Сибирский государственный университет
телекоммуникаций и информатики

ОБРАБОТКА ИНФОРМАЦИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

РОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ

МАТЕРИАЛЫ КОНФЕРЕНЦИИ

Новосибирск
2018

ISBN 978-5-91434-042-8

СОДЕРЖАНИЕ

Секция 1
ИНФОРМАТИКА И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
Подсекция 1.1. НГТУ

УДК 519.24

О свойствах и мощности критериев нормальности Лина-Мудхолкара и Васичека

Б. Ю. Лемешко[1], В. Н. Белоцерковец

Рассмотрено несколько критериев, предназначенных для проверки гипотезы о принадлежности выборки нормальному закону распределения. Исследованы их недостатки и достоинства, проведено сравнение мощности критериев относительно конкурирующих гипотез.

Ключевые слова: критерий нормальности, критерий Васичека, критерий Лина-Мудхолкара.

1. Введение

Нормальный закон играет очень важную роль в приложениях теории вероятностей и математической статистики. Важной особенностью, выделяющей нормальный закон среди иных, является то, что он представляет собой предельный закон, к которому в определённых условиях сходятся другие законы распределения. Встречающиеся на практике случайные величины, например, ошибки измерений, отклонения от цели при стрельбе и другие, можно представить в виде суммы большого числа сравнительно малых слагаемых – элементарных ошибок, которые не зависят друг от друга. Такие ошибки могут подчиняться каким-либо иным законам распределения, в сумме большого числа слагаемых подобные особенности распределений нивелируются, и в результате сумма приближенно подчиняется нормальному закону распределения. При этом влияние каждого фактора на конечный результат должно быть существенно меньше суммарного влияния всех остальных факторов.

Проверка принадлежности наблюдаемых данных нормальному закону представляет собой важное условие для корректного применения множества параметрических методов математической статистики, используемых в задачах обработки измерений, стандартизации и контроля качества.

Настоящие исследования дополняют результаты, представленные в [1-5].

2. Критерий Лина-Мудхолкара

Данный критерий позволяет проверить сложную гипотезу нормальности наблюдаемого закона с неизвестными средним и (или) дисперсией.

Применение данного критерия не требует ни упорядочения, ни преобразования переменных, ни выборочной оценки параметров распределения. Распределение статистики критерия Лина-Мудхолкара при H_0 очень близко к нормальному закону распределения при малых выборках (в других критериях данная особенность чаще всего встречается в асимптотике).

[1] Работа выполнена при поддержке Министерства образования и науки РФ в рамках государственной работы «Обеспечение проведения научных исследований» (№ 1.4574.2017/6.7) и проектной части государственного задания (№ 1.1009.2017/4.6).

Критерий базируется на анализе n средних и дисперсий, которые рассчитываются по n подвыборкам, образуемым с помощью исключения из основной каждый раз одного наблюдения. Несмотря на то, что полученные пары значений $(\bar{x}_i, s_i^2)$ не будут являться независимыми, их можно использовать для построения критерия [5].

Распределение одной из пары величин, а именно распределение оценок дисперсий s_i^2, не является нормальным. Однако критерий предполагает использование коэффициента корреляции для нормальной двумерной совокупности [6-7]. Поэтому для оценки дисперсии применяется нормализующее представление Вилсона-Хилферти

$$y_i = \left\{ \frac{1}{n} \left[\sum_{j \neq i} x_j^2 - \frac{1}{n-1} \left(\sum_{j \neq i} x_j \right)^2 \right] \right\}^{\frac{1}{3}}, \ i = 1, \ldots, n. \qquad (1)$$

Эти случайные величины уже можно рассматривать как нормально распределенные.

Таким образом, исследуется связь между величинами $(\bar{x}_i, y_i)$, оцениваемая коэффициентом корреляции

$$r = \frac{\sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})(y_i - \bar{y})}{\left\{ \sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right\}^{\frac{1}{2}}}. \qquad (2)$$

В качестве статистики критерия рассматривается нормализующее представление

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}. \qquad (3)$$

Критерий двухсторонний. Можно отметить значительную зависимость распределений статистики (3) критерия от объема выборки, которую иллюстрирует рис. 1.

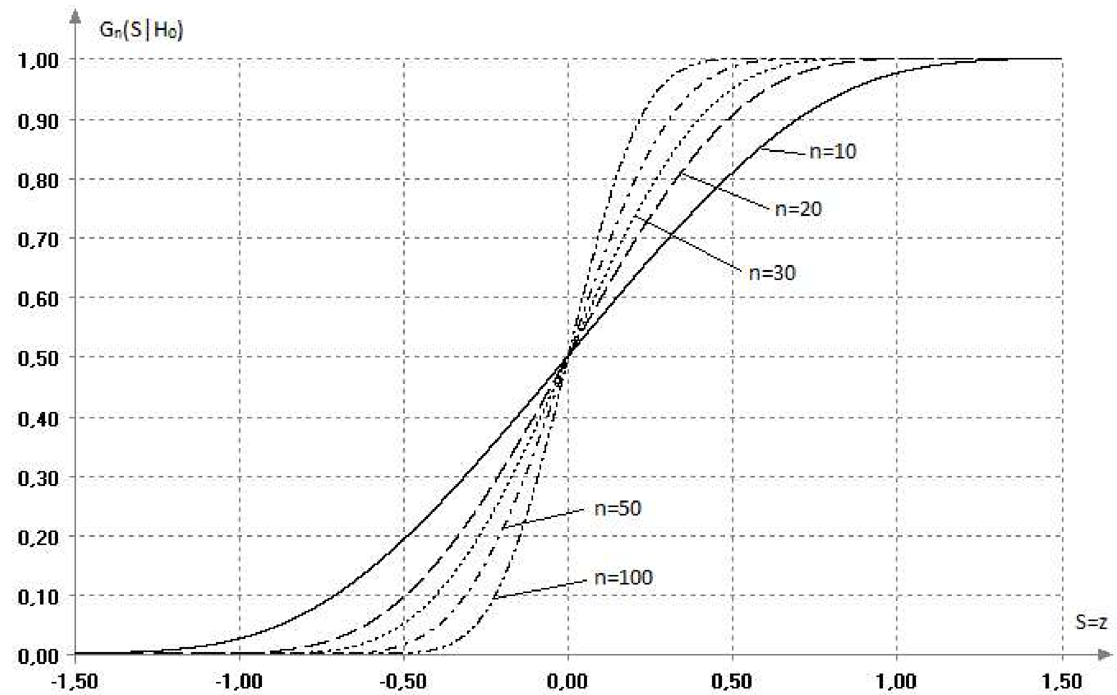

Рис. 1. Распределения статистики z критерия Лина–Мудхолкара в зависимости от объема выборки при n=10, 20, 30, 50, 100

Есть несколько отличающееся нормализующее представление данного критерия со статистикой

$$z = \frac{1}{2\sqrt{3n^{-1}}} \ln \frac{1+r}{1-r}$$

Результаты проверки согласия распределения статистики рассматриваемого критерия с нормальным законом представлены в таблице 1.

Таблица 1. Достигнутые уровни значимости при проверке гипотезы о принадлежности статистики (1) нормальному закону

Критерий	P-value
Колмогорова	0.277
Хи-квадрат Пирсона	0.057
Крамера–Мизеса–Смирнова	0.176
Андерсона–Дарлинга	0.066

3. Критерий Васичека

Энтропийный критерий нормальности Васичека базируется на том, что энтропия нормального закона превышает энтропию любого другого распределения с той же дисперсией [9-10].

Энтропия распределения вероятностей с плотностью $f(x)$ определяется выражением

$$H = -\int_{-\infty}^{\infty} f(x) \log f(x) dx, \tag{4}$$

а ее оценка по выборке соотношением

$$H_{mn} = \frac{1}{n} \sum_{i=1}^{n} \log \left\{ \frac{n}{2m} \left(x_{i+m} - x_{i-m} \right) \right\}, \tag{5}$$

где x_i - i-я порядковая статистика (при $i-m<1$ $x_{i-m}=x_1$; при $i+m>n$ $x_{i+m}=x_n$), m - целое положительное число (размер окна), не превышающее $n/2$.

Статистика критерия Васичека имеет вид

$$K_{mn} = \frac{n}{2ms} \left\{ \prod_{i=1}^{n} \left(x_{i+m} - x_{i-m} \right) \right\}^{\frac{1}{n}}, \tag{6}$$

где $s^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$; $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Критерий левосторонний. Если $K_{mn} < K_{mn}(\alpha)$, где $K_{mn}(\alpha)$ – критическое значение статистики, то гипотеза H_0 о принадлежности выборки нормальному закону отклоняется на уровне значимости α. При $n,m \to \infty$, $m/n \to \infty$ и справедливости проверяемой гипотезы $K_{mn} \to \sqrt{2\pi \cdot e} = 4{,}133$, при этом всегда $0 \le K_{mn} \le 4{,}133$.

Зависимость распределений статистики (6) критерия Васичека от объемов выборок для $m=1$ и $m=2$ показана на рисунках 2 и 3 соответственно. По картине, представленной на рисунках, можно отметить существенную зависимость распределения статистики критерия от объема выборки.

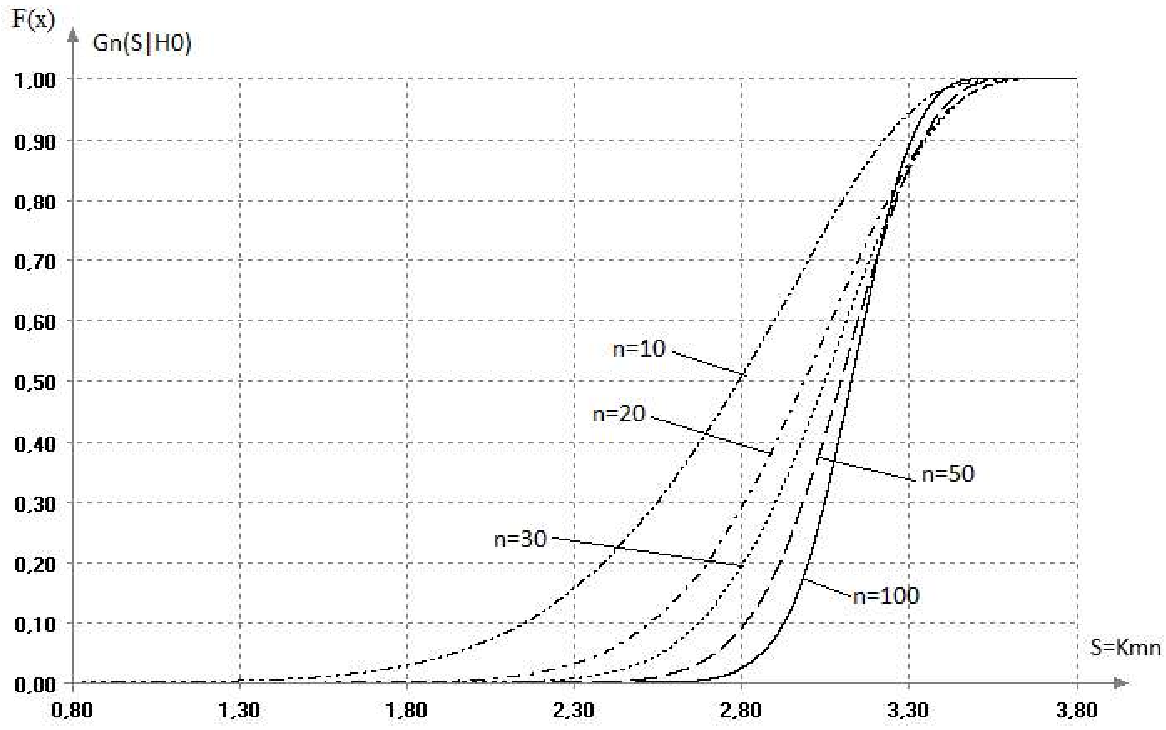

Рис. 2. Распределения статистики K_{1n} критерия Васичека в зависимости от объема выборки при *n*=10, 20, 30, 50, 100

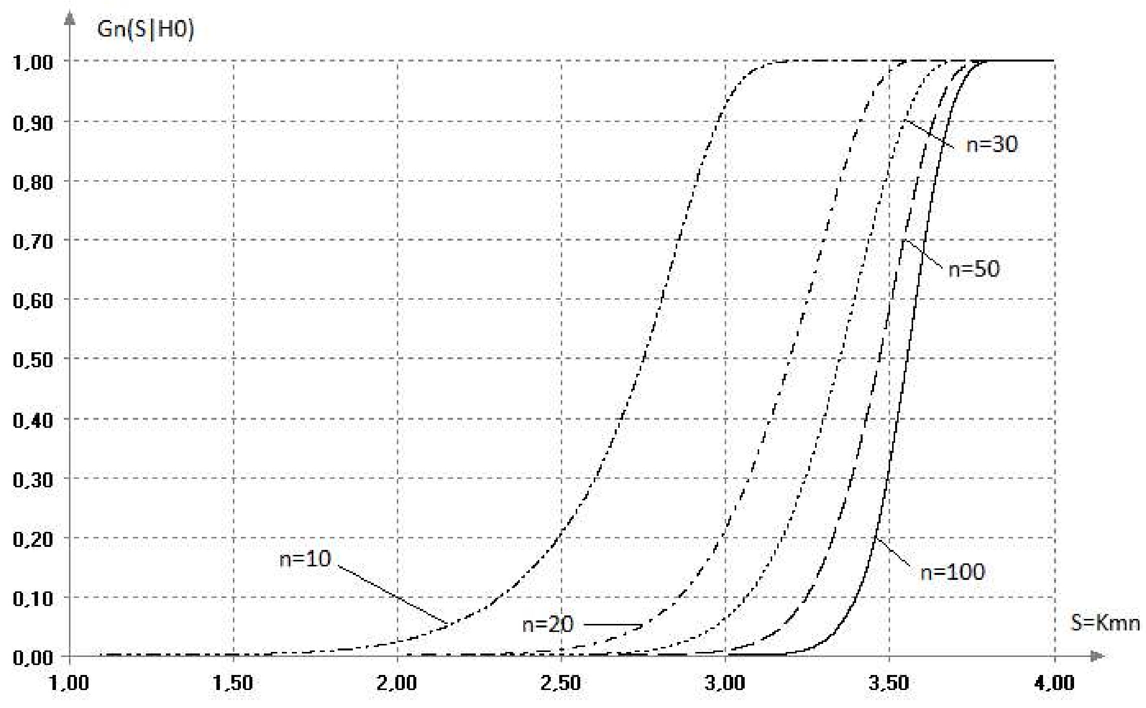

Рис. 3. Распределения статистики K_{2n} критерия Васичека в зависимости от объема выборки при *n*=10, 20, 30, 50, 100

Значения $K_{mn}(\alpha)$ для $\alpha = 0,05$ [9, 5], а также эти же критические значения K_{mn}, но полученные в результате моделирования в программе ISW, приведены в таблицах 2 и 3.

Существенное различие $K_{mn}(\alpha)$, приведенных в [5], от полученных в нашем случае, объясняется плохими качествами используемого датчика псевдослучайных чисел в [9], а также малым количеством имитационных экспериментов, по которым строились $K_{mn}(\alpha)$ в [9].

Таблица 2. Критические значения статистики $K_{mn}(\alpha)$ и их значения, полученные в результате моделирования

n	$K_{1n}(\alpha)$	K_{1n}	n	$K_{2n}(\alpha)$	K_{2n}	n	$K_{3n}(\alpha)$	K_{3n}
3	0.99	0.914	5	1.70	1.461	7	1.87	1.74
4	1.05	1.158	6	1.77	1.655	8	2.05	1.882
5	1.19	1.371	7	1.87	1.816	9	2.13	2.015
6	1.33	1.557	8	1.97	1.948	10	2.21	2.111
7	1.46	1.663	9	2.06	2.064	12	2.36	2.306
8	1.57	1.782	10	2.15	2.154	14	2.49	2.462
9	1.67	1.876	12	2.31	2.344	16	2.60	2.595
10	1.76	1.953	14	2.43	2.484	18	2.69	2.698
12	1.90	2.091	16	2.54	2.591	20	2.77	2.783
14	2.01	2.194	18	2.62	2.672	25	2.93	2.932
16	2.11	2.282	20	2.69	2.747	30	3.04	3.056
18	2.18	2.344	25	2.83	2.868	35	3.13	3.144
20	2.25	2.394	30	2.93	2.970	40	3.19	3.208
			35	3.00	3.043	45	3.25	3.262
						50	3.29	3.304

Таблица 3. Критические значения статистики $K_{mn}(\alpha)$ и их значения, полученные в результате моделирования

n	$K_{4n}(\alpha)$	K_{4n}	n	$K_{5n}(\alpha)$	K_{5n}
18	2.67	2.645	25	2.91	2.868
20	2.76	2.734	30	3.05	3.031
25	2.93	2.918	35	3.16	3.147
30	3.06	3.061	40	3.24	3.227
35	3.16	3.163	45	3.30	3.298
40	3.24	3.234	50	3.35	3.350
45	3.29	3.298			
50	3.34	3.345			

Критерий прост и не нуждается в таблице коэффициентов, как, например, популярный критерий Шапиро-Уилка. Считается, что критерий Васичека при проверке нормальности распределения наиболее эффективен против альтернатив равномерности и экспоненциальности [5]. Отметим, что в последнее время для проверки нормальности предлагаются и другие энтропийные критерии [11].

4. Сравнительный анализ мощности

При сравнительном анализе мощности критериев рассматривались следующие гипотезы:
– в качестве конкурирующей гипотезы H_1 рассмотрено распределение семейства

$$f(x) = \frac{\theta_2}{2\theta_1 \Gamma(1/\theta_2)} \exp\left\{ -\frac{|x-\theta_0|^{\theta_2}}{\theta_1} \right\} \tag{7}$$

с параметром формы $\theta_2 = 1$ (распределение Лапласа), параметрами масштаба $\theta_1 = 1$ и сдвига $\theta_0 = 0$;

– конкурирующей гипотезе H_2 соответствует логистическое распределение с плотностью

$$f(x) = \frac{\pi}{\theta_1\sqrt{3}} \exp\left\{ -\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}} \right\} / \left[1 + \exp\left\{ -\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}} \right\} \right]^2 \tag{8}$$

с параметрами $\theta_1 = 1$ и $\theta_0 = 0$;

– в качестве конкурирующей гипотезы H_3 рассмотрено распределение семейства (7) с параметром формы $\theta_2 = 4$, параметром масштаба $\theta_1 = 1$ и параметром сдвига $\theta_0 = 0$.

Оценки мощности критериев для различных объёмов выборок n получены по результатам статического моделирования условных распределений статистик $G(S|H_0)$ и $G(S|H_i)$ при числе имитационных экспериментов N=1660000, что гарантирует достаточно высокую точность оценок мощности.

Как можно видеть, среди рассмотренных критериев относительно всех конкурирующих гипотез наиболее мощным оказался критерий Васичека со статистикой K_{2n}.

При малых объёмах выборок критерий Васичека оказывается смещённым относительно гипотезы H_3.

Анализ мощности критериев Лина-Мудхолкара и Васичека показал, что они не слишком сильно уступают в мощности наиболее предпочтительным критериям нормальности. В то же время критерий Лина-Мудхолкара не способен различать гипотезы H_0 и H_3.

Таблица 4. Мощность критериев относительно конкурирующей гипотезы H_1

Критерий	n	α				
		0.15	0.1	0.05	0.025	0.01
Васичека K_{1n}	10	0.26	0.19	0.11	0.064	0.031
	20	0.336	0.259	0.166	0.105	0.057
	30	0.390	0.310	0.209	0.139	0.081
	50	0.473	0.389	0.277	0.195	0.123
	100	0.615	0.533	0.412	0.314	0.216
Лина-Мудхолкара	10	0.297	0.233	0.154	0.102	0.059
	20	0.349	0.286	0.204	0.146	0.095
	30	0.372	0.308	0.226	0.167	0.112
	50	0.394	0.331	0.248	0.187	0.130
	100	0.418	0.355	0.271	0.208	0.148
Васичека K_{2n}	10	0.271	0.200	0.120	0.073	0.038
	20	0.373	0.295	0.197	0.131	0.076
	30	0.454	0.0374	0.267	0.190	0.120
	50	0.571	0.492	0.377	0.287	0.197
	100	0.746	0.680	0.571	0.474	0.365

Таблица 5. Мощность критериев относительно конкурирующей гипотезы H_2

Критерий	n	α				
		0.15	0.1	0.05	0.025	0.01
Васичека K_{1n}	10	0.237	0.165	0.088	0.046	0.02
	20	0.242	0.172	0.095	0.052	0.024
	30	0.247	0.177	0.1	0.056	0.026
	50	0.256	0.186	0.108	0.062	0.03
	100	0.277	0.204	0.121	0.072	0.036
Лина-Мудхолкара	10	0.206	0.148	0.085	0.049	0.023
	20	0.236	0.180	0.11	0.068	0.036
	30	0.258	0.196	0.125	0.079	0.043
	50	0.278	0.215	0.14	0.092	0.053
	100	0.300	0.237	0.158	0.107	0.064
Васичека K_{2n}	10	0.279	0.194	0.104	0.056	0.025
	20	0.272	0.195	0.11	0.062	0.029
	30	0.279	0.203	0.119	0.069	0.034
	50	0.295	0.22	0.133	0.081	0.042
	100	0.332	0.255	0.161	0.102	0.056

Таблица 6. Мощность критериев относительно конкурирующей гипотезы H_3

Критерий	n	α				
		0.15	0.1	0.05	0.025	0.01
Лина-Мудхолкара	10	0.119	0.075	0.035	0.016	0.006
	20	0.104	0.064	0.028	0.013	0.004
	30	0.098	0.059	0.026	0.011	0.004
	50	0.092	0.055	0.023	0.01	0.003
	100	0.088	0.052	0.021	0.008	0.002
Васичека K_{1n}	10	0.124	0.083	0.041	0.021	0.008
	20	0.152	0.103	0.053	0.027	0.011
	30	0.174	0.121	0.064	0.034	0.014
	50	0.212	0.15	0.083	0.045	0.02
	100	0.286	0.218	0.123	0.071	0.034
Васичека K_{2n}	10	0.117	0.079	0.039	0.019	0.007
	20	0.171	0.119	0.065	0.035	0.015
	30	0.207	0.148	0.083	0.046	0.021
	50	0.265	0.196	0.115	0.067	0.032
	100	0.377	0.293	0.187	0.117	0.062

8. Заключение

В качестве общего недостатка рассмотренных критериев можно назвать сильную зависимость распределений статистик от объемов выборок и неизвестность аналитических распределений этих статистик. В этой связи на практике для принятия решения о результатах проверки гипотезы вынуждены использовать таблицы критических значений.

В ISW при использовании данных критериев реализована возможность моделирования эмпирических распределений статистик $G(S_n|H_0)$, что устраняет указанный недостаток и позволяет оценивать по $G(S_n|H_0)$ достигнутый уровень значимости p_{value}.

Литература

1. *Лемешко, Б.Ю.* Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход: Монография / Б.Ю. Лемешко, С.Б. Лемешко, С.Н. Постовалов, Е.В. Чимитова. – Новосибирск: Изд-во НГТУ, 2011. – 888 с.
2. *Лемешко, Б.Ю.* Критерии проверки отклонения распределения от нормального закона. Руководство по применению: Монография. – М.: ИНФРА-М, 2015. – 160 с.
3. *Лемешко Б.Ю., Рогожников А.П.* Исследование особенностей и мощности некоторых критериев нормальности // Метрология. 2009. № 4. С. 3-24.
4. *Лемешко Б.Ю., Лемешко С.Б.* Сравнительный анализ критериев проверки отклонения распределения от нормального закона // Метрология. 2005. № 2. С. 3-24.
5. *Кобзарь, А.И.* Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. М.: ФИЗМАТЛИТ, 2006. – 816 с.
6. *Lin Ch.-Ch., Mudholker G.S.* A simple test for normality against asymmetric alternatives // Biometrika. 1980. V. 67, № 2. – P. 455-461.
7. *Mudholkar G. S., Lin C. C.* On two applications characterization theorems to goodness-of-fit // Goodness-of-fit, Amsterdam, 1978. – P. 395-414.
8. *Nelson L. S.* A simple test for normality // J. of Quality Technology. 1983. V. 13. – P. 76-77.
9. *Vacicek O*. A test for normality based on sample entropy // JRSS. 1976. V. 38, № 1. – P. 54-59.
10. *Prescott P.* On a test for normality based on sample entropy // JRSS. 1976. V. 38, № 3. – P. 254-256.
11. *Zamanzade E., Arghami N.R.* Testing normality based on new entropy estimators // Journal of Statistical Computation and Simulation 2012. V. 82, № 11. P. 1701-1713.

Лемешко Борис Юрьевич
Д.т.н., профессор, г.н.с. кафедры теоретической и прикладной информатики Новосибирского государственного технического университета (630073, Новосибирск, просп. К. Маркса, 20), e-mail: `lemeshko@ami.nstu.ru`.

Белоцерковец Валерия Николаевна
Магистрант факультета прикладной математики и информатики Новосибирского государственного технического университета (630073, Новосибирск, просп. К. Маркса, 20), e-mail: `belocerkovecz.v@mail.ru`.

Properties and power of tests of a rejection from the normal law

B. Yu. Lemeshko, V. N. Belocerkovecz

Novosibirsk State Technical University

Several tests are considered, they designed to test the hypothesis of sampling accessories normal distribution. Advantages and disadvantages are studied, power are estimated for the competing hypotheses.

Keywords: normality test, Vasicek test, Lin-Mudholkar test.