МАТЕМАТИКА

УДК 519.25 DOI 10.25730/VSU.0536.22.031

Аппроксимация процентных точек критерия Гири

Заляжных Владимир Васильевич
кандидат технических наук, доцент кафедры стандартизации, метрологии и сертификации,
Северный (Арктический) федеральный университет им. М. В. Ломоносова.
Россия, г. Архангельск. ORCID: 0000-0001-6102-7221. E-mail: zalvladimir@yandex.ru

Аннотация. Критерий Гири является одним из наиболее мощных критериев для проверки гипотезы нормальности распределения. Недостаток его состоит в значительной зависимости процентных точек статистики критерия от объема выборки и величины функции распределения критерия, что осложняет оценку гипотезы нормальности как по допускаемому, так и по достигаемому уровню значимости.

В данной статье методом Монте-Карло получена таблица процентных точек статистики критерия Гири в широких диапазонах объемов выборки и значений функции распределения критерия. По ее данным предложена аппроксимация процентных точек, позволяющая находить их значения с высокой точностью. Это дает возможность без затруднений применять критерий Гири. Показано также, что предлагавшиеся ранее аппроксимации процентных точек неудовлетворительны по точности.

Ключевые слова: критерий Гири, аппроксимация, процентные точки, уровень значимости.

Введение. Для статистического анализа числовых данных весьма желательно описать их распределение какой-либо математической моделью, достаточно хорошо описывающей данные. При этом часто наиболее подходящей моделью является нормальное распределение. Проверку гипотезы о нормальности распределения проводят по тому или иному статистическому критерию. Выбор критерия во многом определяется его мощностью.

По данным [8], одним из наиболее мощных критериев нормальности является критерий Гири, предложенный в [14]. К его преимуществам можно отнести также сравнительную простоту вычислений. Критерий Гири применяется в различных исследованиях [например, 1; 3; 7; 9; 11].

Статистика критерия:

$$d = \frac{1}{ns}\sum_{i=1}^{n}|x_i - \bar{x}|, \qquad (1)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, n – объем выборки, x_i – значения выборки.

Критерий двусторонний, гипотеза нормальности не отвергается, если

$$d_{\alpha/2} \le d \le d_{1-\alpha/2}, \qquad (2)$$

где α – допускаемый (задаваемый) уровень значимости. Значения $\alpha/2$ и $1\text{-}\alpha/2$ равны соответствующим значениям F функции распределения критерия Гири.

Часто при проверке гипотез статистическими критериями рассчитывают достигаемый уровень значимости p (или *p-value*), который более информативен, чем проверка по допускаемому уровню значимости α. Для критерия Гири как двустороннего критерия

$$p = 2min\{F(n,d), 1 - F(n,d)\}, \qquad (3)$$

где $F(n,d)$ – значение функции распределения статистики критерия Гири.

Процентные точки $d(n,F)$ статистики (1) для некоторых n и F приведены в [2; 6; 8; 14]. Эти значения сильно зависят от n и F, что осложняет в общем случае оценку по (2), а также нахождение p по (3). Предложены различные аппроксимации $d(n,F)$ нормальным распределением с математическим ожиданием E и дисперсией D. Таким образом,

$$d(n,F) \approx E + u(F)\cdot\sqrt{D},$$

где $u(F)$ – квантиль стандартного нормального распределения. По [12]

$$E = 0,\ 7978845 + \frac{0,199471}{n} + \frac{0,024934}{n^2} - \frac{0,031168}{n^3}$$
$$D = \frac{0,04507}{n} - \frac{0,084859}{n^2} + \frac{0,006323}{n^3} \quad (4)$$

В соответствии с [13]

$$E = 0,\ 7978845608 + \frac{0,19947114}{n} + \frac{0,02493389}{n^2} - \frac{0,03116737}{n^3}$$
$$D = \frac{0,04507034}{n} + \frac{0,07957747}{n^2} - \frac{0,03978874}{n^3} \quad (5)$$

В соответствии с [6]

$$E = 0,\ 79788\frac{n-0,875}{n-1,125}$$
$$D = \frac{1}{n}(0,04507 - \frac{0,0796}{n} \quad (6)$$

Однако значения $d(n,F)$, полученные по (5), (6), (7), могут существенно отличаться от действительных значений, особенно при сравнительно небольших n, как это указано в [5; 6; 8].

Задачи исследования. В задачи исследования входило получение таблицы процентных точек $d(n,F)$ для статистики d критерия Гири при варьировании функции распределения критерия и объема выборки n в широких пределах, нахождение достаточно точной аппроксимации для полученных табличных значений $d(n,F)$, оценка точности найденной аппроксимации в сравнении с известными, нахождение значений $F(n,d)$ при любых n и d с последующим расчетом достигаемого уровня значимости.

Методика исследования. Таблицу процентных точек получали моделированием методом Монте-Карло [4] в MS Excel, реализуя макрос, содержащий циклическую структуру. Моделировали N = 2 млн. выборок из нормального стандартного распределения при каждом исследованном объеме выборки n. По каждой выборке рассчитывали статистику (1). Из полученных массивов статистик (1) находили процентные точки критерия Гири $d(n,F)$. При этом 99,73 %-й доверительный интервал для истинного значения F определяется выражением

$$F \pm 3\sqrt{F(1-F)/N}\ .$$

Некоторые доверительные интервалы:

0,005±0,00015 0,01±0,0002 0,05±0,0005 0,3±0,001 0,5±0,0011 0,7±0,001 0,95±0,0005 0,99±0,0002 0,995±0,00015.

Аппроксимацию полученных табличных значений, учитывая определенную близость распределений F к нормальным распределениям, получали исходя из функций вида $d = f(u[F])$. Расчеты проводили в MS Excel.

Аппроксимация процентных точек $d(n,F)$. Полученные методом Монте-Карло процентные точки $d(n,F)$ приведены в таблице 1.

Таблица 1

Процентные точки $d(n,F)$ для статистики (1) критерия Гири

n	$F = \alpha/2$								
	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,4	0,5
10	0,6443	0,6632	0,6913	0,7147	0,7408	0,7709	0,7919	0,8093	0,8249
15	0,6630	0,6794	0,7031	0,7229	0,7448	0,7703	0,7880	0,8025	0,8155
20	0,6773	0,6918	0,7126	0,7298	0,7490	0,7713	0,7868	0,7995	0,8111
30	0,6976	0,7094	0,7261	0,7400	0,7557	0,7739	0,7866	0,7971	0,8067
40	0,7104	0,7204	0,7347	0,7467	0,7602	0,7760	0,7869	0,7961	0,8045
50	0,7192	0,7281	0,7409	0,7516	0,7636	0,7776	0,7875	0,7957	0,8032
60	0,726	0,7341	0,7457	0,7553	0,7661	0,7789	0,7879	0,7954	0,8022
70	0,7314	0,7387	0,7493	0,7582	0,7682	0,7800	0,7883	0,7953	0,8016
80	0,7356	0,7424	0,7523	0,7605	0,7699	0,7809	0,7887	0,7952	0,8012
90	0,7392	0,7456	0,7548	0,7626	0,7713	0,7817	0,789	0,7951	0,8008
100	0,7424	0,7484	0,7570	0,7643	0,7726	0,7824	0,7893	0,7952	0,8005
150	0,7526	0,7574	0,7644	0,7702	0,7769	0,7849	0,7905	0,7953	0,7996

Окончание табл. 1

200	0,7587	0,7628	0,7688	0,7738	0,7796	0,7864	0,7913	0,7954	0,7992
300	0,7659	0,7692	0,774	0,7781	0,7828	0,7883	0,7923	0,7957	0,7988
500	0,7731	0,7757	0,7793	0,7825	0,7861	0,7904	0,7934	0,796	0,7984
700	0,7771	0,7792	0,7822	0,7849	0,7879	0,7915	0,794	0,7962	0,7983
1000	0,7805	0,7822	0,7847	0,7869	0,7894	0,7925	0,7946	0,7964	0,7982
n	*F=1-α/2*								
	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	
10	0,8398	0,8550	0,8720	0,8941	0,9111	0,9244	0,9387	0,9476	
15	0,8282	0,8414	0,8561	0,8754	0,8903	0,9026	0,9160	0,9247	
20	0,8223	0,8340	0,8471	0,8644	0,8779	0,8891	0,9015	0,9096	
30	0,8161	0,8258	0,8368	0,8516	0,8632	0,8729	0,8839	0,8910	
40	0,8126	0,8212	0,8309	0,844	0,8545	0,8632	0,8729	0,8794	
50	0,8105	0,8183	0,8271	0,8389	0,8484	0,8564	0,8653	0,8713	
60	0,8090	0,8161	0,8242	0,8352	0,8440	0,8513	0,8598	0,8654	
70	0,8079	0,8145	0,8221	0,8323	0,8404	0,8474	0,8553	0,8605	
80	0,8070	0,8132	0,8203	0,8300	0,8377	0,8443	0,8518	0,8568	
90	0,8064	0,8122	0,819	0,8281	0,8354	0,8416	0,8487	0,8534	
100	0,8058	0,8114	0,8178	0,8265	0,8335	0,8394	0,8461	0,8507	
150	0,8040	0,8086	0,8138	0,821	0,8269	0,8318	0,8375	0,8413	
200	0,8029	0,8069	0,8115	0,8178	0,8229	0,8273	0,8323	0,8357	
300	0,8019	0,8051	0,8089	0,8141	0,8183	0,8219	0,8261	0,8288	
500	0,8008	0,8034	0,8063	0,8104	0,8137	0,8165	0,8198	0,8220	
700	0,8003	0,8024	0,8049	0,8084	0,8112	0,8136	0,8164	0,8183	
1000	0,7998	0,8017	0,8037	0,8066	0,8090	0,8111	0,8134	0,8150	

При каждом n зависимость $d = f(u[F])$ с высокой достоверностью аппроксимации R^2 можно описать полиномом третьей степени, как это показано для некоторых n на рис. 1, т. е.

$$d = a(n)\cdot(u[F])^3 + b(n)\cdot(u[F])^2 + c(n)\cdot(u[F]) + z(n) \qquad (7)$$

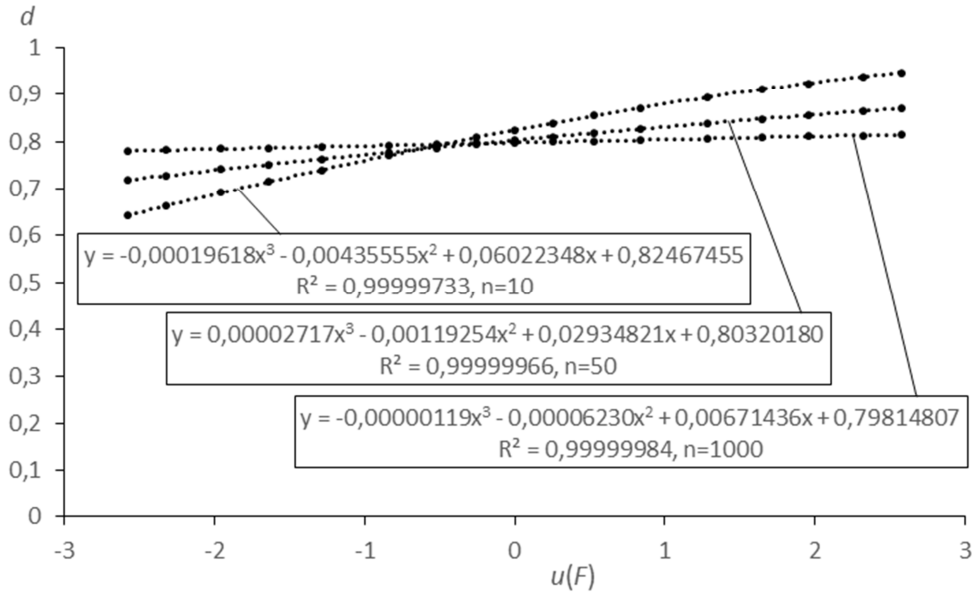

Рис. 1. Аппроксимация d = f(u[F]) полиномом третьей степени

При этом значения коэффициентов $a(n)$, $b(n)$, $c(n)$ и $z(n)$ в зависимости от n с высокой достоверностью аппроксимации могут быть описаны выражениями вида

$$k_1 n^{6r} + k_2 n^{r5r} + k_3 n^{4r} + k_4 n^{3r} + k_5 n^{2r} + k_6 n^r + k_7,$$

получаемыми как полиномы шестой степени от n^r.

Высокая достоверность аппроксимации достигается:
для $a(n)$ при r = -0,14,
для $b(n)$ при r = -0,42,
для $c(n)$ при r = -0,41,
для $z(n)$ при r = 0,05.

В результате:

$$a(n) = -4{,}01173n^{-0{,}84}+12{,}71271n^{-0{,}7}-16{,}74653n^{-0{,}56}+11{,}725466n^{-0{,}42}-4{,}59679n^{-0{,}28}+0{,}955631n^{-0{,}14}-0{,}082229;$$

$$b(n) = -10{,}0553n^{-2{,}52}+11{,}535n^{-2{,}1}-4{,}92286n^{-1{,}68}+1{,}02015n^{-1{,}26}-0{,}149316n^{-0{,}84}+0{,}0077n^{-0{,}42}-0{,}000166;$$

$$c(n) = 25{,}6685n^{-2{,}46}-33{,}3996n^{-2{,}05}+17{,}2824n^{-1{,}64}-4{,}8713n^{-1{,}23}+0{,}8653n^{-0{,}82}+0{,}0737n^{-0{,}41}+0{,}000183;$$

$$z(n) = 234{,}650624n^{0{,}3}-1834{,}91616n^{0{,}25}+5979{,}1209n^{0{,}2}-10393{,}10368n^{0{,}15}+10165{,}33465n^{0{,}1}-5305{,}2972n^{0{,}05}+1155{,}25314.$$

Подставляя выражения для $a(n)$, $b(n)$, $c(n)$ и $z(n)$ в (7), получаем выражение аппроксимации для $d(n,F)$.

Оценка точности аппроксимаций. Находили абсолютные отклонения $d(n,F)$, рассчитанных по (4), (5), (6) и (7), от $d(n,F)$, полученных методом Монте-Карло при n и F, приведенных в таблице 1. По этим отклонениям определяли средние арифметические отклонения $\Delta_{ср}$, максимальные отклонения $\Delta_{макс}$, соответствующие им табличные n_Δ, F_Δ и p_Δ, а также, по данным моделирования методом Монте-Карло, фактические $p_{апп}$.

Таблица 2

Оценка точности различных аппроксимаций

Аппроксимация	$\Delta_{макс}$	$\Delta_{ср}$	n_Δ	F_Δ	p_Δ	$p_{апп}$
По (4)	0,02473	0,00240	10	0,005	0,01	0,0245
По (5)	0,02817	0,00239	10	0,995	0,01	0,00038
По (6)	0,02963	0,00243	10	0,995	0,01	0,000228
По (7)	0,00030	0,000039	10	0,005	0,01	0,0101

Из таблицы 2 видно, что аппроксимации (4), (5) и (6) неудовлетворительны ввиду значительных различий между p_Δ и $p_{апп}$. В то же время предложенная в данной работе аппроксимация (7) обеспечивает высокую точность и вполне приемлема.

Аппроксимация функции распределения $F(n,d)$ и расчет достигаемого уровня значимости. Получить приемлемую аппроксимацию для $F(n,d)$ тем же методом, как для $d(n,F)$, не удается. Значения $F(n,d)$ могут быть получены из (7) итерацией или решением кубического уравнения вида

$$a(n)(u[F])^3+b(n)(u[F])^2+c(n)(u[F])+[z(n)-\mathrm{d}] = 0. \qquad (8)$$

Корнями кубического уравнения могут быть как действительные, так и комплексные числа [8]. В данном случае приемлемыми корнями могут быть только действительные значения $u(F)$, функция стандартного нормального распределения от которых находится в пределах 0,005…0,995.

В [8] определены достигаемые уровни значимости p при проверке гипотезы нормальности различными критериями, полученные по данным некоторых классических экспериментов: Кавендиша по определению плотности Земли, Милликена по измерению заряда электрона, Майкельсона по измерению скорости света, Ньюкомба по уточнению результатов Майкельсона. В частности, были определены методом Монте-Карло значения p по критерию Гири. В таблице 3 приведены значения p для критерия Гири по [8], а также рассчитанные по (8).

Таблица 3

Проверка нормальности по данным классических экспериментов

Эксперимент	n	d по [8]	d по (1)	p по [8]	p по (8)
Кавендиша	29	0,8008	0,8008	0,874	0,873
Милликена	58	0,7977	0,7977	0,864	0,863
Майкельсона	100	0,7790	0,7790	0,320	0,320
Ньюкомба	64	0,7745	0,7745	0,309	0,309

Из таблицы 2 видно, что расчет достигаемых уровней значимости, предложенный в данной работе, дает для классических экспериментов практически те же результаты, что и метод Монте-Карло.

Заключение. Предложенная в статье аппроксимация позволяет с высокой точностью рассчитывать процентные точки статистики критерия Гири и достигаемые уровни значимости и может быть рекомендована для практического применения, в том числе при автоматизированной обработке данных. В то же время предложенные ранее аппроксимации неудовлетворительны.

Список литературы

1. *Александровская Л. Н., Кириллин А. В.* Рекомендации по применению ряда критериев проверки отклонения распределения вероятностей от нормального закона в практике инженерного статистического анализа // Известия самарского научного центра российской академии наук. 2017. Т. 19. № 1. С. 82–90.

2. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М. : Наука, 1983. 416 с.

3. *Долганов Д. В., Долганова Т. И., Самылов В. В.* Оценка нарушений постуральной функции позвоночника в ортостатических стереотипах // Гений Ортопедии. 2018. Т. 24. № 3. С. 357–364.

4. *Ермаков С. М.* Метод Монте-Карло в вычислительной математике (вводный курс). СПб. : Невский диалект, 2009. 192 с.

5. *Зыков С. В., Незнанов А. А., Максименкова О. В.* Критерии отклонения распределения случайных величин от нормального в математическом обеспечении программных систем поддержки измерений в образовании // Программные системы: теория и приложения. 2018. Т. 9. № 4 (39). С. 199–218.

6. *Кобзарь А. И.* Прикладная математическая статистика. М. : Физматлит, 2012. 816 с.

7. *Ласкин М. Б., Пупенцова С. В.* Логарифмически нормальное распределение цен на объекты недвижимости // Имущественные отношения в Российской Федерации. 2014. № 5 (152). С. 52–59.

8. *Лемешко Б. Ю.* Критерии проверки отклонения распределения от нормального закона : руководство по применению. М. : ИНФРА-М, 2015. 160 с.

9. Методика валидации данных долгосрочных наблюдений скорости ветра для целей ветроэнергетики / Иванченко И. В., Пепелов А. В., Петренко Е. В., Тучинский Б. Г. // Международный научный журнал Альтернативная энергетика и экология. 2013. № 3 (121). С. 72–78.

10. *Смирнов В. И.* Курс высшей математики. СПб. : БХВ-Петербург, 2008. Т. 1. 624 с.

11. *Шаталов К. В., Черепанова А. Д.* Проверка гипотезы о соответствии погрешностей результатов анализа нефтепродуктов нормальному закону распределения случайной величины // Измерительная техника. 2019. № 10. С. 52–60.

12. *Geary R. C.* Moments of the ratio of the mean deviation to the standard deviation for normal samples // Biometrika. 1936. № 28. Pp. 295–307.

13. *Geary R. C.* Testing for normality // Biometrika. 1947. Vol. 34. № 3/4. Pp. 209–242.

14. *Geary R. C.* The ratio of the mean deviation to the standard deviation as a test of normality // Biometrika. 1935. V. 27. Pp. 310–322.

Approximation of the percentage points of the Geary criterion

Zalyazhnykh Vladimir Vasilyevich

PhD in Technical Sciences, associate professor of the Department of Standardization, Metrology and Certification,
M. V. Lomonosov Northern (Arctic) Federal University.
Russia, Arkhangelsk. ORCID: 0000-0001-6102-7221. E-mail: zalvladimir@yandex.ru

Abstract. The Geary criterion is one of the most powerful criteria for testing the hypothesis of the normality of the distribution. Its disadvantage consists in a significant dependence of the percentage points of the criterion statistics on the sample size and the value of the criterion distribution function, which complicates the assessment of the normality hypothesis both by the permissible and by the achieved significance level.

In this article, the Monte Carlo method has obtained a table of percentage points of the Geary criterion statistics in wide ranges of sample sizes and values of the criterion distribution function. According to her data, an approximation of percentage points is proposed, which makes it possible to find their values with high accuracy. This makes it possible to apply the Kettlebell criterion without difficulty. It is also shown that the previously proposed approximations of percentage points are unsatisfactory in accuracy.

Keywords: Geary criterion, approximation, percentage points, significance level.

References

1. *Aleksandrovskaya L. N., Kirillin A. V. Rekomendacii po primeneniyu ryada kriteriev proverki otkloneniya raspredeleniya veroyatnostej ot normal'nogo zakona v praktike inzhenernogo statisticheskogo analiza* [Recommendations on the application of a number of criteria for checking the deviation of the probability distribution from the normal law in the practice of engineering statistical analysis] // *Izvestiya samarskogo nauchnogo centra rossijskoj akademii nauk* – Proceedings of the Samara Scientific Center of the Russian Academy of Sciences. 2017. Vol. 19. No. 1. Pp. 82–90.

2. *Bol'shev L. N., Smirnov N. V. Tablicy matematicheskoj statistiki* [Tables of mathematical statistics]. M. Nauka (Science). 1983. 416 p.

3. *Dolganov D. V., Dolganova T. I., Samylov V. V. Ocenka narushenij postural'noj funkcii pozvonochnika v ortostaticheskih stereotipah* [Assessment of disorders of the postural function of the spine in orthostatic stereotypes] // *Genij Ortopedii* – Genius of Orthopedics. 2018. Vol. 24. No. 3. Pp. 357–364.

4. *Ermakov S. M. Metod Monte-Karlo v vychislitel'noj matematike (vvodnyj kurs)* [Monte Carlo method in computational mathematics (introductory course)]. SPb. Nevsky Dialect. 2009. 192 p.

5. *Zykov S. V., Neznanov A. A., Maksimenkova O. V. Kriterii otkloneniya raspredeleniya sluchajnyh velichin ot normal'nogo v matematicheskom obespechenii programmnyh sistem podderzhki izmerenij v obrazovanii* [Criteria for deviation of the distribution of random variables from normal in the mathematical support of software systems for measurement support in education] // *Programmnye sistemy: teoriya i prilozheniya* – Software systems: theory and applications. 2018. Vol. 9. No. 4 (39). Pp. 199–218.

6. *Kobzar' A. I. Prikladnaya matematicheskaya statistika* [Applied mathematical statistics]. M. Fizmatlit. 2012. 816 p.

7. *Laskin M. B., Pupencova S. V. Logarifmicheski normal'noe raspredelenie cen na ob'ekty nedvizhimosti* [Logarithmically normal distribution of prices for real estate objects] // *Imushchestvennye otnosheniya v Rossijskoj Federacii* – Property relations in the Russian Federation. 2014. No. 5 (152). Pp. 52–59.

8. *Lemeshko B. Yu. Kriterii proverki otkloneniya raspredeleniya ot normal'nogo zakona : rukovodstvo po primeneniyu* [Criteria for checking the deviation of the distribution from the normal law : application guide]. M. INFRA-M. 2015. 160 p.

9. *Metodika validacii dannyh dolgosrochnyh nablyudenij skorosti vetra dlya celej vetroenergetiki* – Methodology for validating data from long-term observations of wind speed for wind energy purposes / Ivanchenko I. V., Pepelov A. V., Petrenko E. V., Tuchinsky B. G. // International Scientific Journal Alternative Energy and Ecology. 2013. No. 3 (121). Pp. 72–78.

10. *Smirnov V. I. Kurs vysshej matematiki* [Course of higher mathematics]. SPb. BHV-Petersburg. 2008. Vol. 1. 624 p.

11. *Shatalov K. V., Cherepanova A. D. Proverka gipotezy o sootvetstvii pogreshnostej rezul'tatov analiza nefteproduktov normal'nomu zakonu raspredeleniya sluchajnoj velichiny* [Verification of the hypothesis about the correspondence of the errors of the results of the analysis of petroleum products to the normal distribution law of a random variable] // *Izmeritel'naya tekhnika* – Measuring technique. 2019. No. 10. Pp. 52–60.

12. *Geary R. C.* Moments of the ratio of the mean deviation to the standard deviation for normal samples // Biometrika. 1936. No. 28. Pp. 295–307.

13. *Geary R. C.* Testing for normality // Biometrika. 1947. Vol. 34. No. 3/4. Pp. 209–242.

14. *Geary R. C.* The ratio of the mean deviation to the standard deviation as a test of normality // Biometrika. 1935. V. 27. Pp. 310–322.