Записки научных
семинаров ПОМИ
Том 441, 2015 г.

В. В. Литвинова, Я. Ю. Никитин

КРИТЕРИИ КОЛМОГОРОВА ДЛЯ ПРОВЕРКИ НОРМАЛЬНОСТИ, ОСНОВАННЫЕ НА ВАРИАНТАХ ХАРАКТЕРИЗАЦИИ ПОЙА

§1. Введение и постановка задачи

В 1923 г. Дж. Пойа доказал знаменитую теорему, в которой была найдена одна из первых характеризаций нормального закона.

Теорема Пойа [13]. *Пусть X_1 и X_2 – центрированные независимые одинаково распределенные случайные величины* (н.о.р.с.в.)*, такие что X_1 и $(X_1+X_2)/\sqrt{2}$ одинаково распределены. Тогда X_1 имеет нормальное распределение.*

Одним из применений этой теоремы в статистике может служить интегральный критерий нормальности, построенный в [10]. В этой работе показано, как вычислить его локальную бахадуровскую эффективность и произведены ее вычисления. Она оказалась весьма высокой. Например, для альтернативы сдвига ее значение равно 0.96.

Обобщением теоремы Пойа является следующее утверждение.

Теорема 1. *Пусть $X_1, X_2, \ldots, X_m$ – центрированные н.о.р.с.в. с функцией распределения* (ф.р.) *F, а вещественные постоянные $a_1, a_2, \ldots, a_m$ таковы, что $0 < a_i < 1$ и что $\sum\limits_{i=1}^{m} a_i^2 = 1$. Тогда статистики X_1 и $\sum\limits_{i=1}^{m} a_i X_i$ одинаково распределены тогда и только тогда, когда $F \in N(0, \tau^2)$ с некоторой дисперсией $\tau^2 > 0$.*

Этот результат является частным случаем более общей теоремы из монографии Кагана, Линника и Рао [6, §13.7], см. также [8] и [7, §2.1]. Теорема Пойа является частным случаем теоремы 1 для $m = 2$ и $a_1 = a_2 = \frac{1}{\sqrt{2}}$.

Ключевые слова: характеризация Пойа, критерий нормальности, бахадуровская эффективность, альтернатива сдвига.

Исследование второго автора поддержано грантом РФФИ 13-01-00172, грантом НШ No. 2504.2014.1 и грантом СПбГУ No. 6.38.672.2013.

Пусть $X_1, \ldots, X_n$ – н.о.р.с.в., имеющие ф.р. G. Будем проверять сложную гипотезу согласия H_0, согласно которой $G \in N(0, \tau^2)$ с некоторой неопределенной дисперсией τ^2 против альтернативы H_1, состоящей в том, что H_0 не выполняется.

Обозначим через $G_n(t)$, обычную эмпирическую ф.р., а именно

$$G_n(t) = n^{-1} \sum_{i=1}^{n} I\{X_i < t\},$$

и рассмотрим также так называемую V-статистическую эмпирическую ф.р.

$$V_{m,n}(t) = n^{-m} \sum_{i_1,\ldots,i_m=1}^{n} \frac{1}{m!} \Big(\sum_{\sigma} I\{a_{\sigma(i_1)} X_{i_1} + \cdots + a_{\sigma(i_m)} X_{i_m} < t\} \Big), \quad t \in \mathbf{R}^1,$$

где внутреннее суммирование берется по всем перестановкам σ индексов $i_1, \ldots, i_m$.

В дальнейшем понадобится также близкая к ней U-статистическая эмпирическая ф.р.

$$U_{m,n}(t) = \binom{n}{m}^{-1} \sum_{1 \leqslant i_1 < \cdots < i_m \leqslant n} \frac{1}{m!} \Big(\sum_{\sigma} I\{a_{\sigma(i_1)} X_{i_1} + \ldots + a_{\sigma(i_m)} X_{i_m} < t\} \Big), \quad t \in \mathbf{R}^1.$$

Введем в рассмотрение интегральную статистику

$$I_n = \int_{\mathbf{R}^1} (V_{m,n}(t) - G_n(t))\, dG_n(t),$$

которая может служить для проверки H_0 против H_1. Дело в том, что по теореме Гливенко–Кантелли для U и V-эмпирических функций распределения [5], подынтегральное выражение равномерно по t стремится п.н. к разности

$$\mathbf{P}\Big\{ \sum_{i=1}^{m} a_i X_i < t \Big\} - \mathbf{P}\{X_1 < t\},$$

которая при H_0 равна нулю в силу рассматриваемой характеризации. Поэтому статистика I_n при основной гипотезе должна быть мала.

В работе [9] были изучены асимптотические свойства этой статистики в двух простых частных случаях выбора констант $a_1, \ldots, a_m$, а именно:

1. $a_1^2 + a_2^2 = 1, \quad 0 < a_1, a_2 < 1.$
2. $a_1 = \cdots = a_m = 1/\sqrt{m}.$

В частности, были найдены предельные распределения и вычислена локальная бахадуровская эффективность (ЛБЭ). В первом случае, для параметра сдвига и при подходящем выборе параметров (например, $a_1 = \frac{7}{25}, a_2 = \frac{24}{25}$) она достигает 0.99. При этом неожиданно оказалось, что наихудшим является как раз классический случай Пойа, когда $a_1 = a_2 = \frac{1}{\sqrt{2}}$. Во втором случае ЛБЭ быстро стремится к 1 с ростом m и уже при $m = 10$ достигает 0.988.

Цель настоящей работы – вычислить в указанных двух случаях ЛБЭ статистик колмогоровского типа

$$K_n = \sup_t |G_n(t) - U_{m,n}(t)|.$$

В момент написания работы [9] это не представлялось возможным ввиду отсутствия информации о больших уклонениях изучаемых статистик. Использование U-статистической эмпирической ф.р. несколько упрощает статистики и позволяет воспользоваться результатами [12] об их больших уклонениях.

§2. Статистика Колмогорова первого типа

Не ограничивая общности, можно считать при H_0, что $\tau = 1$. Рассмотрим здесь статистику K_n при наличии только двух констант a и b, подчиненных соотношению $a^2 + b^2 = 1$, $0 < a, b < 1$. Предельное распределение этой статистики неизвестно, но критические значения для конкретных a и b можно найти с помощью моделирования.

Нашу статистику можно рассматривать как супремум по t модуля семейства (по t) U-статистик с ядрами

$$\Psi_1(X, Y; t) = \frac{1}{2}\big(I\{aX+bY < t\}+I\{aY+bX < t\}-I\{X < t\}-I\{Y < t\}\big).$$

В силу рассматриваемой характеризации эти ядра центрированы.

Имея в виду применить результат о больших уклонениях таких статистик из [12], вычислим проекцию ядра:

$$\begin{aligned}\psi_1(s,t) &:= \mathbf{E}\big(\Psi_1(X,Y;t) \,\big|\, Y=s\big)\\ &= \frac{1}{2}\left(\mathbf{P}\{aX+bs<t\}+\mathbf{P}\{as+bX<t\}-\mathbf{P}\{X<t\}-I\{s<t\}\right)\\ &= \frac{1}{2}\left(\Phi\Big(\frac{t-bs}{a}\Big)+\Phi\Big(\frac{t-as}{b}\Big)-\Phi(t)-I\{s<t\}\right).\end{aligned}$$

Тогда получаем

$$\begin{aligned}4\,\psi_1^2(X,t) &= \Phi^2\Big(\frac{t-bX}{a}\Big)+\Phi^2\Big(\frac{t-aX}{b}\Big)+\Phi^2(t)+I\{X<t\}\\ &\quad+2\,\Phi(t)\,I\{X<t\}+2\,\Phi\Big(\frac{t-bX}{a}\Big)\Phi\Big(\frac{t-aX}{b}\Big)\\ &\quad-2\,\Phi\Big(\frac{t-bX}{a}\Big)I\{X<t\}-2\,\Phi\Big(\frac{t-bX}{a}\Big)\Phi(t)\\ &\quad-2\,\Phi\Big(\frac{t-bX}{a}\Big)I\{X<t\}-2\,\Phi\Big(\frac{t-aX}{b}\Big)\Phi(t)\\ &\quad-2\,\Phi\Big(\frac{t-aX}{b}\Big)I\{X<t\}.\end{aligned}$$

Это позволяет вычислить функцию дисперсии $\sigma_1^2(t,a) := \mathbf{E}\,\psi_1^2(X,t)$ рассматриваемого семейства ядер как функцию от t и a:

$$\begin{aligned}\sigma_1^2(t,a) &= \frac{1}{4}\Big\{\int\limits_{-\infty}^{\infty}\Big(\Phi\Big(\frac{t-bx}{a}\Big)+\Phi\Big(\frac{t-ax}{b}\Big)\Big)^2\,d\Phi(x)\\ &\quad-2\int\limits_{-\infty}^{t}\Big(\Phi\Big(\frac{t-bx}{a}\Big)+\Phi\Big(\frac{t-ax}{b}\Big)\Big)\,d\Phi(x)-\Phi^2(t)+\Phi(t)\Big\}.\end{aligned}$$

Эта функция двух переменных (напомним, что $b=\sqrt{1-a^2}$) не поддается аналитическому исследованию, однако построение графика соответствующей двумерной поверхности с помощью пакета MAPLE привело к выводу, что при всех $a\in(0,1)$ максимум функции дисперсии достигается при $t=0$ и равен

$$\begin{aligned}\Delta_1^2(a,b) &= \sup_t\ \sigma_1^2(t,a)\\ &= \frac{1}{16}-\frac{1}{4\pi}\left(\operatorname{arctg}\frac{a}{\sqrt{1+b^2}}+\operatorname{arctg}\frac{b}{\sqrt{1+a^2}}-\operatorname{arctg}\frac{ab}{\sqrt{1-a^2b^2}}\right).\end{aligned}$$

Подтверждением того, что максимум функции дисперсии достигается в нуле, служит тот факт, что производная по t функции $\sigma_1^2(t,a)$ при фиксированном a (мы опустим это громоздкое выражение) обращается в нуль при $t=0$. Значение полученного выражения для $\Delta_1^2(a,b)$ в случае Пойа $a=b=\frac{1}{\sqrt{2}}$ совпадает со значением $\frac{1}{48}$, найденным в [12].

Поскольку семейство ядер центрировано и ограничено, то в соответствии с [12] мы можем написать при справедливости H_0

$$\lim_{n\to\infty} n^{-1}\ln \mathbf{P}\{K_n > z\} = h(z) \sim -\frac{z^2}{2\,m^2\,\Delta_1^2(a,b)}, \quad z\to 0,$$

где h – некоторая непрерывная функция, у которой важна асимптотика в нуле.

Отсылая читателя за необходимыми определениями и техникой вычисления ЛБЭ к монографиям [3] и [11], а также к работе [9], мы вычислим локальный бахадуровский наклон и ЛБЭ статистики K_n при альтернативе сдвига, т.е. в случае, когда альтернативная ф.р. $G(x;\theta) = \Phi(\frac{x+\theta}{\rho})$, $\theta, \rho > 0$.

Сначала необходимо вычислить вспомогательную величину

$$\begin{aligned} b_{1.1}(\theta;a,b;t) &:= \mathbf{E}_\theta \Psi_1(X,Y;t) = \mathbf{E}_\theta\left(I\{aX+bY<t\} - I\{X<t\}\right)\\ &= \mathbf{P}_\theta\{aX+bY<t\} - \mathbf{P}_\theta\{X<t\}\\ &= \mathbf{P}_\theta\Big\{a\frac{X+\theta}{\rho} + b\frac{Y+\theta}{\rho} < \frac{t+\theta(a+b)}{\rho}\Big\} - \mathbf{P}_\theta\{X<t\}\\ &= \Phi\Big(\frac{t+\theta(a+b)}{\rho}\Big) - \Phi\Big(\frac{t+\theta}{\rho}\Big)\\ &\sim \theta(a+b-1)\,\varphi(t/\rho), \quad \theta\to 0. \end{aligned}$$

Теперь, опираясь на сделанные вычисления, можно написать для локального точного наклона статистики Колмогорова K_n выражение при $\theta\to 0$

$$c_{1.1}(\theta;a,b) \sim \frac{b_{1.1}^2(\theta;a,b,0)}{4\Delta_1^2(a,b)} = \frac{\theta^2(a+b-1)^2}{8\pi\rho^2\Delta_1^2(a,b)}.$$

Асимптотика верхней границы наклона $K(\theta,\rho)$ в терминах информации Кульбака–Лейблера [3, 11] известна:

$$K(\theta,\rho) \sim \frac{\theta^2}{2\rho^2}, \quad \theta\to 0,$$

см. [9], следовательно, ЛБЭ в рассматриваемом случае составляет

$$\mathrm{eff}_{1.1}(a) = (a+b-1)^2 \Big/ \Big(\frac{\pi}{2} - 2\,\mathrm{arctg}\,\frac{a}{\sqrt{1+b^2}} - 2\,\mathrm{arctg}\,\frac{b}{\sqrt{1+a^2}} + 2\,\mathrm{arctg}\,\frac{ab}{\sqrt{1-a^2b^2}} \Big).$$

Аналитическое исследование этой функции от a затруднительно, но можно построить ее график, из которого видно, что максимальная эффективность довольно умеренна и составляет 0.328. По-видимому, она достигается в "случае Пойа" $a = b = \frac{\sqrt{2}}{2}$.

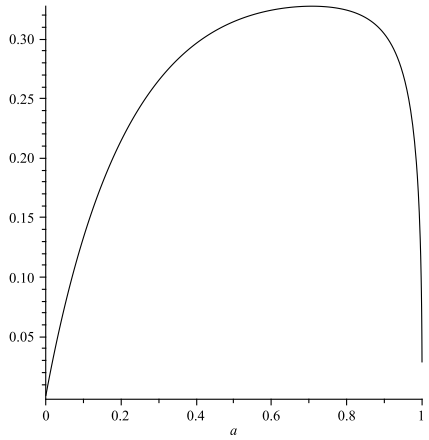

Рис. 1. График ЛБЭ статистики K_n как функции от a.

Теперь мы вычислим локальный бахадуровский наклон и локальную эффективность статистики K_n при скошенной альтернативе [1,2], т.е. если альтернативная плотность распределения $g(x;\theta) = 2\,\varphi(x)\,\Phi(\theta x)$.

Похожие вычисления дают нам

$$b_{1.2}(\theta; a, b; t) := \mathbf{P}_\theta\{aX + bY < t\} - \mathbf{P}_\theta\{X < t\}$$
$$= \int_{-\infty}^{+\infty} g(y;\theta)\, dy \int_{-\infty}^{\frac{t-by}{a}} g(x;\theta)\, dx - G(t;\theta)$$
$$\sim \sqrt{\frac{2}{\pi}}\, \varphi(t)\, (1 - a - b)\, \theta, \quad \theta \to 0.$$

Отсюда вытекает асимптотика локального точного наклона

$$c_{1.2}(\theta; a, b) \sim \frac{b^2_{1.2}(\theta; a, b, 0)}{4\, \Delta_1^2(a, b)} = \frac{\theta^2 (a + b - 1)^2}{4\, \pi^2\, \Delta_1^2(a, b)}.$$

Поскольку для скошенной альтернативы $K(\theta) \sim \frac{\theta^2}{\pi}$, см. [4], то ЛБЭ при скошенной альтернативе совпадает с эффективностью при альтернативе сдвига, вычисленной выше. Это – обычное явление, которое наблюдается также для других статистик [4].

Теперь рассмотрим лемановскую альтернативу, т.е. когда ф.р. $G(x;\theta) = \Phi^{1+\theta}(x)$. Аналогично предыдущему получаем

$$b_{1.3}(\theta; a, b; t) := \mathbf{P}_\theta\{aX + bY < t\} - \mathbf{P}_\theta\{X < t\}$$
$$= \int_{-\infty}^{+\infty} \Phi^{1+\theta}\Big(\frac{t - by}{a}\Big)\, d\Phi^{1+\theta}(y) - \Phi^{1+\theta}(t),$$

так что

$$c_{1.3}(\theta; a, b) \sim \frac{b^2_{1.3}(\theta; a, b, 0)}{4\Delta_1^2(a, b)}, \quad \theta \to 0.$$

Используя, что для альтернативы Лемана $K(\theta) \sim \frac{\theta^2}{2}$, см. [9], вычислим бахадуровскую эффективность. Вычисления проводились численно. Максимальная эффективность достигается, по-видимому, снова при $a = b = \frac{1}{\sqrt{2}}$, и она примерно равна 0,283. Сравнительно низкая ЛБЭ статистики Колмогорова в сравнении с интегральными статистиками неудивительна. Она частично компенсируется состоятельностью статистики Колмогорова против любых альтернатив.

§3. Статистика Колмогорова второго типа

Теперь проведем вычисления для статистики

$$K_{m,n} = \sup_t \big|G_n(t) - U_{m,n}(t)\big|, \quad m > 2.$$

Семейство ядер для статистики $K_{m,n}$ будет выглядеть следующим образом:

$$\Psi_2(X_1,\dots,X_m;t) = I\{X_1+\dots+X_m < t\sqrt{m}\} - \frac{1}{m}\big(I\{X_1<t\}+\dots+I\{X_m<t\}\big).$$

Вычислим проекцию ядра при фиксированном t:

$$\psi_2(s,t) = \mathbf{E}\big(\Psi_2(X_1,\dots,X_m;t)\mid X_m=s\big) = \mathbf{P}_\theta\{X_1+\dots+X_{m-1}<t\sqrt{m}-s\} - \frac{1}{m}\big(\mathbf{P}_\theta\{X_1<t\}+\dots+\mathbf{P}_\theta\{X_{m-1}<t\}+I\{s<t\}\big) = \Phi\Big(\frac{\sqrt{m}t-s}{\sqrt{m-1}}\Big) - \frac{m-1}{m}\,\Phi(t) - \frac{1}{m}\,I\{s<t\}.$$

Далее получаем

$$\psi_2^2(X,t) = \Phi^2\Big(\frac{\sqrt{m}\,t-X}{\sqrt{m-1}}\Big) + \frac{(m-1)^2}{m^2}\,\Phi^2(t) + \frac{1}{m^2}\,I\{X<t\} - \frac{2\,(m-1)}{m}\,\Phi(t)\,\Phi\Big(\frac{\sqrt{m}\,t-X}{\sqrt{m-1}}\Big) - \frac{2}{m}\,\Phi\Big(\frac{\sqrt{m}\,t-X}{\sqrt{m-1}}\Big)\,I\{X<t\} + \frac{2\,(m-1)}{m^2}\,\Phi(t)\,I\{X<t\}.$$

Найдем соответствующую функцию дисперсии $\sigma^2_{\psi_2}(t) = \mathbf{E}\,\psi_2^2(X,t)$:

$$\sigma^2_{\psi_2}(t) = \int\limits_{-\infty}^{\infty} \Phi^2\Big(\frac{\sqrt{m}t-x}{\sqrt{m-1}}\Big)\,d\Phi(x) - \frac{2}{m}\int\limits_{-\infty}^{t} \Phi\Big(\frac{\sqrt{m}t-x}{\sqrt{m-1}}\Big)\,d\Phi(x) - \frac{(m-1)^2}{m^2}\,\Phi^2(t) + \frac{1}{m^2}\,\Phi(t).$$

Исследуя функцию дисперсии графически, мы видим, что ее максимум достигается при $t=0$ и равен

$$\Delta_2^2(m) = \sup_t \sigma^2_{\psi_2}(t) = \frac{m^2+1}{4m^2} - \frac{1}{\pi}\operatorname{arctg}\frac{\sqrt{m-1}}{\sqrt{m+1}} - \frac{1}{\pi m}\operatorname{arctg}\frac{1}{\sqrt{m-1}}.$$

Вычислим локальный бахадуровский наклон и ЛБЭ статистики $K_{m,n}$ при альтернативе сдвига, т.е. если альтернативная ф.р. $G(x;\theta) = \Phi(\frac{x+\theta}{\rho})$, $\theta, \rho > 0$. Сначала находим

$$\begin{aligned} b_{2.1}(\theta; m; t) &:= \mathbf{E}_\theta \Psi_2(X_1, \dots, X_m; t) \\ &= \mathbf{E}_\theta \left(I\{X_1 + \cdots + X_m < t\sqrt{m}\} - I\{X < t\} \right) \\ &= \mathbf{P}_\theta\{X_1 + \cdots + X_m < t\sqrt{m}\} - \mathbf{P}_\theta\{X < t\} \\ &= \mathbf{P}_\theta\Big\{ \frac{X_1 + \theta}{\sqrt{m}\tau} + \cdots + \frac{X_m + \theta}{\sqrt{m}\tau} < \frac{t + \theta\sqrt{m}}{\tau} \Big\} - \mathbf{P}_\theta\{X < t\} \\ &= \Phi\Big(\frac{t + \theta\sqrt{m}}{\tau}\Big) - \Phi\Big(\frac{t + \theta}{\tau}\Big) \sim \theta(\sqrt{m} - 1)\, \varphi(t\rho), \end{aligned}$$

и, следовательно,

$$c_{2.1}(\theta; m) \sim \frac{b^2_{2.1}(\theta; m, 0)}{m^2\, \Delta_2^2(m)}.$$

Асимптотика верхней границы наклона известна [9]: $K(\theta, \rho) \sim \frac{\theta^2}{2\rho^2}$,

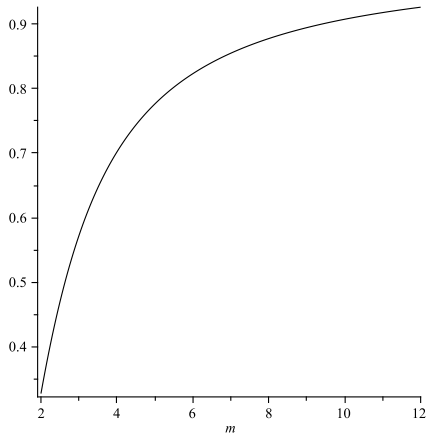

Рис. 2. График ЛБЭ статистики $K_{m,n}$ как функции от m.

$\theta \to 0$, следовательно, ЛБЭ составляет

$$\mathrm{eff}_{2.1}(m) = (\sqrt{m}-1)^2 \Big/ \Big(\frac{\pi(m^2+1)}{2} - 2\,m^2 \operatorname{arctg} \frac{\sqrt{m-1}}{\sqrt{m+1}} - 2\,m \operatorname{arctg} \frac{1}{\sqrt{m-1}} \Big).$$

Легко проверяется, что предел эффективности на бесконечности равен 1. График этой функции изображен на рисунке 2.

Заметим, что если взять $m = 10$, то эффективность принимает неожиданно высокое для статистик Колмогорова значение $\mathrm{eff}_{2.1}(10) \approx 0,907$, а при $m = 100$ получаем $\mathrm{eff}_{2.1}(100) \approx 0,993$. Случай $m = 10$ представляется вполне реалистическим для использования соответствующего критерия нормальности на практике.

Литература

1. A. Azzalini, *A class of distributions which includes the normal ones.* — Scand. J. Statist. **12** (1985), 171–178.
2. A. Azzalini, with the collaboration of A. Capitanio, *The Skew-normal and Related Families.* Cambridge University Press, New York, 2014.
3. R. R. Bahadur, *Some Limit Theorems in Statistics.* SIAM, Philadelphia, 1971.
4. A. Durio, Ya. Yu. Nikitin, *Local asymptotic efficiency of some goodness-of-fit tests under skew alternatives.* — J. Statist. Plann. Infer. **115**, No. 1 (2003), 171–179.
5. R. Helmers, P. Janssen, R. Serfling, *Glivenko–Cantelli properties of some generalized empirical DF's and strong convergence of generalized L-statistics.* — Probab. Theory Relat. Fields **79** (1988), 75–93.
6. A. M. Kagan, Yu. V. Linnik, C.R. Rao, *Characterization Theorems of Mathematical Statistics.* Wiley, New York, 1973.
7. A. V. Kakosyan, L. B. Klebanov, J. A. Melamed, *Characterization of Distributions by the Method of Intensively Monotone Operators.* Lecture Notes in Mathematics, **1088**, Springer, Berlin, 1984.
8. R. G. Laha, E. Lukacs, *On a linear form whose distribution is identical with that of a monomial.* — Pacific J. Math. **15** (1965), 207–214.
9. В. В. Литвинова, Я. Ю. Никитин, *Два семейства критериев нормальности, основанных на характеризации Пойа, и их асимптотическая эффективность.* — Зап. научн. семин. ПОМИ **328** (2005), 147–159.
10. P. Muliere, Ya. Yu. Nikitin, *Scale-invariant test of normality based on Polya's characterization.* — Metron **LX**, No. 1–2 (2002), 21–33.
11. Ya. Nikitin, *Asymptotic Efficiency of Nonparametric Tests.* Cambridge University Press, New York, 1995.
12. Ya. Yu. Nikitin, *Large deviations of U-empirical Kolmogorov–Smirnov tests, and their efficiency.* — J. Nonparam. Statist. **22** (2010), 649–668.
13. G. Polya, *Herleitung des Gauss'schen Fehlergesetzes aus einer Funktionalgleichung.* — Math. Zeitschrift **18** (1923), 96–108.

Litvinova V. V., Nikitin Ya. Yu. Kolmogorov tests of normality based on some variants of Polya's characterization.

Two variants of Kolmogorov-type U-empirical tests of normality are studied. They are based on the variants of famous Polya's characterization of the normal law. We calculate their local Bahadur efficiency against location, skew and Lehmann alternatives and find that the integral tests are usually more efficient.

С.-Петербургский государственный
университет путей сообщения и
Национальный университет
Высшая школа экономики
Союза Печатников, 16,
Санкт-Петербург 190008, Россия
E-mail: vikulenka@gmail.com

С.-Петербургский
государственный университет
Университетский пр. 28, Петродворец,
198504 Санкт-Петербург, Россия
E-mail: yanikit47@mail.ru

Поступило 2 ноября, 2015 г.