau$) и $t_k \in \Delta_k$ имеем

$$\int_{A_3} |\xi_n(t)|^2\varphi_0(t)\,dt = \sum_{k=1}^{N} |\xi_n(t_k)|^2 \int_{\Delta_k} \varphi_0(t)\,dt + \alpha_1, \tag{10}$$

где в силу (6),(7)

$$E|\alpha_1| \le \sum_{k=1}^{N} \int_{\Delta_k} E||\xi_n(t)|^2 - |\xi_n(t_k)|^2|\varphi_0(t)\,dt \le$$

$$\le 8\left(\tau E|X_1|) + \sqrt{2\tau E|X_1|}\right) \int_{A_3} \varphi_0(t)\,dt \underset{\tau\to 0}{\longrightarrow} 0.$$

Представление, аналогичное (10), справедливо и для случайной величины (сл.в.) $\int_{A_3} |\xi(t)|^2\varphi_0(t)\,dt$. Далее, в силу сходимости конечномерных распределений сл.пр. $\xi_n(t)$ к конечномерным распределениям сл.пр. $\xi(t)$ имеем, что

$$\sum_{k=1}^{N} |\xi_n(t_k)|^2 \int_{\Delta_k} \varphi_0(t)\,dt \underset{n\to\infty}{\xrightarrow{D}} \sum_{k=1}^{N} |\xi(t_k)|^2 \int_{\Delta_k} \varphi_0(t)\,dt. \tag{11}$$

Объединяя теперь (8)–(11), получаем

$$Q = \int_{R^d} |\xi(t)|^2 \varphi_0(t)\, dt = D - \lim_{n\to\infty} \int_{R^d} |\xi_n(t)|^2 \varphi_0(t)\, dt,$$

$$Q = D - \lim_{\substack{\varepsilon\to 0\\ \tau\to 0}} Q_N, \quad Q_N = \sum_{k=1}^{N} |\xi(t_k)|^2 \int_{\Delta_k} \varphi_0(t)\, dt \tag{12}$$

для некоторого разбиения Δ_k множества A_3 с $\tau \underset{\varepsilon\to 0}{\longrightarrow} 0$. Поэтому, в частности, сл.в. Q есть слабый предел квадратичных форм Q_N от центрированных гауссовских сл.в. и, следовательно, мы можем применить неравенство из [15]:

$$P\left\{\frac{Q}{EQ} > \left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2\right\} \le \alpha, \quad \forall \alpha \le 0.21515...$$

Определим тестовую статистику равенством

$$T_n = \frac{\sum_{i,j=1} |X_i + X_j| - \sum_{i,j=1} |X_i - X_j|}{2\sum_{i=1} |X_i|} = \frac{nJ(f_n, \varphi_0)}{G(f_n)}.$$

Выше было показано, что при любой альтернативе

$$\frac{T_n}{n} \underset{n\to\infty}{\xrightarrow{P}} const > 0,$$

а при нулевой гипотезе H_{01}

$$T_n \underset{n\to\infty}{\xrightarrow{D}} Q,$$

где Q есть слабый предел квадратичных форм от центрированных гауссовских сл.в., $EQ = 1$.

Таким образом, справедлива следующая теорема.

Теорема 1. *Критерий, отвергающий гипотезу H_{01}, при*

$$T_n \ge \Lambda, \quad \Lambda = \left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2 \tag{13}$$

состоятелен против всех альтернатив, он имеет асимптотический уровень значимости не более чем $\alpha, \forall \alpha \le 0.21515....$

Рассмотрим теперь асимптотическое поведение мощности β_n критерия (13). Запишем

$$1 - \beta_n = P\{T_n \le \Lambda\} = P\{U_n - EU_n \ge \nu + (1-\Lambda)\frac{2}{n-1}\widetilde{X}\}, \tag{14}$$

где

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} \left(|X_i - X_j| - |X_i + X_j|\right), \widetilde{X} = \frac{1}{n}\sum_{i=1}^{n} |X_i|,$$

$$\nu = -EU_n = E\left(|X_1 + X_2| - |X_1 - X_2|\right),$$

заметим, что $\nu > 0$ при альтернативах и в случае $P\{|X_1| = 0\} > 0$ мы определяем мощность β_n как правую часть (14).

Пусть выполнено условие

C1) $\forall H > 0 : E\exp\{H|X_1|^2\} < \infty$,

тогда по экспоненциальному неравенству Чебышева

$$P\left\{\frac{2(1-\Lambda)\widetilde{X}}{n-1} > \frac{\nu}{\sqrt{n}}\right\} \le P\left\{\sum_{k=1}^{n} |X_k| \ge \frac{(n-1)\sqrt{n}\nu}{2(1+\Lambda)}\right\} \le$$

$$\le \exp\left\{-\frac{(n-1)\sqrt{n}\nu}{2(1+\Lambda)}\right\} (E\exp\{|X_1|\})^n \le e^{-Cn\sqrt{n}} \tag{15}$$

для некоторой положительной константы C. С другой стороны, при условии С1) по теореме о больших уклонениях Р. Дасгупты для невырожденных U-статистик (см. теорему 1.6.4. в [1]) для любой последовательности $\gamma_n \underset{n\to\infty}{\longrightarrow} 0$

$$\lim_{n\to\infty} \frac{1}{n} \ln P\{U_n - EU_n > \nu + \gamma_n\} = u\left(\frac{\nu}{2\sigma}\right), \tag{16}$$

где $\sigma^2 = D\psi, \psi = |X_1 + X_2| - |X_1 - X_2|$,

$$u(x) = \ln \inf_{t\geq 0}(e^{-tn} E e^{t(\nu-\psi)}).$$

Объединяя (15) и (16), получаем

Предложение 1. *Для всех альтернатив, удовлетворяющих С1)*

$$\lim_{n\to\infty} \frac{1}{n} \ln(1-\beta_n) = u\left(\frac{\nu}{2\sigma}\right).$$

Заметим, что $-2u(\frac{\nu}{2\sigma})$ есть показатель Ходжеса-Лемана критерия (13) и

$$u(\frac{\nu}{2\sigma}) = -\frac{\nu^2}{8\sigma^2}(1+o(1)),$$

когда $\nu \to 0, \sigma > const > 0$ (см. [1], § 1.2).

Замечание 1. Определим коэффициент диагональной симметричности как:

$$SYM = \frac{T_n}{n} = \frac{\sum_{i,j=1} |X_i + X_j| - \sum_{i,j=1} |X_i - X_j|}{2n \sum_{i=1} |X_i|}.$$

Ясно, что $0 \leq SYM \leq 1$ и $SYM = 1$ тогда и только тогда, когда сл.в. $X_i \equiv const \neq 0, \forall i$. Некоторые из его свойств даны в (4), теореме 1 и предложении 1. Заметим, что коэффициент SYM инвариантен к ортогональным преобразованиям данных и изменению масштаба.

Замечание 2. Если требование $E|X_1| < \infty$ для нулевой гипотезы и альтернатив не выполнено, то тогда можно заметить, что гипотезы H_{01} и H_{02} эквивалентны аналогичным гипотезам с повторной выборкой $Y_1, Y_2, \ldots, Y_n$, где $Y_k = X_k|X_k|^{-1} \arctan|X_k|$, $k = 1, 2, \ldots, n$, (если сл.в. X_1 не имеет моментов) или $Y_k = X_k|X_k|^{\gamma-1}$ (если мы знаем, что $E|X_1|^\gamma < \infty$). Заметим, что в первом случае соответствующее значение T_n не инвариантно к изменению масштаба.

Замечание 3. Наши рассуждения остаются справедливыми и для коэффициента

$$T'_n = \frac{\sum_{i,j=1} |X_i + X_j|^\gamma - \sum_{i,j=1} |X_i - X_j|^\gamma}{2\sum_{i=1} |X_i|^\gamma}, \qquad \gamma \in (0,2).$$

Выясним теперь вопрос: "которое γ лучше?" в некотором специальном случае и в некотором специальном смысле. А именно, рассмотрим локальную ($\nu \to 0$) эффективность по Ходжесу-Леману, сдвиговые альтернативы с $X_1 \overset{D}{=} N(\theta, E)$, где $\theta = EX_1$ — параметр сдвига, E — единичная матрица. Заметим, что сл.в. $|X_1 - \theta|^2$ имеет распределение с d-степенями свободы. Можно подсчитать, что в данном случае при $|\theta| \to 0$ показатель Ходжеса-Лемана есть

$$\frac{\nu^2}{4\sigma^2} = |\theta|^4 G(\gamma)(1+o(1)),$$

где

$$G(\gamma) = \frac{\nu^2}{8d^2}\left(\frac{\Gamma(\frac{d}{2})\Gamma(\frac{d}{2}+\nu)}{\Gamma^2(\frac{d+\nu}{2})} - 1\right)^{-1}.$$

Ниже мы доказываем, что $G(\gamma)$ возрастает по γ при $\gamma \in (0,1]$ так, что ответ такой: чем больше γ, тем лучше.

Лемма 1. $G'(\gamma) > 0$, $\forall \gamma \in (0,1]$, $\forall d$.

Доказательство. Это неравенство эквивалентно следующему:

$$1-\frac{\Gamma^2(n+x)}{\Gamma(n)\Gamma(n+2x)}\geq x\left(\frac{\Gamma'(n+2x)}{\Gamma(n+2x)}-\frac{\Gamma'(n+x)}{\Gamma(n+x)}\right), \tag{17}$$

где $n=d/2, x=\gamma/2$. Известна следующая формула [16]:

$$\Psi(1+z)\stackrel{def}{=}\frac{\Gamma'(1+z)}{\Gamma(1+z)}=-C_0+\sum_{m=1}^{\infty}\frac{z}{m(m+z)}, \tag{18}$$

где C_0 — абсолютная константа. Теперь (17) может быть переписано как

$$1-\exp\{\int_n^{n+x}\Psi(z)\,dz-\int_{n+x}^{n+2x}\Psi(z)\,dz\}\geq x\left(\Psi(n+2x)-\Psi(n+x)\right),$$

или в силу (18) для $v=m+n+x-1$

$$1-\prod_{m=1}^{\infty}\left(1-\frac{x^2}{v^2}\right)\geq x^2\sum_{m=1}^{\infty}\frac{1}{v(v+x)}. \tag{19}$$

Применяя здесь неравенство $\prod(1-a_i)\leq 1-\sum a_i+\sum_{i<j}a_ia_j$ для $a_i\geq 0$, мы видим, что (19) верно, если

$$\sum_{m=1}^{\infty}\frac{1}{v^2(v+x)}+\frac{x}{2}\sum_{m=1}^{\infty}\frac{1}{v^4}\geq\frac{x}{2}\left(\sum_{m=1}^{\infty}\frac{1}{v^2}\right)^2. \tag{20}$$

Далее, оценивая суммы соответствующими интегралами, мы можем записать, что

$$\sum_{m=1}^{\infty}\frac{1}{v^2(v+x)}\geq\frac{1}{(n+x)^2(n+2x)}+\int_{n+x+1}^{\infty}\frac{dz}{z^2(z+x)}\geq$$

$$\geq\frac{1}{(n+x)^2(n+2x)}+\frac{1}{2(n+x+1)^2}-\frac{x}{3(n+x+1)^3},$$

$$\sum_{m=1}^{\infty}\frac{1}{v^4}\geq\frac{1}{(n+x)^4}+\frac{1}{3(n+x+1)^3},\quad\sum_{m=1}^{\infty}\frac{1}{v^2}\leq\frac{1}{(n+x)^2}+\frac{1}{n+x}.$$

Применяя эти неравенства, мы можем доказать (20).

Лемма доказана.

Замечание. Аналогично можно показать, что $\forall\gamma\in(1,2)\ \exists d_0=d_0(\gamma):\forall d>d_0$ $G(\gamma)=\sup\{G(s)|s\leq\gamma\}$.

§ 2. Сферическая симметричность

Гипотеза о сферической симметричности распределения выборки состоит в следующем: $H_{02}: X_1\stackrel{D}{=}CX_1$ для всех ортогональных матриц C.

В одномерном случае ($d=1$) $H_{01}\equiv H_{02}$, поэтому мы предполагаем, что $d\geq 2$. Как легко видеть,

$$H_{02}\Longleftrightarrow f(t)=f(Ct),\text{ для всех } t\in R^d \text{ и ортогональных матриц } C\Longleftrightarrow$$

$$\Longleftrightarrow R(f,\varphi)\stackrel{def}{=}\frac{1}{2}\int\int_{R^d}|f(t)-f(C^Tt)|^2\varphi(t)\,dt\,d\chi(C)=0$$

для любой весовой функции $\varphi(t)$ такой, что интеграл существует и $\varphi(t)>0$ п.в., здесь $\chi(C)$ — нормализованная мера Хаара на множестве ортогональных матриц, $\int d\chi(C)=1$.

Пусть

$$f_n(t)=n^{-1}\sum_{k=1}^{n}\exp\{i(t,X_k)\},\quad f_n(t,C)=n^{-1}\sum_{k=1}^{n}\exp\{i(t,CX_k)\}$$

— выборочные характеристические функции. Рассмотрим статистику

$$R(f_n,\varphi_0)=\frac{1}{2}\int\int_{R^d}|f_n(t)-f_n(t,C)|^2\varphi_0(t)\,dt\,d\chi(C)=$$

$$=C(d,1)n^{-2}\sum_{i,j=1}^{n}\left(g(|X_i|,|X_j|)-|X_i-X_j|\right),$$

где

$$g(u,v)=\frac{w_{d-1}}{w_d}\int\limits_{-1}^{1}\sqrt{u^2-2uvx+v^2}(1-x^2)^{\frac{d-3}{2}}\,dx,\quad w_1\stackrel{def}{=}2$$

и мы использовали формулу (3) § 1 вместе с равенством: $\forall a,b\in R^d$

$$\int|a-Cb|\,d\chi(C)=\int\sqrt{|a|^2-2(a,Cb)+|b|^2}\,d\chi(C)=$$

$$=E\sqrt{|a|^2-2|a||b|\xi+|b|^2}=g(|a|,|b|),$$

где ξ — первая координата случайного вектора равномерно распределенного на единичной сфере в R^d. Заметим, что $g(u,v)\le|u|+|v|$ и

$$g(u,u)=\mu_d|u|,\qquad \mu_d=\frac{2^{d-1}\Gamma(\frac{d}{2})\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(d-\frac{1}{2})},$$

здесь $\Gamma(.)$ есть гамма-функция и мы использовали формулу: $w_d=2\pi^{d/2}/\Gamma(\frac{d}{2})$. Для $u\neq v$ функция $g(u,v)$ может быть вычислена при помощи формулы 3.665.2 из [13].

По закону больших чисел для статистик Мизеса мы имеем, также как и в § 1,

$$R(f_n,\varphi_0)\xrightarrow[n\to\infty]{P}R(f,\varphi_0). \tag{21}$$

Далее, при нулевой гипотезе

$$nER(f_n,\varphi_0)=C(d,1)Eg(|X_1|,|X_1|)=$$

$$=P-\lim_{n\to\infty}V_n,\quad V_n=C(d,1)\frac{\mu_d}{n}\sum_{k=1}^{n}|X_k|.$$

Окончательно определим тестовую статистику формулой

$$Q_n=\frac{nR(f_n,\varphi_0)}{V_n}=\frac{\sum_{i,j=1}^{n}\left(g(|X_i|,|X_j|)-|X_i-X_j|\right)}{\mu_d\sum_{k=1}^{n}|X_k|}.$$

Можно показать, как и ранее, что при $n\to\infty$ статистика Q_n слабо сходится к квадратичной форме от центрированных гауссовских сл.в., $EQ=1$. С другой стороны, в силу (21) мы имеем для любой альтернативы к гипотезе H_{02}, что: $\frac{1}{n}Q_n\xrightarrow[n\to\infty]{P}const>0$.

Таким образом, справедлива следующая теорема.

Теорема 2. *Критерий, отвергающий гипотезу H_{02}, при $Q_n\ge\Lambda$ состоятелен против всех альтернатив, он имеет асимптотический уровень значимости не более чем $\alpha,\forall\alpha\le$ 0.21515....*

§ 3. Проверка симметричности с неизвестными центром симметрии

Для неизвестного параметра $a\stackrel{def}{=}EX_1$ мы рассматриваем гипотезы:

$H^*_{01}: X_1-a\stackrel{D}{=}-(X_1-a)$,

$H^*_{02}: X_1-a\stackrel{D}{=}C(X_1-a)$ для всех ортогональных матриц C.

Идея состоит в применении методов §§ 1-2 к центрированным сл.в. $X_k - \overline{X}$, $k = 1, 2, \ldots, X_n$, где $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ — выборочное среднее. Предположим дополнительно, что

C2) $E|X_1|^2 < \infty$, матрица $R = Cov(X_1, X_1)$ невырождена.

C3) $\int_{R^d} |t|^2 f^2(t)\varphi_0(t)\, dt < \infty$.

Рассмотрим подробнее условие C3). Известно равенство (см. [17], гл. 1, § 5, стр. 63)

$$\int_{S_\rho} e^{i(s,X)}\, ds = \left(\frac{2\pi\rho}{|X|}\right)^{d/2} |X| J_{\frac{d-2}{2}}(\rho|X|),$$

где ρ — радиус сферы S_ρ с центром в нуле в R^d, $d \geq 2$, $J_m(x)$ — функция Бесселя первого рода m-го порядка. Так что для $Z = X_1 - X_2$

$$\int_{R^d} |t|^2 f^2(t)\varphi_0(t)\, dt = \int_0^\infty \left(\int_{S_y} f^2(t)\, dt\right) \frac{dy}{y^{d-1}} =$$

$$= E\int_0^\infty \left(\frac{2\pi y}{|Z|}\right)^{d/2} |Z| J_{\frac{d-2}{2}}(y|Z|)\, \frac{dy}{y^{d-1}} = C_0 E|Z|^{-1},$$

где (см. [13], форм. 6.511, 6.561.14)

$$C_0 = (2\pi)^{d/2} \int_0^\infty y^{1-d/2} J_{\frac{d-2}{2}}(y)\, dy = \frac{2\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d-1}{2})} = C(d,1)(d-1).$$

Другими словами, условие C3) эквивалентно свойству $E|X_1 - X_2|^{-1} < \infty$, которое не является неестественным в многомерном случае: $d > 1$.

Рассмотрим сначала гипотезу о диагональной симметричности. Пусть

$$\widehat{R} = \frac{1}{n}\sum_{k=1}^{n} (X_k - \overline{X})(X_k - \overline{X})^T$$

— выборочная ковариационная матрица и $\widetilde{X_i} = \widehat{R}^{-1/2}(X_i - \overline{X})$ — нормированные выборочные значения.

Обозначим через D_n тестовую статистику:

$$D_n = \frac{\sum_{i,j=1}^{n} (|\widetilde{X_i} + \widetilde{X_j}| - |\widetilde{X_i} - \widetilde{X_j}|),}{2\sum_{k=1}^{n} |\widetilde{X_k}| + 4\frac{d-1}{n-1}\sum_{i<j} |\widetilde{X_i} - \widetilde{X_j}|^{-1} - \frac{2}{n-1}\sum_{i,j=1}^{n} |\widetilde{X_i} - \widetilde{X_j}|}.$$

Можно показать, что $D_n \xrightarrow[n\to\infty]{P} \infty$ при всех альтернативах удовлетворяющих C2), C3).

Справедлива следующая теорема.

Теорема 3. *При условиях C2),C3) критерий, отвергающий гипотезу H_{01}^*, при $D_n \geq \Lambda$ состоятелен против всех альтернатив, он имеет асимптотический уровень значимости не более чем $\alpha, \forall\alpha \leq 0.21515....$*

Рассмотрим гипотезу о сферической симметричности. Обозначим

$$S_n = \frac{\sum_{i,j=1}^{n} \left[g(|X_i - \overline{X}|, |X_j - \overline{X}|) - |X_i - X_j|\right]}{\mu_d \sum_{i=1}^{n} |X_i - \overline{X}| + Tr\widehat{R}\, \frac{2(d-1)}{d(n-1)} \sum_{i<j} |X_i - X_j|^{-1} - \frac{1}{n-1}\sum_{i,j=1}^{n} |X_i - X_j|},$$

где функция g была определена ранее.

Справедлива теорема.

Теорема 4. *При условиях C2),C3) критерий, отвергающий гипотезу H_{02}^*, при $S_n \geq \Lambda$ состоятелен против всех альтернатив, он имеет асимптотический уровень значимости не более чем $\alpha, \forall\alpha \leq 0.21515....$*

§ 4. Эллиптическая симметричность

Пусть выполнено условие

C4) $E|X_1|^4 < \infty$.

Рассмотрим гипотезу об эллиптической симметричности распределения выборки или эквивалентно:

H_{03} : распределение $R^{-1/2}X_1$ сферически симметрично, $EX_1 = 0$.

Здесь $R = Cov(X_1, X_1)$. Обозначим $\widehat{R} = \frac{1}{n}\sum_{k=1}^n X_k X_k^T$, $Y_i = \widehat{R}^{-1/2}X_i$ и

$$g_n(t) = n^{-1}\sum_{k=1}^n \exp\{i(t, Y_k)\}, \quad g_n(t, C) = n^{-1}\sum_{k=1}^n \exp\{i(t, CY_k)\}.$$

Рассмотрим статистику

$$P(g_n, \varphi_0) = \frac{1}{2}\int\int_{R^d} |g_n(t) - g_n(t, C)|^2 \varphi_0(t)\, dt\, d\chi(C) =$$

$$= C(d, 1)n^{-2}\sum_{i,j=1}^n \left(g(|Y_i|, |Y_j|) - |Y_i - Y_j|\right),$$

где функция $g(u, v)$ определена в § 2. Мы анализируем статистику $P(g_n, \varphi_0)$. Также, как и ранее, сделаем только некоторые замечания.

Во-первых, для $V_k = R^{-1/2}X_k$ и $\Delta = \widehat{R}^{-1/2}R^{1/2} - E$ по формуле Тейлора и закону больших чисел

$$\sqrt{n}(g_n(t) - g_n(t, C)) \approx \frac{1}{\sqrt{n}}\sum_{k=1}^n \left[e^{i(t,V_k)} - e^{i(t,CV_k)} + i(t, \Delta V_k)e^{i(t,V_k)} -\right.$$

$$\left. - i(t, C\Delta V_k)e^{i(t,CV_k)}\right] \approx \frac{1}{\sqrt{n}}\sum_{k=1}^n \left[e^{i(t,V_k)} - e^{i(t,CV_k)}\right] +$$

$$+(t, \sqrt{n}\Delta f'(t)) - (C^T t, \sqrt{n}\Delta f'(C^T t)).$$

С другой стороны, для $\delta \stackrel{def}{=} \widehat{R} - R = R^{1/2}(\widehat{V} - E)R^{1/2}$, где $\widehat{V} = n^{-1}\sum_{k=1}^n V_k V_k^T$ мы имеем в силу C2) асимптотическое равенство $\widehat{R}^{1/2} = R^{1/2} + L(\delta) + o(\delta), \delta \to 0$ для некоторой симметричной матрицы $L(\delta)$ линейной по δ. Возводя его в квадрат, получаем $\delta = R^{1/2}L(\delta) + L(\delta)R^{1/2}$ или

$$\widehat{V} - E = R^{-1/2}L(\delta) + L(\delta)R^{-1/2}. \tag{22}$$

При справедливости H_{03} имеем $f(t) = g\left(|t|^2/2\right)$ для некоторой функции $g : R^1 \to R^1$. Далее, так как $\Delta = -\widehat{R}^{-1/2}(\widehat{R}^{1/2} - R^{1/2}) = -R^{-1/2}L(\delta) + o(\delta)$, то мы имеем в силу (22) для любых $t \in R^d$

$$(t, \sqrt{n}\Delta f'(t)) = (t, \sqrt{n}\Delta t)g'\left(|t|^2/2\right) = -\frac{\sqrt{n}}{2}(t, (\widehat{V} - E)t)g'\left(|t|^2/2\right) +$$

$$+o(\delta) = o(\delta) - \frac{1}{2\sqrt{n}}\sum_{k=1}^n (t, (V_k V_k^T - E)f'(t)).$$

Это поможет нам проанализировать главную часть разности $g_n(t) - g_n(t, C)$. Также, как и ранее, мы можем вывести при справедливости H_{03} и условий C2),C3),C4) что

$$nP(g_n, \varphi_0) \xrightarrow[n\to\infty]{D} Q,$$

где Q есть квадратичная форма от центрированных гауссовских сл.в. и для $Z = R^{-1/2}X_1$

$$EQ = \frac{1}{2}\int\int_{R^d} |\Delta_1(t) - \frac{1}{2}\Delta_2(t)|^2\varphi_0(t)\,dt\,d\chi(C),$$

где

$$\Delta_1(t) = \exp\{i(t,Z)\} - \exp\{i(t,CZ)\},$$
$$\Delta_2(t) = (t,(ZZ^T - E)f'(t)) - (C^Tt,(ZZ^T - E)f'(C^Tt)).$$

Рассуждая, как и ранее, получаем

$$E\int\int_{R^d} |\Delta_1(t)|^2\varphi_0(t)\,dt\,d\chi(C) = 2C(d,1)\mu_d E|Z| \tag{23}$$

и

$$(t,f'(t)) = |t|^2 g'\left(|t|^2/2\right), \quad \Delta_2(t) = ((t,Z)^2 - (t,CZ)^2)g'\left(|t|^2/2\right),$$
$$E(\Delta_2(t))^2 = (g')^2(2E(t,Z)^4 - 2E(t,Z)^2(t,CZ)^2).$$

Рассмотрим слагаемые в последнем выражении. В силу сферической симметричности ф.р. сл.в. $Z = (Z_1, Z_2, \ldots, Z_d)$

$$E(t,Z)^4 = |t|^4 E(|t|/t,Z)^4 = |t|^4 EZ_1^4 = |t|^4 E|Z|^4\frac{3}{d(d+2)},$$

здесь последнее равенство выполнено для всех сферически симметричных распределений.

С другой стороны,

$$\int (t,CZ)^2\,d\chi(C) = |t|^2|Z|^2 E\xi^2 = \frac{1}{d}|t|^2|Z|^2, \tag{24}$$

где ξ есть первая координата случайного вектора, равномерно распределенного на единичной сфере в R^d. Итак,

$$E\int (\Delta_2(t))^2\,d\chi(C) = 2(g')^2(|t|^4 EZ_1^4 - \frac{1}{d}|t|^2E|Z|^2(t,Z)^2) = \tag{25}$$
$$= 2(t,f'(t))^2\varrho_d E|Z|^4,$$

где $\varrho_d = 2(d-1)d^{-2}(d+2)^{-1}$. Пусть Y есть независимая копия сл.в. Z, обозначим

$$J = \int_{R^d} (t,f'(t))^2\varphi_0(t)\,dt, \quad v(x,y) = |x-y|^{-3}(|x|^2|y|^2 - (x,y)^2)$$

и предположим выполненным условие

C5) $E|X_1|^2|X_1 - X_2|^{-1} < \infty$.

Теперь нам потребуется следующая лемма.

Лемма 2. $J = C(d,1)Ev(Z,Y)$.

Доказательство. Нетрудно доказать следующую цепочку равенств:

$$J = E\int_{R^d} (t,Z)(t,Y)\exp\{i(t,Z-Y)\}\varphi_0(t)\,dt =$$
$$= -E\int_{R^d} \frac{\partial^2}{\partial\alpha\partial\beta}\bigg|_{\alpha=\beta=1} (1 - \exp\{i(t,\alpha Z - \beta Y)\})\varphi_0(t)\,dt =$$
$$= -E\frac{\partial^2}{\partial\alpha\partial\beta}\bigg|_{\alpha=\beta=1} C(d,1)|\alpha Z - \beta Y| = C(d,1)Ev(Z,Y),$$

здесь мы использовали формулу (3). Все интегралы понимаются в смысле их главного значения (на бесконечности), $Ev(Z,Y) < \infty$ в силу C5).

Лемма доказана.

Следствие. *В силу леммы 4 и (25)*

$$E\int\int_{R^d}|\Delta_2(t)|^2\varphi_0(t)\,dt\,d\chi(C)=2C(d,1)\varrho_d Ev(Z,Y)E|Z|^4. \tag{26}$$

Можно также показать, что

$$ReE\int\int_{R^d}\Delta_1(t)\overline{\Delta_2(t)}\varphi_0(t)\,dt\,d\chi(C)=$$

$$=E\int\int_{R^d}[\exp\{i(t,Z)\}-\exp\{i(t,CZ)\}]((t,Z)^2-(t,CZ)^2)) \tag{27}$$

$$g'\left(|t|^2/2\right)\varphi_0(t)\,dt\,d\chi(C)=$$

$$=2E\int_{R^d}\exp\{i(t,Z)\}((t,Z)^2-\frac{1}{d}|t|^2|Z|^2)g'\left(|t|^2/2\right)\varphi_0(t)\,dt,$$

здесь мы применили формулу (24). Действуя, как ранее, мы можем преобразовать (27):

$$E\int_{R^d}\exp\{i(t,Z)\}(t,Z)^2g'\left(|t|^2/2\right)\varphi_0(t)\,dt=$$

$$=E\int_{R^d}\exp\{i(t,Z)\}(t,Z)(Z,f'(t))\varphi_0(t)\,dt= \tag{28}$$

$$=E(Z,Y)\int_{R^d}\frac{\partial}{\partial\alpha}\Big|_{\alpha=1}(1-\exp\{i(t,\alpha Z-Y)\})\varphi_0(t)\,dt=C(d,1)v_1(Z,Y),$$

где $v_1(x,y)=(x,y)(|x|^2-(x,y))|x-y|^{-1}$. Рассуждая аналогично, получаем

$$E\int_{R^d}\exp\{i(t,Z)\}|t|^2|Z|^2g'\left(|t|^2/2\right)\varphi_0(t)\,dt=$$

$$=E|Z|^2\int_{R^d}\exp\{i(t,Z)\}(t,f'(t))\varphi_0(t)\,dt=C(d,1)v_2(Z,Y), \tag{29}$$

где $v_2(x,y)=|x|^2((x,y)-|y|^2)|x-y|^{-1}$.

Объединяя (23),(26)–(29) получаем, что при гипотезе H_{03} для $Y_i=\widehat{R}^{-1/2}X_i$

$$EQ=C(d,1)P-\lim_{n\to\infty}R_n,$$

$$R_n=R_n(Y_1,Y_2,\dots,Y_n)=\frac{\mu_d}{n}\sum_{i=1}^n|Y_i|+\frac{1}{2n(n-1)}\sum_{i<j}v(Y_i,Y_j)\frac{\varrho_d}{n}\sum_{i=1}^n|Y_i|^4-$$

$$-\frac{2}{n(n-1)}\sum_{i<j}(v_1(Y_i,Y_j)-\frac{1}{d}v_2(Y_i,Y_j)).$$

Окончательно определим тестовую статистику формулой $G_n=nP(g_n,\varphi_0)/R_n$.

Итак, верна следующая теорема.

Теорема 5. *Путь выполнены условия C2),C4) и C5). Критерий, отвергающий гипотезу H_{03}, при $G_n\geq\Lambda$ состоятелен против всех альтернатив, он имеет асимптотический уровень значимости не более чем $\alpha,\forall\alpha\leq 0.21515....$*

Рассмотрим гипотезу:

H^*_{03} : распределение $R^{-1/2}(X_1-\theta)$ сферически симметрично при некотором неизвестном $\theta\in R^d$.

Обозначим $\widehat{R}=\frac{1}{n}\sum_{k=1}^n(X_k-\overline{X})(X_k-\overline{X})^T$, $Z_i=\widehat{R}^{-1/2}(X_i-\overline{X})$,

$$G^*_n=\frac{\frac{1}{n}\sum_{i,j=1}^n\left(g(|Z_i|,|Z_j|)-|Z_i-Z_j|\right)}{R^*_n},$$

$$R_n^* = R_n(Z_1, Z_2, \dots, Z_n) + \frac{2(d-1)}{n(n-1)} \sum_{i<j} |Z_i - Z_j|^{-1} + \frac{4}{n(n-1)} \sum_{i<j} v_3(Z_i, Z_j),$$

где $v_3(x,y) = ((x,y) - |y|^2)|x-y|^{-1}$.

Теорема 6. *Пусть выполнены условия C2),C4) и C5). Критерий, отвергающий гипотезу H_{03}^*, при $G_n^* \geq \Lambda$ состоятелен против всех альтернатив, он имеет асимптотический уровень значимости не более чем $\alpha, \forall \alpha \leq 0.21515....$*

3. Проверка однородности

Пусть $X_1, X_2, \dots X_n$ и $Y_1, Y_2, \dots Y_m$, $X_i, Y_i \in R^d$ — две независимые повторные выборки $X_1, Y_1 \not\equiv 0, a.s.$. Проверка гипотезы H_0 об однородности (совпадении) распределений X-ов и Y-ов в случае $d = 1$ традиционно производится на основе критериев типа Колмогорова-Смирнова, w^2, хи-квадрат, Манна-Уитни, Вилкоксона и других. В многомерном случае возникают трудности с определением асимптотических уровней значимости их аналогов, эти аналоги не инвариантны к линейным преобразованиям исходных данных. В настоящем параграфе работы предлагаются алгоритмы проверки однородности распределений двух повторных выборок в различных постановках и изучаются их свойства.

Рассмотрим следующий коэффициент:

$$HOM = H_{m,n} = 1 - \frac{S_1 + S_2}{2S_3}, \tag{30}$$

где

$$S_1 = \frac{1}{n^2} \sum_{i,j=1}^{n} |X_i - X_j|, S_2 = \frac{1}{m^2} \sum_{i,j=1}^{m} |Y_i - Y_j|, S_3 = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} |X_i - Y_j|,$$

здесь $|.|$ — евклидова норма в R^d и в случае $S_3 = 0$ (что влечет $S_1 = S_2 = 0$) мы полагаем по определению, что $H_{m,n} = 0$. Коэффициент (30) может быть представлен через выборочные характеристические функции:

$$f_n(t) = \frac{1}{n} \sum_{k=1}^{n} \exp\{i(t, X_k)\}, \quad g_m(t) = \frac{1}{m} \sum_{k=1}^{m} \exp\{i(t, Y_k)\},$$

(где $(..)$ — скалярное произведение в R^d), а именно:

$$H_{m,n} = \frac{\int_{R^d} |f_n(t) - g_m(t)|^2 \frac{dt}{|t|^{1+d}}}{\int_{R^d} (1 - |f_n(t)|^2 + 1 - |g_m(t)|^2 + |f_n(t) - g_m(t)|^2) \frac{dt}{|t|^{1+d}}}.$$

Справедлива следующая теорема.

Теорема 7. *1) Пусть $E(|X_1| + |Y_1|)\ln^{\delta}(1 + |X_1| + |Y_1|) < \infty$, для некоторого $\delta > 1$, тогда почти, наверное*

$$H \stackrel{def}{=} \lim_{m,n \to \infty} H_{m,n} = 1 - \frac{E|X_1 - X_2| + E(|Y_1 - Y_2|}{2E|X_1 - Y_1|};$$

2) $0 \leq H \leq 1$, $H = 1 \Longleftrightarrow X_1 \equiv C_0 = const, Y_1 \equiv C_1 = const, C_0 \neq C_1$,

$H = 0 \Longleftrightarrow$ выборки X-ов и Y-ов однородны;

3) $0 \leq H_{m,n} \leq 1$, $H_{m,n} = 1 \Longleftrightarrow X_i \equiv C_0 = const, Y_j \equiv C_1 = const, C_0 \neq C_1, \forall i, j$, если $m = n$, то тогда $H_{n,n} = 0 \Longleftrightarrow$ выборки X-ов и Y-ов совпадают (без учета их порядка), если $m \neq n$, то тогда $H_{m,n} = 0 \Longleftrightarrow X_i = Y_j, \forall i, j$;

4) В условиях пункта 1) при справедливости нулевой гипотезы H_0 об однородности X-ов и Y-ов имеет место слабая сходимость распределений:

$$D - \lim_{m,n \to \infty} \frac{2mn}{m+n} H_{m,n} = Q, \qquad EQ = 1,$$

где Q — неотрицательная квадратичная форма от центрированных гауссовских случайных величин.

Доказательство. Пункт 1) следует из закона бошьших чисел для одновыборочных и двух-выборочных статистк Мизеса [16, стр. 69, 73]. Второй и третий пункты можно легко доказать, используя интегральное представление для $H_{m,n}$ через выборочные характеристические функции приведенное выше, и, пункт 4) доказывается так же как это делалось выше в §1.

Замечание 1. Для квадратичных форм Q, упомянутых в пункте 4), равномерно по X_1, Y_1 имеет место неравенство: $P\{Q \geq \Lambda\} \leq \alpha \, \forall \alpha \leq 0{,}21515\ldots$, где $\Lambda = (\Phi^{-1}(1 - \frac{\alpha}{2}))^2$, $\Phi(.)$ — ф.р. нормальной (0,1) сл.в., $\Phi^{-1}(.)$ — обратная функция. Таким образом, критерий, отвергающий гипотезу H_0 об однородности X-ов иY-ов при

$$\frac{2mn}{m+n} H_{m,n} \geq \Lambda, \tag{31}$$

состоятелен против всех альтернатив, удовлетворяющих условиям пункта 1), он имеет асимптотический уровень значимости, равный $\alpha, \forall \alpha \leq 0{,}21515\ldots$.

Замечание 2. Задачу проверки гипотезы об однородности распределений одномерных случайных величин X_k, Y_k, принимающих конечное число значений, можно свести к проверке однородности распределений выборок:

$$X'_k = \begin{pmatrix} \chi(X_k = a_1) \\ \chi(X_k = a_2) \\ \ldots \\ \chi(X_k = a_d) \end{pmatrix}, Y'_k = \begin{pmatrix} \chi(Y_k = a_1) \\ \chi(Y_k = a_2) \\ \ldots \\ \chi(Y_k = a_d) \end{pmatrix} \quad k = 1, 2, \ldots n,$$

здесь $\chi(.)$ — индикатор соответствующего множества и $\{a_i\}_{i=1}^d$ — вероятные значения случайных величин X_k, Y_k. Рассуждая, как и ранее, можно получить формулу

$$H_{m,n} = \frac{\sum_{k=1}^d (p_k - q_k)^2}{2 - 2\sum_{k=1}^d p_k q_k},$$

где $p_k, (q_k), k \geq 1$ — выборочные частоты значений a_k для выборки X-ов (Y-es). Действительно, пусть $t = (t_1, t_2, \ldots, t_d)$, тогда

$$f_n(t) = \frac{1}{n} \sum_{j=1}^n e^{i(t, X'_k)} = \sum_{k=1}^d p_k e^{it_k}, \quad g_m(t) = \frac{1}{m} \sum_{j=1}^m e^{i(t, Y'_k)} = \sum_{k=1}^d q_k e^{it_k},$$

$$|f_n(t) - g_m(t)|^2 = \sum_{k=1}^d (p_k - q_k)^2 + \sum_{k \neq l}^d (p_k - q_k)(p_l - q_l) e^{i(t_k - t_l)} =$$

$$= \sum_{k \neq l}^d (p_k - q_k)(p_l - q_l) \left(e^{i(t_k - t_l)} - 1\right),$$

следовательно, числитель $H_{m,n}$ равен

$$2S_3 - S_1 - S_2 = \frac{1}{C(d,1)} \int_{R^d} \frac{|f_n(t) - g_m(t)|^2}{|t|^{d+1}} d\, t = -r \sum_{k \neq l}^d (p_k - q_k)(p_l - q_l) =$$

$$= r \sum_{k=1}^d (p_k - q_k)^2$$

для некоторой константы r и, с другой стороны,

$$1 - f_n(t)\overline{g_m(t)} = 1 - \sum_{k,l=1}^{d} p_k q_l e^{i(t_k - t_l)} = 1 - \sum_{k=1}^{d} p_k q_k - \sum_{k \neq l}^{d} p_k q_l e^{i(t_k - t_l)} =$$

$$= \sum_{k \neq l}^{d} p_k q_l \left(1 - e^{i(t_k - t_l)}\right),$$

следовательно знаменатель $H_{m,n}$ равен

$$2S_3 = 2Re \frac{1}{C(d,1)} \int_{R^d} \frac{1 - f_n(t)\overline{g_m(t)}}{|t|^{d+1}} d\, t = 2r \sum_{k \neq l}^{d} p_k q_l = 2r \left(1 - \sum_{k=1}^{d} p_k q_k\right).$$

Мы видим, что вклад относительно малых p_k, q_k в $H_{m,n}$ будет относительно мал. В этом — одно из их отличий $\frac{2mn}{m+n} H_{m,n}$ от хорошо известного варианта статистики хи-квадрат критерия

$$\frac{2mn}{m+n} \sum_{k=1}^{d} \frac{(p_k - q_k)^2}{p_k + q_k}$$

для проверки однородности ([20], стр. 88).

Отметим также, что статистика

$$H_{m,n} = \frac{\sum_{k=1}^{\infty} (p_k - q_k)^2}{2 - 2\sum_{k=1}^{\infty} p_k q_k}$$

применима для проверки однородности, когда случайные величины X_k, Y_k изменяются на счетном множестве значений.

§ 5. Асимптотика функции мощности

Оценим теперь мощность β критерия (31) при увеличении объемов выборок. Пусть $m = n, d = 1$ и выполнено условие Г. Крамера: $\exists L > 0 : E \exp\{L(|X_1| + |Y_1|)\} < \infty$. Имеем

$$1 - \beta = P\{n(2S_3 - S_1 - S_2) \leq 2S_3\Lambda\} \leq P\{(n - \Lambda)(2S_3 - S_1 - S_2) \leq$$

$$\leq 2\Lambda \log n\} + P\{S_1 \geq \log n\} + P\{S_2 \geq \log n\}. \tag{32}$$

Оценим два последних слагаемых. Ясно, что $S_1 \leq \frac{2}{n} \sum_{i=1}^{n} |X_i|$ (аналогичное соотношение выполнено для S_2), следовательно, в силу экспоненциального неравенства для сумм независимых случайных величин, удовлетворяющих условию Г. Крамера ([18], стр. 81): $\exists n_0 : \forall n \geq n_0$

$$P\{S_1 \geq \log n\} + P\{S_2 \geq \log n\} \leq \exp\{-n\sqrt{\log n}\}.$$

Обозначим единым символом C все положительные константы, зависящие только от размерности данных d и моментов $E|X_1|, E|Y_1|$. Далее, имеем

$$C(2S_3 - S_1 - S_2) = \int_{R^d} |f_n(t) - g_n(t)|^2 \frac{dt}{|t|^{1+d}} \geq \Delta - S_n, \tag{33}$$

где

$$\Delta = \int_{R^d} |f(t) - g(t)|^2 \frac{dt}{|t|^{1+d}}, \quad S_n = \frac{1}{n} \sum_{k=1}^{n} \xi_k, \quad E\xi_k = 0,$$

$$\xi_k = 2Re \int_{R^d} (\overline{g}(t) - \overline{f}(t))(e^{i(t,X_k)} - f(t) - e^{i(t,Y_k)} + g(t)) \frac{dt}{|t|^{1+d}}$$

(черта сверх означает комплексное сопряжение). Итак, для некоторой неслучайной последовательности $\alpha_n \underset{n\to\infty}{\longrightarrow} 0$

$$1-\beta \le \exp\{-n\sqrt{\log n}\} + P\{S_n \ge \Delta(1+\alpha_n)\}.$$

Легко видеть, что

$$\int_{R^d} (1-\exp\{i(t,X)\})\frac{dt}{|t|^{1+d}} \equiv C|X|.$$

Интеграл понимается в смысле главного значения

$$\xi_k^2/\Delta + |\xi_k| \le C(|X_k| + E|X_k| + |Y_k| + E|Y_k|),$$

поэтому, в частности, для случайных величин ξ_k условие Г. Крамера выполнено, следовательно ([19], стр. 208)

$$\nu = \limsup_{n\to\infty} \frac{1}{n}\ln(1-\beta) \le -h(\Delta) = -\sup_x(\Delta x - \ln Ee^{x\xi_1}).$$

В соответствии с общей теорией больших уклонений $h(\Delta) < \infty$ для достаточно малых Δ функция $h(.)$ неотрицательна, выпукла и $h(0) = 0$ ([19], стр. 204).

Пусть теперь $\Delta \to 0$ и $\tau^* = \sup\{\tau | E\exp\{\tau|X_1| + \tau|Y_1|\}\} > 0$ равномерно для достаточно малых Δ, тогда $h(\Delta) = (1+o(1))\Delta^2/(2\sigma^2)$ (см. [19], стр. 204, 208), здесь $\sigma^2 \stackrel{def}{=} E\xi_1^2$.

Рассмотрим далее простые сдвиговые альтернативы $H_1 : g(t) = e^{i\theta t}f(t), \forall t$. Предположим, что существует плотность распределения $p(x)$, (соответствующая характеристической функции $f(t)$), принадлежащая $L_2(R^1)$, положительная, абсолютно непрерывная, обладающая ненулевой, конечной фишеровской информацией I и пусть при $\theta \to 0$

$$-\ln \int_{-\infty}^{\infty} \sqrt{p(x)p(x+\theta)}\, dx \sim \frac{\theta^2 I}{8}.$$

Из результатов работы [20](стр. 88) следует, что

$$eff = \limsup_{\theta\to 0}\left(-\frac{8\nu}{\theta^2 I}\right) \le 1,$$

здесь eff — локальная, относительная асимптотическая эффективность критерия (31) по Ходжесу-Леману. Далее, при $\theta \to 0$

$$(1+o(1))\Delta^2 = (\theta^2\int_{-\infty}^{\infty}|f(t)|^2\,dt)^2 = (2\pi\theta^2\int_{-\infty}^{\infty}p^2(x)\,dx)^2,$$

$$(1+o(1))\sigma^2 = 8\theta^2 E(\mathrm{Im}\int_{R^d} t^{-1}\overline{f}(t)(e^{itX_1}-f(t))\,dt)^2 = \frac{8}{3}\pi^2\theta^2,$$

следовательно, в рассматриваемом случае для критерия (31)

$$eff \ge 6\left(\int_{-\infty}^{\infty}p^2(x)\,dx\right)^2\left(\int_{-\infty}^{\infty}\frac{(p'(x))^2}{p(x)}\,dx\right)^{-1},$$

поэтому, например, для гауссовских распределений $eff \ge \frac{3}{2\pi} = 0{,}47\ldots$, для распределения Лапласа: $p(x) = \lambda\exp\{-2\lambda|x|\}, eff \ge 0{,}375$ и для логистического: $p(x) = e^x(1+e^x)^{-2}, eff \ge 0{,}5$. Отметим здесь, что $eff = 1$ для критериев Колмогорова-Смирнова и w^2, более того, они асимптотически оптимальны по Ходжесу-Леману (см. [20]).

§ 6. Проверка внутренней однородности выборки

Рассмотрим задачу проверки гипотезы H_{01} о внутренней однородности как таковой выборки из независимых случайных векторов $X_0, X_1, X_2, \dots X_n$, $E|X_j| < \infty$, при H_{01} распределения X_k-х совпадают. В параметрической постановке она рассматривалась, например, А.Д. Бернштейном (против альтернатив, сближающихся с нулевой гипотезой при $n \to \infty$). Обозначим $f_k(t)$ характеристические функции случайных векторов X_k. Ясно, что

$$H_{01} \Longleftrightarrow f_k(t) = f_l(t), \forall k, l, t \Longleftrightarrow$$

$$U(f) \stackrel{def}{=} \frac{\int_0^1 x(1-x) \int_{R^d} |f(t,x) - g(t,x)|^2, \frac{dt}{|t|^{1+d}}\, dx}{\int_0^1 \int_{R^d} (1 - Ref(t,x)\overline{g(t,x)}, \frac{dt}{|t|^{1+d}}\, dx} = 0,$$

где

$$f(t,x) = \frac{1}{[nx]+1} \sum_{k=0}^{[nx]} f_k(t) \quad g(t,x) = \frac{1}{n-[nx]} \sum_{k=[nx]+1}^{n} f_k(t),$$

здесь [.] — целая часть числа, таким образом, $U(f) \neq 0$ для всех альтернатив для фиксированного n.

Проверка гипотезы H_{01} может быть произведена на основе статистики $H_n = U(f_n)$, полученной подстановкой в $U(f)$ вместо $f(t,x)$ и $g(t,x)$ соответственно их выборочных значений:

$$f_n(t,x) = \frac{1}{[nx]+1} \sum_{k=0}^{[nx]} \exp\{i(t, X_k)\}, \qquad g_n(t,x) = \frac{1}{n-[nx]} \sum_{k=[nx]+1}^{n} \exp\{i(t, X_k)\}$$

и заменой в $U(f_n)$ интегралов на соответствующие интегральные суммы:

$$H_n = \frac{\frac{1}{n}\sum_{k=1}^{n} x_k(1-x_k)(2S_3(k) - S_1(k) - S_2(k))}{\frac{1}{n}\sum_{k=1}^{n} S_3(k)},$$

где

$$S_1(k) = \frac{1}{k^2} \sum_{i,j=0}^{k-1} |X_i - X_j|, \quad S_2(k) = \frac{1}{(n-k+1)^2} \sum_{i,j=k}^{n} |X_i - X_j|,$$

$$S_3(k) = \frac{1}{k(n-k+1)} \sum_{i=0}^{k-1} \sum_{j=k}^{n} |X_i - X_j|, \quad x_k = \frac{k}{n}.$$

Ясно, что $0 \leq H_n \leq 0{,}5$ с вероятностью 1 и что коэффициент H_n инвариантен к изометрическим преобразованиям и изменению масштаба данных, так что он может рассматриваться как мера внутренней однородности выборки. Можно также заметить, что $H_n = 0 \Longleftrightarrow X_1 = X_2 = \dots = X_n$, п.в. При справедливости нулевой гипотезы H_{01} можно показать (также, как и в [18], § 1, § 2), что $D - \lim_{n\to\infty} nH_n = Q^*, EQ^* = 1$, где Q^* — неотрицательная квадратичная форма от центрированных гауссовских случайных величин, так что критерий, отвергающий H_{01} при

$$nH_n \geq \Lambda, \tag{34}$$

имеет асимптотический уровень значимости не более чем $\alpha, \forall \alpha \leq 0{,}21515\dots$.

Обозначим теперь $U_0(f) = \int_0^1 x(1-x) \int_{R^d} |f(t,x) - g(t,x)|^2, \frac{dt}{|t|^{1+d}}\, dx$. Для последовательности альтернатив:

$$H_{1n} : nU_0(f) \underset{n\to\infty}{\longrightarrow} \infty, \sup_{i,j,n} E|X_i - X_j| < \infty$$

имеем $E\frac{1}{n}\sum_{k=1}^{n} S_3(k) \le C$, поэтому аналогично (32), (33) мы можем получить: $\forall K > 0$

$$1-\beta \le P\{n\int_0^1 x(1-x)\int_{R^d} |f_n(t,x)-g_n(t,x)|^2 \frac{dt}{|t|^{1+d}}\,dx \le C\Lambda K\}+$$
$$+\frac{C}{K} \le P\{nS_n^* \ge nU_0(f) - C\Lambda K\} + \frac{C}{K},$$

где β — мощность критерия (34),

$$S_n^* = 2Re\int_0^1 x(1-x)\int_{R^d} \xi_n(t)(\overline{g}(t)-\overline{f}(t))\frac{dt}{|t|^{1+d}}\,dx$$
$$\xi_n(t) = f_n(t,x) - f(t,x) - g_n(t,x) + g(t,x).$$

По неравенству Коши-Буняковского

$$E(S_n^*)^2 \le 4U_0(f)\int_0^1 x(1-x)\int_{R^d} E|\xi_n(t)|^2\frac{dt}{|t|^{1+d}}\,dx \le CU_0(f)/n,$$

поэтому $\beta \underset{n\to\infty}{\longrightarrow} 1$ и, значит, критерий (34) состоятелен против последовательности альтернатив H_{1n}.

Пример. Рассмотрим проверку нулевой гипотезы $H_0 : X_k = \xi_k,\ k \ge 1$ где $\xi_k,\ k \ge 1$ повторная выборка против альтернативы $H_1 : X_k = \xi_k + F(\frac{k}{n}), k \ge 1$, где $F(x)$ — функция, локально интегрируемая по Риману, и $F(x)$ не есть константа почти всюду по отношению к мере Лебега. В этом случае

$$U_0(f) \underset{n\to\infty}{\longrightarrow}$$
$$\underset{n\to\infty}{\longrightarrow} \int_0^1 x(1-x)\int_{R^d}\left|\frac{1}{x}\int_0^x e^{i(t,F(s))}d\,s - \frac{1}{1-x}\int_x^1 e^{i(t,F(s))}d\,s\right|^2 \frac{|f(t)|^2 d\,t}{|t|^{1+d}}\,dx,$$

что не равно нулю, так как иначе мы имели бы для всех достаточно малых $|t|$, что интегранд равен нулю или после элементарных преобразований

$$\int_0^x e^{i(t,F(s))}d\,s = x\int_0^1 e^{i(t,F(s))}d\,s$$

почти всюду по отношению к мере Лебега. Производная по x интеграла в левой части здесь существует и равна интегранду или $e^{i(t,F(x))} = C(t)$, что невозможно, так как $F(x)$ не есть константа.

Таким образом, рассматриваемый критерий состоятелен против всех сдвиговых альтернатив с ф.р. F. К примеру, мы можем тестировать линейный тренд данных.

§ 7. Проверка некоторых линейных гипотез

1. Однородность к сдвигам. Рассмотрим гипотезу $H_0 : F_Y(x) = F_X(x-\theta), \forall x \in R^d$, для некоторого $\theta \in R^d$ другими словами, распределения X-ов и Y-ов принадлежат одному сдвиговому семейству распределений или эквивалентны $X - EX \overset{D}{=} Y - EY$. Пусть имеются две независимые повторные выборки $X_1, X_2, \ldots, X_n$ и $Y_1, Y_2, \ldots, Y_n$ равных объемов. В соответствии с используемым нами методом рассмотрим эмпирические характеристические функции

$$f_n(t) = \frac{1}{n}\sum_{k=1}^{n}\exp\{i(t, X_k - \overline{X})\}, \quad g_n(t) = \frac{1}{n}\sum_{k=1}^{n}\exp\{i(t, Y_k - \overline{Y})\}.$$

Пусть выполнены условия:
C) $E|X_1|^2 + E|Y_1|^2 < \infty$, матрицы $Cov(X_1, X_1)$ и $Cov(Y_1, Y_1)$ невырождены, обозначим $Z_1 = X_1 - EX_1 - Y_1 + EY_1$, в случае $d \ne 1$ $E|R^{-1/2}(Z_1)|^{-1} < \infty$ и если $d = 1$, то тогда сл.в. Z_1 имеет непрерывную плотность распределения $p(x)$ такую, что $p(0) \ne 0$.

Обозначим $f(t)$ х.ф. сл.в. $X_1 - EX_1$ и рассмотрим случайный процесс

$$\xi_n(t) = \sqrt{n}\,(f_n(t) - g_n(t))\,.$$

Нетрудно подсчитать, что при справедливости H_0

$$E|\xi_n(t)|^2 = 2(1-|f(t)|^2) + 2n\left(|f(t)|^2 - \left|f(t-\frac{t}{n})f^{n-1}(-\frac{t}{n})\right|^2\right) \underset{n\to\infty}{\longrightarrow}$$

$$\underset{n\to\infty}{\longrightarrow} 2(1-|f(t)|^2) + 2Re\left[(t,f'(t))f(-t) - f(t)(t,f'(-t))\right] + |f(t)|^2(t,Rt) = V(t),$$

где $R = Cov(X_1, X_1)$. Обозначим $X_1^* = R^{-1/2}(X_1 - EX_1), Y_1^* = R^{-1/2}(Y_1 - EY_1)$, рассуждая далее, как и в § 1 (проверка симметричности), мы можем подсчитать $\forall d \neq 1$

$$\frac{1}{C(d,1)}\int_{R^d}\frac{V(t)}{(t,Rt)^{\frac{d+1}{2}}}d\,t = 2detR^{-1/2}E|X_1^* - Y_1^*| +$$

$$+2E\frac{1}{C(d,1)}\int_{R^d}\frac{i(t,X_1-EX_1-Y_1+EY_1)e^{i(t,X_1-EX_1-Y_1+EY_1)}}{(t,Rt)^{\frac{d+1}{2}}}d\,t +$$

$$+\frac{2detR^{-1/2}}{C(d,1)}\int_{R^d}\frac{|t|^2|f(R^{-1/2}t)|^2}{|t|^{d+1}}d\,t = 2detR^{-1/2}E|X_1^* - Y_1^*|\left(1 - \frac{C^*(d,1)}{C(d,1)}\right) +$$

$$+2\frac{C_0(d,1)detR^{-1/2}}{C(d,1)}E|X_1^* - Y_1^*|^{-1} = 2(d-1)detR^{-1/2}E|X_1^* - Y_1^*|^{-1} =$$

$$= P - \lim_{n\to\infty} U_n,$$

где $C_0(d,1) = (d-1)C(d,1)$ (см. §1 (проверка симметричности)) и

$$C^*(d,1) = \int_{R^d}\frac{s_1\sin s_1}{|s|^{d+1}}d\,s = C(d,1),$$

что может быть показано также, как и при выводе формулы для $C(d,1)$:

$$U_n = 2(d-1)\frac{detR^{-1/2}}{n^2}\sum_{i,j=1}^{n}|X_i^* - Y_j^*|^{-1}, \quad X_i^* = \widehat{R}^{-1/2}(X_i - \overline{X}),$$

$$Y_j^* = \widehat{R}^{-1/2}(Y_j - \overline{Y}), \quad \widehat{R} = \frac{1}{2n}\sum_{i=1}^{n}\left\{(X_i-\overline{X})(X_i-\overline{X})^T + (Y_i-\overline{Y})(Y_i-\overline{Y})^T\right\}.$$

В случае $d = 1$ мы имеем

$$\frac{1}{C(d,1)}\int_{R^d}\frac{V(t)}{(t,Rt)^{\frac{d+1}{2}}}d\,t = \frac{2detR^{-1/2}}{C(d,1)}\int_{R^d}|f(R^{-1/2}t)|^2 d\,t = \frac{4\pi detR^{-1/2}}{C(d,1)}p(0),$$

где $p(x)$ — плотность распределения $X_1^* - Y_1^*$. Рассмотрим следующую тестовую статистику:

$$Q_n = \frac{n}{U_n}\times\frac{1}{C(d,1)}\int_{R^d}\frac{|f_n(t)-g_n(t)|^2}{(t,\widehat{R}t)^{\frac{d+1}{2}}}d\,t =$$

$$= n\times\frac{\frac{2}{n^2}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{n}|X_i^* - Y_k^*| - \frac{1}{n^2}\sum\limits_{i,j=1}^{n}|X_i^* - X_j^*| - \frac{1}{n^2}\sum\limits_{k,l=1}^{n}|Y_k^* - Y_l^*|}{\frac{2(d-1)}{n^2}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{n}|X_i^* - Y_k^*|^{-1}},$$

где в случае $d = 1$ мы заменяем знаменатель на $4\pi\widehat{p}(0)$ для некоторой состоятельной оценки плотности$p(x)$ в нулевой точке. Рассуждая, как и ранее, можно показать, что $Q_n \underset{n\to\infty}{\longrightarrow} Q$ слабо, где Q — квадратичная форма от центрированных гауссовских случайных величин,

$EQ = 1$. Соответственно, асимптотический уровень значимости будет иметь заданное значение $\alpha, \forall \alpha \le 0.21515\ldots$ и тест будет состоятельным против всех альтернатив, подчиненных условиям C).

Если условие C) не выполнено, то тогда можно рассмотреть альтернативный вариант для проверки H_0, используя статистику:

$$Q_{m,n} = \inf_{A \in R^d} Q_{m,n}(A), \quad Q_{m,n}(A) = \frac{2mn}{m+n} \times$$

$$\times \frac{\frac{2}{mn}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{m}|X_i - A - Y_k| - \frac{1}{n^2}\sum\limits_{i,j=1}^{n}|X_i - X_j| - \frac{1}{m^2}\sum\limits_{k,l=1}^{m}|Y_k - Y_l|}{\frac{1}{n^2}\sum\limits_{i,j=1}^{n}|X_i - X_j| + \frac{1}{m^2}\sum\limits_{k,l=1}^{m}|Y_k - Y_l|}.$$

Можно заметить, что минимум $Q_{m,n}(A)$ по A легко найти, перебирая значения $A = X_i - Y_j$ и $Q_{m,n}(A) \le Q_{m,n}(\theta) \xrightarrow[m,n\to\infty]{D} Q$, $EQ = 1$, где Q есть неотрицательная квадратичная форма от центрированных гауссовских случайных величин. Таким образом, мы получаем верхнюю оценку для асимптотического уровня значимости:

$$\lim_{m,n\to\infty} P\{Q_{m,n} \ge (\Phi^{-1}(1 - \frac{\alpha}{2}))^2\} \le \lim_{m,n\to\infty} P\{Q_{m,n}(\theta) \ge (\Phi^{-1}(1 - \frac{\alpha}{2}))^2\} \le$$

$$\le P\{Q \ge (\Phi^{-1}(1 - \frac{\alpha}{2}))^2\} = \alpha \,\forall\, \alpha \le 0{,}21515\ldots.$$

Соответствующий критерий будет состоятельным против всех альтернатив, подчиненных условию $E\left(|X_1| + |Y_1|\right) < \infty$. Действительно, рассмотрим

$$R_{m,n}(A) \stackrel{def}{=} C(d,1) \times$$

$$\times \left(\frac{2}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}|X_i - A - Y_k| - \frac{1}{n^2}\sum_{i,j=1}^{n}|X_i - X_j| - \frac{1}{m^2}\sum_{k,l=1}^{m}|Y_k - Y_l| \right),$$

которая выпукла и ограничена: $R_{m,n}(A) \le 2|A| + \frac{4}{n}\left(\sum\limits_{i=1}^{n}|X_i| + \sum\limits_{j=1}^{m}|Y_j|\right)$. Применяя закон больших чисел, мы получаем

$$\lim_{m,n\to\infty} \inf_{A\in R^d} R_{m,n}(A) = \inf_{A\in R^d} C(d,1)(2|X_1 - A - Y_1| - E|X_1 - X_2| -$$

$$-E|Y_1 - Y_2|) = \inf_{A\in R^d} \int_{R^d} \frac{|e^{-i(t,A)}f(t) - g(t)|^2}{|t|^{d+1}} d\,t,$$

правая часть здесь неотрицательна и равна нулю только при справедливости H_0, таким образом, для всех альтернатив $Q_{m,n} \xrightarrow[m,n\to\infty]{} \infty, \quad a.s.$ что означает состоятельность критерия (здесь $f(t), g(t)$ есть соответствующие характеристические функции).

2. Принадлежность сдвиго-масштабному семейству распределений. Пусть X, Y есть случайные вектора со значениями в R^d. Рассмотри гипотезу $H_0 : Y \stackrel{D}{=} A + BX$ для некоторых неслучайных вектора A и матрицы B. Рассмотрим сначала одномерный случай: пусть $E|X_1|^2 + E|Y_1|^2 < \infty$. Рассуждая, как и ранее, рассмотрим тестовую статистику

$$L_{m,n} = \inf_{A\in R^1, B>0} L_{m,n}(A,B), \quad L_{m,n}(A,B) = \frac{2mn}{m+n} \times$$

$$\times\frac{\frac{2}{mn}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{m}|Y_k-A-BX_i|-\frac{1}{n^2}\sum\limits_{i,j=1}^{n}B|X_i-X_j|-\frac{1}{m^2}\sum\limits_{k,l=1}^{m}|Y_k-Y_l|}{\frac{1}{n^2}\sqrt{\frac{\widehat{S}_Y}{\widehat{S}_X}}\sum\limits_{i,j=1}^{n}|X_i-X_j|+\frac{1}{m^2}\sum\limits_{k,l=1}^{m}|Y_k-Y_l|},$$

где $\widehat{S}_Y, \widehat{S}_X$ есть соответствующие выборочные дисперсии. Заметим, что $L_{m,n}(A,B)$ есть выпуклая функция от A, B, что облегчает вычисление инфинума. Можно показать, что критерий, отвергающий H_0 в случае

$$L_{m,n} \geq (\Phi^{-1}(1-\frac{\alpha}{2}))^2,$$

имеет асимптотический уровень значимости менее чем $\alpha\,\forall\,\alpha \leq 0{,}21515\ldots$ и является состоятельным против всех альтернатив, подчиненных условию $E\left(|X_1|+|Y_1|\right)<\infty$.

Рассмотрим теперь случай $d \neq 1$ и специальный случай матрицы B: $B = bB'$ для некоторой положительной константы b и ортогональной матрицы B', этот случай соответствует ситуации, когда одна из выборок подвергается сдвигам, вращениям и изменению масштаба (что соответствует различным способам регистрации данных). Обозначим $\widehat{S}_Y, \widehat{S}_X$ выборочные дисперсии:

$$\widehat{S}_Y=\frac{1}{m}\sum_{k=1}^{m}(Y_k-\overline{Y})(Y_k-\overline{Y})^T,\quad \widehat{S}_X=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{Y})(X_i-\overline{Y})^T$$

и

$$M_{m,n}=\inf_{A\in R^d,b>0,B'}M_{m,n}(A,B),\quad M_{m,n}(A,B)=\frac{2mn}{m+n}\times$$

$$\times\frac{\frac{2}{mn}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{m}|Y_k-A-bB'X_i|-\frac{1}{n^2}\sum\limits_{i,j=1}^{n}b|X_i-X_j|-\frac{1}{m^2}\sum\limits_{k,l=1}^{m}|Y_k-Y_l|}{\frac{1}{n^2}\sqrt{\frac{Tr\,\widehat{S}_Y}{Tr\,\widehat{S}_X}}\sum\limits_{i,j=1}^{n}|X_i-X_j|+\frac{1}{m^2}\sum\limits_{k,l=1}^{m}|Y_k-Y_l|}.$$

Можно показать, что критерий, отвергающий H_0 в случае

$$M_{m,n} \geq (\Phi^{-1}(1-\frac{\alpha}{2}))^2,$$

имеет асимптотический уровень значимости менее чем $\alpha\,\forall\,\alpha \leq 0{,}21515\ldots$ и является состоятельным против всех альтернатив, подчиненных условию $E\left(|X_1|+|Y_1|\right)<\infty$.

СПИСОК ЛИТЕРАТУРЫ

1. Никитин Я.Ю. *Асимптотическая эффективность непараметрических критериев.* М.: Наука. 1995. 238 с.
2. M.L. Puri, P.K. Sen *On the theory of rank order tests for location in the multivariate one-sample problem* // Ann. Math. Statist. V. 38. 1967. P. 1216–1228.
3. M. Huskova *Asymptotic distribution of rank statistics used for multivariate symmetry* // J. Multiv.Analysis. V. 1. No. 1. 1971. P. 461–484.
4. J. Mottonen, T.P. Hettsmansperger, H. Oja, J. Tienari *On the efficiency of affine invariant multivariate rank test* // J. Multiv. Analysis. V. 66. 1998. P. 118–132.
5. K-T. Fang, L-X. Zhu, P.M. Bentler *A nessesary test of goodness of fit for symmetry* // J. Multiv. Analysis. V. 45. 1993. P. 34–55.
6. V.I. Koltchinskii, L. Li *Testing for spherical symmetry of a multivariate distributions* // J. Multiv. Analysis. V. 65. 1998. P. 218–244.
7. J.C. Lee, T.C. Chang, P.R. Krishnaiah *Approximation of the distribution s of the likelihood ratio statistics for testing certain structures of the covariance matrix of real multivariate normal populations.* // APL TR 75–167, Aerospace Research Laboratory, Wright-Patterson, Ohio.
8. S. Csorgo, C.R. Heathcote *Testing for symmetry* // Biometrika. V. 74, No. 1. 1987. P. 177–184.
9. L. Baringhaus // Ann. Statist. V. 19, No. 2. 1991. P. 899–917.

10. G. Neuhaus, L-X. Zhu *Permutation test for reflected symmetry* // J. Multiv. Analysis. V. 67. 1998. P. 129–153.
11. Bakirov N.K., Rizzo M.L. , Szekely G.J. *A multivariate nonparametric test of independence*// Journal of Multivariate Analysis. 2006. V.97. Issue 8. P.1742-1756.
12. Bakirov N.K., Rizzo M.L. , Szekely G.J. *Measuring and Testing Dependence by Correlation of Distances* // The Annals of Mathematical Statistics. 2007. V.35. No. 6. P. 2769-2794.
13. Градштейн И.С., Рыжик И.М. *Таблицы интегралов, сумм, рядов и произведений, 4-е изд.* М.: Физматгиз. 1962. 1100 с.
14. R.M. Dudley *Gaussian processes on several parameters* // Ann. Math. Statist. 1965. V. 36, No. 3. P. 771–788.
15. N.K. Bakirov, G.J. Szekely *Extremal properties for Gaussian quadratic forms"Probability theory and related fields* // Probability theory and related fields. 2003. V. 126. No. 2. P. 184–202.
16. M. Abramovitz, I. Stegun *Handbook of mathematical functions.* // National Bureau of Standards. 1964.
17. Гихман И.И., Скороход А.В. *Введение в теорию случайных процессов.* М.: Наука. 1965. 654 с.
18. Петров В.В. *Предельные теоремы для сумм независимых случайных величин.* М.: Наука. 1987. 320 с.
19. Боровков А.А. *Теория вероятностей, 2-е изд.* М.: Наука. 1986. 432 с.
20. Никитин Я.Ю. *Об асимптотической эффективности по Ходжесу-Леману непараметрических критериев согласия и однородности* // Теория вероятностей и ее применения. 1987. Т. 32, № 1. С. 82–91.

Наиль Кутлужанович Бакиров,
Институт математики с ВЦ УНЦ РАН,
ул. Чернышевского, 112,
450008, г. Уфа, Россия
E-mail: bakirovnk@rambler.ru