Материалы X международной конференции “Актуальные проблемы электронного приборостроения”
АПЭП-2010. Т.6, Новосибирск, 2010. – С.42-47.

Непараметрические критерии проверки гипотезы о значимости коэффициента парной корреляции

Анна И. Коденко, Борис Ю. Лемешко, Алексей В. Танасейчук
Новосибирский Государственный Технический Университет, Новосибирск, Россия
annakodenko@gmail.com

Аннотация **– В работе рассматриваются непараметрические критерии проверки гипотезы об отсутствии корреляции, исследуются распределения статистик критериев для различных законов, проводится сравнительный анализ мощности критериев.**

Ключевые слова **– непараметрические критерии, ранги, порядковые статистики, мощность, нормальное распределение**

I. ВВЕДЕНИЕ

ОДНИМ ИЗ ПЕРВООЧЕРЕДНЫХ ЭТАПОВ анализа многомерных данных является корреляционный анализ, в процессе которого, в том числе, проверяются гипотезы о значимости коэффициента парной корреляции.

В этих целях применяются параметрические и непараметрические критерии. Параметрические критерии построены в предположении о нормальности многомерного распределения, для них известны предельные распределения, получены процентные точки [1], исследованы рамки применимости критерия [2]. С другой стороны, непараметрические критерии обладают повышенной устойчивостью к отклонениям закона распределения наблюдений от нормального.

В данной работе проводится сравнительный анализ непараметрических критериев проверки гипотезы о значимости корреляции между двумя величинами. Исследуемые критерии рассматриваются с точки зрения их эффективности по сравнению с классическим критерием проверки значимости коэффициента парной корреляции Пирсона (*t*-критерий) [2]. В данном случае под эффективностью будем понимать относительную мощность критериев по сравнению с мощностью классического критерия при соблюдении предположения о нормальности. Кроме того вызывает интерес степень устойчивости распределений статистик непараметрических критериев к существенным отклонениям наблюдаемого многомерного закона от нормального.

Непараметрические критерии можно разделить на две группы: критерии, использующие порядковые статистики, и ранговые. Среди рассматриваемых в данной работе критериев к первой группе относятся квадрантный критерий [3], приближенный критерий Шахани [4] и сериальный критерий Шведа-Эйзенхарта [5]. К ранговым критериям относятся критерий Гёфдинга [6], критерий Ширахате [7] и критерий Фишера-Йейтса [8].

II. ПОСТАНОВКА ЗАДАЧИ

A. *Проверяемая гипотеза*

Пусть даны две выборки случайных величин x, y характеризующиеся некоторой совместной функцией распределения. Проверяемая гипотеза имеет вид

$$H_0 : r_{xy} = 0 . \qquad (1)$$

При описании статистик критериев используются следующие обозначения: $\tilde{x}$, $\tilde{y}$ – выборочные медианы для x и y соответственно; R_i^x – ранг i - го наблюдения (порядковый номер в упорядоченной по возрастанию выборке x), R_i^y - аналогично для y; $x_{(i)}$, $y_{(i)}$ – i-е порядковые статистики, полученные по выборкам x и y.

B. *Методика моделирования*

При исследовании распределений статистик критериев использовалась методика статистического моделирования [9].

Моделирование псевдослучайных нормально распределенных векторов проводилось с помощью хорошо зарекомендовавшего себя преобразования [2]: пусть мы имеем совокупность случайных величин $\{Z_i\}$, $i = 1,...,m$, где Z_i – подчиняется стандартному нормальному закону с параметрами $(0;1)$. Тогда вектор $\overline{X}$, распределенный по многомерному нормальному закону с вектором математического ожидания $\overline{M}$ и ковариационной матрицей Σ, получается в результате линейного преобразования вида

$$\overline{X} = A\overline{Z} + \overline{M} . \qquad (2)$$

Где A – нижняя треугольная матрица, коэффициенты a_{ij} которой определяются рекуррентной процедурой (3).

Процедуру моделирования многомерных величин, распределенных по законам, отличным от нормального, с заданными математическим ожиданием и ковариационной матри-

цей предложено [9] реализовать в соответствии с описанным выше алгоритмом. При этом совокупность $\{Z_i\}$, $i=1,...,m$, формируется уже не по стандартному нормальному закону, а в соответствии с некоторым одномерным законом распределения с нулевым математическим ожиданием и единичной дисперсией. Затем заданная матрица раскладывается по формуле (3) и осуществляется преобразование (2). На выходе мы имеем некоторый многомерный закон, отличный от нормального закона, с известным математическим ожиданием, но, вообще говоря, с неизвестной ковариационной матрицей, так как ковариационная матрица смоделированного закона не совпадает с используемой при моделировании матрицей Σ.

$$a_{ij}=\frac{\sigma_{ij}-\sum_{k=1}^{j-1}a_{ik}a_{jk}}{\sqrt{\sigma_{jj}-\sum_{k=1}^{j-1}a_{jk}^2}},\ 1\le j\le i\le m. \qquad (3)$$

Для моделирования различных совокупностей $\{Z_i\}$, $i=1,...,m$, распределенных по законам, отличающимся от нормальных, удобно использовать двустороннее экспоненциальное распределение (ДЭ($\theta_0;\theta_1;\lambda$)) с плотностью

$$f(x;\theta_0,\theta_1,\lambda)=\frac{\lambda}{2\sqrt{2}\theta_1\Gamma\left(1/\lambda\right)}\exp\left(-\left(\frac{|x-\theta_0|}{\sqrt{2}\theta_1}\right)^{\lambda}\right), \qquad (4)$$

где λ – параметр формы, так как оно охватывает целый класс симметричных распределений. Частными случаями данного закона являются распределение Лапласа (при $\lambda=1$), нормальное ($\lambda=2$), предельными – распределение Коши ($\lambda\to 0$) и равномерное ($\lambda\to\infty$). Рис. 1 иллюстрирует изменение функции плотности двустороннего экспоненциального распределения при изменении параметра формы от 0,5 до 10. Меняя параметр формы λ, мы можем задавать непрерывное «удаление» моделируемого (наблюдаемого) многомерного закона от нормального, делая его более плосковершинным по сравнению с нормальным при $\lambda>2$ или более островершинным при $0<\lambda<2$. При $\lambda=2$ будут формироваться псевдослучайные векторы в соответствии с нормальным законом

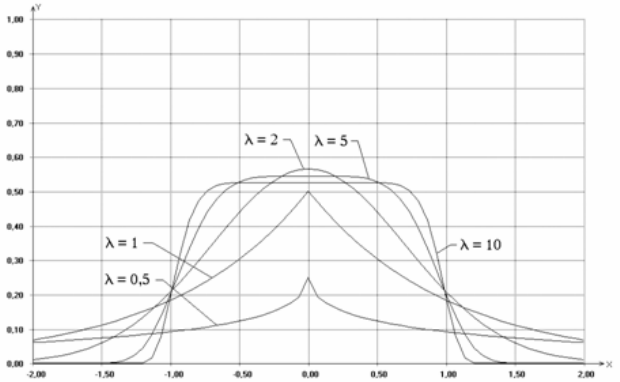

Рис. 1. Функции плотности семейства (4).

III. ИССЛЕДУЕМЫЕ КРИТЕРИИ

A. *Классический критерий*

В классическом корреляционном анализе при проверке гипотез о коэффициенте парной корреляции используют оценку коэффициента парной корреляции [2]:

$$\hat{r}_{ij}=\frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}, \qquad (5)$$

где $\hat{\sigma}_{ij}$ - элементы оценки ковариационной матрицы $\hat{\Sigma}$. Тогда статистика для проверки гипотезы (1) имеет вид [2]

$$t=\frac{\sqrt{n-2}\hat{r}_{ij}}{\sqrt{1-\hat{r}_{ij}^2}}, \qquad (6)$$

которая при справедливости гипотезы H_0 подчиняется распределению Стьюдента с $n-2$ степенями свободы.

B. *Критерии, использующие порядковые статистики*

Квадрантный критерий [3]. Критерий называется квадрантным, так как статистика критерия S основана на числе наблюдений в квадрантах, на которые делится плоскость $x0y$ прямыми $x=\tilde{x}$ и $y=\tilde{y}$. Статистика критерия имеет вид

$$S=\sum_{i=1}^{n}s(x_i,y_i), \qquad (7)$$

где $s(x_i,y_i)$ - «вес» i-го наблюдения, который вычисляется следующим образом:

$$s(x_i,y_i)=\begin{cases}1, & \text{если } x_i>\tilde{x},\ y_i>\tilde{y};\\ \frac{1}{2}, & \text{если } x_i=\tilde{x},\ y_i>\tilde{y};\\ \frac{1}{2}, & \text{если } x_i>\tilde{x},\ y_i=\tilde{y};\\ \frac{1}{4}, & \text{если } x_i=\tilde{x},\ y_i=\tilde{y};\\ 0, & \text{в остальных случаях.}\end{cases}$$

В случае нечетного количества наблюдений медиана исключается из рассмотрения. При четном n, очевидно, значения функции $s(x_i,y_i)$ принимают значения 0 или 1. Гипотеза H_0 не отклоняется если $S_{\alpha/2}<S<S_{1-\alpha/2}$.

Распределение статистики является дискретным и существенно зависит от n. Таблицы процентных точек приведены, например, в [1]. В пределе статистика распределена по нормальному закону с параметрами [3]:

$$E(S)=\frac{n}{4},\quad D(S)=\begin{cases}\frac{n^2}{16\ n-1}, & n=2k,\\ \frac{n-1}{16}, & n=2k-1.\end{cases}$$

Однако дискретностью статистики пренебрегать нельзя (см. рис.2).

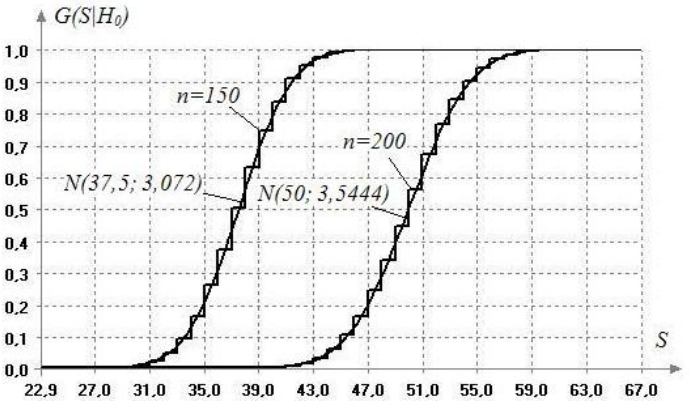

Рис. 2. Условные распределения $G(S_q \mid H_0)$ статистики квадрантного критерия в случае двумерного нормального закона их аппроксимации в случае $n > 100$.

Приближенный критерий Шахани [4]. Статистика критерия вычисляется следующим образом:

$$S = \frac{|a_1 + a_3 - a_2 - a_4|}{\sqrt{\sum_{i=1}^{4} a_i}}, \qquad (8)$$

где

$$a_k = \sum_{i=1}^{n} \alpha_k \ \ x_i, y_i \ , \ \ k = 1,2,3,4;$$

$$\alpha_1 \ \ x, y \ = \begin{cases} 1, \ \textit{если} \ x > x_{(0.7n)} \ \ u \ \ y > y_{(0.7n)}, \\ 0 \ \ \textit{иначе}; \end{cases}$$

$$\alpha_2 \ \ x, y \ = \begin{cases} 1, \ \textit{если} \ x \le x_{(0.3n)} \ \ u \ \ y > y_{(0.7n)}, \\ 0 \ \ \textit{иначе}; \end{cases}$$

$$\alpha_3 \ \ x, y \ = \begin{cases} 1, \ \textit{если} \ x \le x_{(0.3n)} \ \ u \ \ y \le y_{(0.3n)}, \\ 0 \ \ \textit{иначе}; \end{cases}$$

$$\alpha_4 \ \ x, y \ = \begin{cases} 1, \ \textit{если} \ x > x_{(0.3n)} \ \ u \ \ y \le y_{(0.3n)}, \\ 0 \ \ \textit{иначе}. \end{cases}$$

Здесь $x_{(0,3n)}$, $x_{(0,7n)}$ – $0,3n$-е и $0,7n$-е порядковые статистики соответственно. Распределение статистики является дискретным, но аппроксимируется полунормальным законом с параметрами $(0;1)$. Однако и в данном случае дискретностью распределения статистики пренебрегать не следует (см. рис.3). Проверяемая гипотеза H_0 отвергается, если $S > u_\alpha$, где u_α - квантиль полунормального распределения с параметрами $(0;1)$.

Сериальный критерий Шведа-Эйзенхарта [5]. Для вычисления статистики критерия упорядочим совокупность наблюдений (x_i, y_i) по x_i. Пару (x_i, y_i), $y_i > \tilde{y}$ обозначим через a, а (x_i, y_i), $y_i < \tilde{y}$ через b. При четном n, наблюдение, соответствующее $y_i = \tilde{y}$, исключается из рассмотрения. Исходная выборка преобразуется к последовательности вида $a,b,a,a,b,b,b,\ldots$.

Последовательность из элементов одного вида, ограниченная с двух сторон элементами другого вида, называется серией. Статистикой m критерия Шведа-Эйзенхарта является количество серий. Гипотеза H_0 отвергается если $m \le m_\alpha$, критические значения m_α приведены, например, в [1]. Очевидно, что распределение статистики, подобно распределению статистики квадрантного критерия, является дискретным и зависит от n.

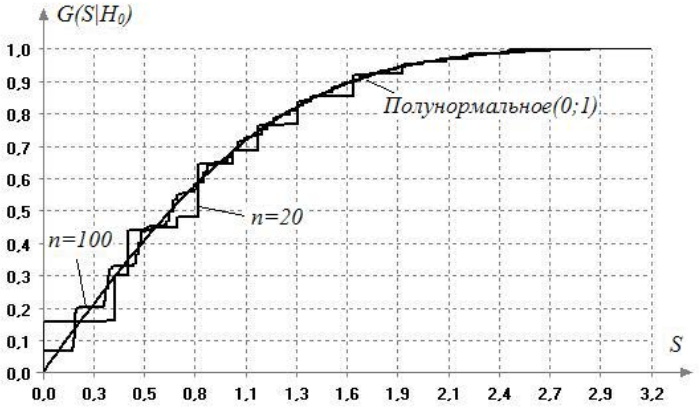

Рис. 3. Условные распределения $G(S_{Shah} \mid H_0)$ статистики квадрантного критерия в случае двумерного нормального закона и полунормальное распределение

C. Ранговые критерии

Критерий Гёфдинга [6]. Статистика критерия вычисляется следующим образом:

$$C_i = \sum_{\substack{v=1 \\ v \ne i}}^{n} \varphi(x_v, x_i)\varphi(y_v, y_i), \ i = 1, \ldots, n,$$

$$\varphi(a,b) = \begin{cases} 1, \ \textit{если} \ \ a < b; \\ 1/2, \ \textit{если} \ \ a = b; \\ 0, \ a > b. \end{cases}$$

$$Q = \sum_{i=1}^{n} (R_i^x - 1)(R_i^x - 2)(R_i^y - 1)(R_i^y - 2);$$

$$K = \sum_{i=1}^{n} C_i (R_i^x - 2)(R_i^y - 2), S = \sum_{i=1}^{n} C_i (C_i - 1);$$

$$D = \frac{Q - 2(n-2)K + (n-2)(n-3)S}{n(n-1)(n-2)(n-3)(n-4)}. \qquad (9)$$

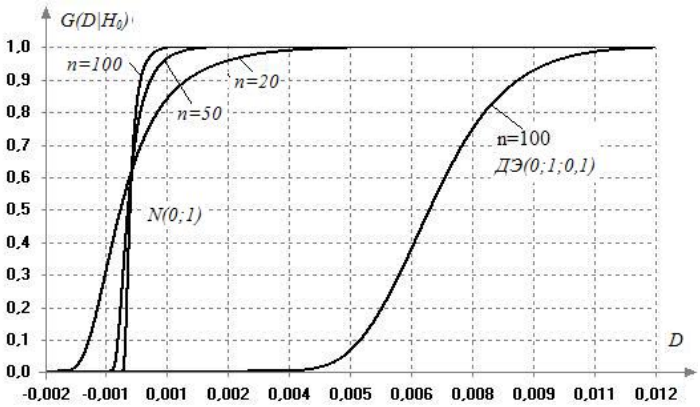

Рис. 4. Условные распределения $G(D \mid H_0)$ статистики критерия Гёфдинга как при выполнении предположения о нормальности ($N(0;1)$) для различных n, как и при ее нарушении (ДЭ(0;1;0,1)) для $n = 100$.

Гипотеза H_0 отвергается, если $D \geq D_\alpha$. Критические значения D_α приведены, например, в [1]. Распределение статистики является дискретным, однако дискретностью распределения можно пренебречь уже при достаточно малых n, распределение статистики зависит от объема рассматриваемой выборки.

Критерий Ширахате [7]. Статистика критерия имеет вид:

$$S = \sum_{i=1}^{n} \left[\left(R_i^x + \sum_{j=1}^{n} sign\ \ x_i - y_i \right) \times \right. \\ \left. \times \left(R_i^y + \sum_{j=1}^{n} sign\ \ y_i - x_i \right) \right], \quad (10)$$

где

$$sign(x) = \begin{cases} 1, & если\ \ x \geq 0, \\ 0, & если\ \ x < 0. \end{cases} \quad (11)$$

H_0 не отвергается если $S_{\alpha/2} < S < S_{1-\alpha/2}$. Таблица процентных точек приведена в [1]. Распределение статистики критерия очень сильно зависит от объема выборок n.

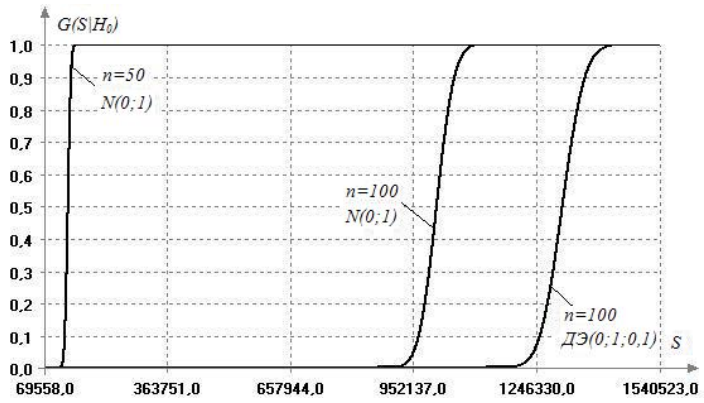

Рис. 5. Условные распределения $G(S \mid H_0)$ статистики критерия Ширахате, как при выполнении предположении о нормальности ($N(0;1)$) для различных *n*, как и при ее нарушении (ДЭ(0;1;0,1)) для $n = 100$.

Критерий корреляции Фишера-Йэйтса [8]. Статистика имеет вид:

$$\rho^* = \frac{\sum_{i=1}^{n} a_n(R_i^x) a_n(R_i^y)}{\sum_{i=1}^{n} a_n^2(i)}, \quad (12)$$

где $a_n(i)$ - математическое ожидание i-й порядковой статистики в выборке объема n из стандартного нормального распределения. Распределение статистики сходится к стандартному нормальному закону. Проверяемая гипотеза H_0 отвергается, если $|\rho^*| > u_{1-\alpha/2}$, где $u_{1-\alpha/2}$ – квантиль стандартного нормального распределения.

Распределение статистики уже при достаточно малых значениях n является непрерывным и сходится к предельному.

Распределения данной статистики устойчиво к отклонениям многомерной выборки от нормального закона. Заметные отклонения распределений статистики от стандартного нормального закона отмечается лишь в случае тяжелых хвостов наблюдаемого многомерного закона (см. рис. 6).

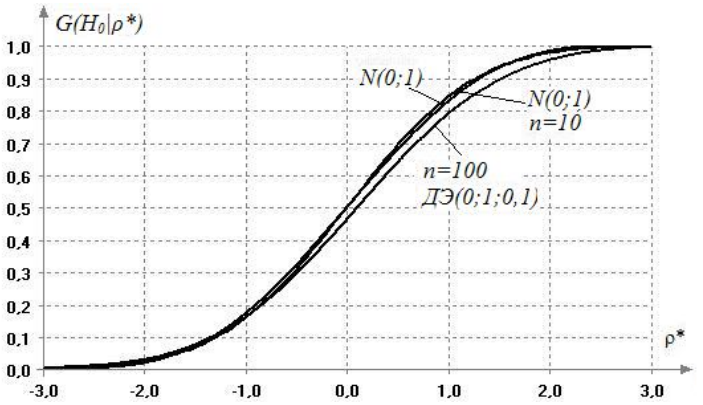

Рис. 6. Условные распределения $G(\rho^* \mid H_0)$ статистики критерия Фишера-Йейтса как при выполнении предположении о нормальности (N(0;1)) для *n*=10, как и при ее нарушении (ДЭ(0;1;0,1)) для $n = 100$, а так же предельное распределение $G(\rho^* \mid H_0)$.

IV. РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

A. Исследование распределений статистик в случае нарушения предположений о нормальности

Исследование распределений статистик критериев проводилось как в случае принадлежности наблюдений многомерному нормальному закону, так и при существенных отклонения о него. В последнем случае в процедуре моделирования многомерных величин использовалось распределение (4). Рассматривались ситуации с очень тяжелыми хвостами при использовании в процедуре генерации псевдослучайных векторов (ДЭ(0; 1; 0,1)), а также случай более плосковершинных по сравнению с нормальным законов при использовании (ДЭ(0; 1; 4)), $n = 100$.

Результаты моделирования показали, что критерии, в которых оценивание коэффициента корреляции осуществляется при помощи порядковых статистик, устойчивы к отклонению наблюдаемого закона от нормального (даже при очень тяжелых хвостах).

Напомним, что для классический критерий со статистикой (5) также устойчив к отклонениям наблюдаемого закона от нормального [2].

Ранговые критерии демонстрируют чувствительность к тяжелым хвостам (см. рис. 4, 5, 6). Распределения статистик ранговых критериев в случае плосковершинного закона на рисунках не приведены, так как полученные распределения статистик *практически совпадают* со случаем для нормального закона.

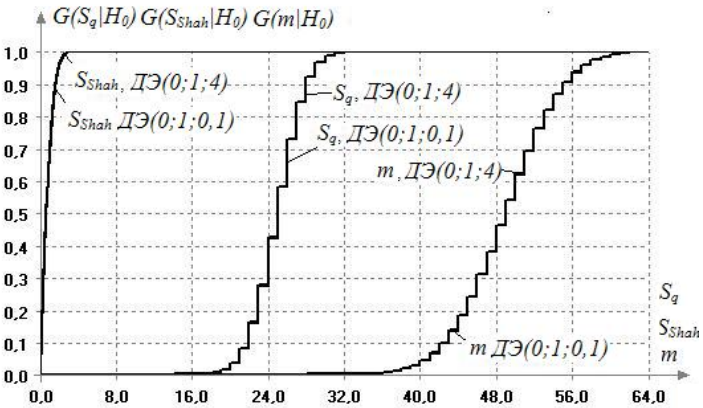

Рис. 7. Условные распределения статистик критериев, использующих порядковые статистики, в случае нарушения предположения о нормальности

B. Сравнительный анализ мощности критериев при выполнении предположений о нормальности

Исследование мощности критериев проводилось для конкурирующих гипотез $H_1: r = 0{,}1$ и $H_2: r = 0{,}5$. Ранговые критерии лишь немногим уступают по мощности классическому критерию (см. рис. 8). Исключение касается критерия Гёфдинга, который заметно уступает остальным.

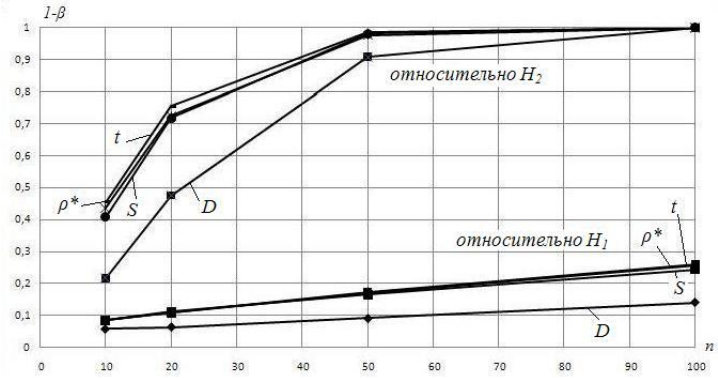

Рис. 8. Мощность ранговых критериев относительно конкурирующих гипотез H_1, H_2, в зависимости от объема выборки n при $\alpha = 0{,}05$ двумерного нормального закона.

А критерии, основанные на порядковых статистиках, по мощности относительно близких конкурирующих гипотез существенно уступает классическому t-критерию (см. рис. 9).

Наихудшие результаты показал сериальный критерий Шведа-Эйзенхарта.

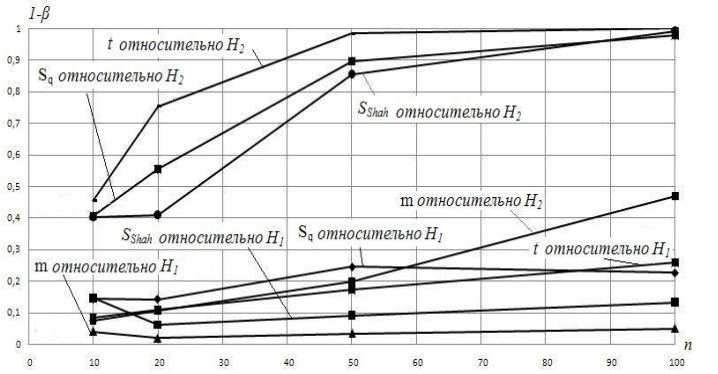

Рис. 9. Мощность критериев, использующих порядковые статистики, относительно конкурирующих гипотез H_1, H_2, в зависимости от объема выборки n при $\alpha = 0{,}05$ двумерного нормального закона.

V. ВЫВОДЫ И ЗАКЛЮЧЕНИЕ

В работе рассмотрен ряд ранговых критериев и критериев, основанных на порядковых статистиках, предназначенных для проверки гипотезы о значимости корреляции между случайными величинами. Методами статистического моделирования исследованы распределения статистик критериев и проведен сравнительный анализ мощностей критериев.

Результаты показали, что ранговые критерии лишь немногим, но уступают по мощности классическому t-критерию. Это естественно, так как t-критерий показывает высокую устойчивость к отклонениям наблюдаемого закона от многомерного нормального [2]. А в таких ситуациях, как правило, параметрические критерии лишь немногим превосходят непараметрические. Критерии же, основанные на порядковых статистиках, заметно уступают по мощности классическому критерию. Наихудший результат по мощности показали критерий Шведа-Эйзенхарта и критерий Гёфдинга.

Настоящие исследования выполнены при поддержке Российского фонда фундаментальных исследований (проект № 09-01-00056-а) и ФЦП "Развитие научного потенциала высшей школы" на 2009-2013 годы.

СПИСОК ЛИТЕРАТУРЫ

[1] Кобзарь А.И. Прикладная математическая статистика. – М.: ФИЗМАТЛИТ, 2006. – 816 с.

[2] Помадин С.С. Исследование распределений статистик многомерного анализа данных при нарушении предположений о нормальности, диссертация на соискание ученой степени кандидата технических наук. – Новосибирск: НГТУ, 2004 – 136с.

[3] Гаек Я., Шидак З. Теория ранговых критериев / Пер. с англ. – М.: Наука, 1971.

[4] Shahani, A.K. A simple graphical test of association for large samples. // Applied Statistics 1969 . 18 P. 185-190.

[5] F. S. Swed , C. Eisenhart, Tables for Testing Randomness of Grouping in a Sequence of Alternatives. // Ann. Math. Stat. 1943. V. 14, № 1, 66-87.

[6] Hoeffding W. A non-parametric test of independence. // AMS. 1961 V. 19. P. 546-557.

[7] Shirahate S. Intraclass test of independence. // Biometrika. 1981. V. 68, № 2 P. 451-456.

[8] Fieller E. C., Pearson E. S. Tests for rank correlation coefficients. 2 // Biometrika. 1961. V. 48, №1-2 P. 29-40.

[9] Лемешко Б.Ю., Постовалов С.Н., Чимитова Е.В., Помадин С.С., Французов А.В. Компьютерные методы исследований статистических закономерностей // Тезисы докладов всероссийской НТК «Информационные системы и технологии ИСТ-2001». – Нижний Новгород, 2001. – с. 87-89.

Коденко Анна Игоревна, магистрант кафедры прикладной математики Новосибирского государственного технического университета.

Лемешко Борис Юрьевич, д.т.н., профессор кафедры прикладной математики Новосибирского государственного технического университета, декан факультета прикладной математики и информатики. Область научных интересов – прикладная математическая статистика, компьютерные методы анализа данных и исследования статистических закономерностей. Имеет около 300 публикаций.

Танасейчук Алексей Владимирович, аспирант кафедры прикладной математики Новосибирского государственного технического университета. Область научных интересов – компьютерные методы анализа данных и исследования статистических закономерностей. Имеет 6 публикаций.

.