УДК 519.2

01.00.00 Физико-математические науки

**АСИМПТОТИКА ОЦЕНОК ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ**

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н., профессор
РИНЦ SPIN-код: 4342-4994
*Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5, prof-orlov@mail.ru*

Непараметрические оценки плотности распределения вероятностей в пространствах произвольной природы - один из основных инструментов нечисловой статистики. Рассмотрены их частные случаи – ядерные оценки плотности в пространствах произвольной природы, гистограммные оценки и оценки типа Фикс-Ходжеса. Цель настоящей статьи - завершение цикла работ, посвященного математическому изучению асимптотических свойств различных видов непараметрических оценок плотности распределения вероятности в пространствах общей природы. Тем самым подводится математический фундамент под применения таких оценок в нечисловой статистике. Начинаем с рассмотрения среднего квадрата ошибки ядерной оценки плотности и - с целью максимизации порядка его убывания - выбор ядерной функции и последовательности показателей размытости. Основные понятия - круговая функция распределения и круговая плотность. Порядок сходимости в общем случае тот же, что и при оценивании плотности числовой случайной величины, но основные условия наложены не на плотность случайной величины, а на круговую плотность. Далее рассматриваем другие виды непараметрических оценок плотности - гистограммные оценки и оценки типа Фикс-Ходжеса. Затем изучаем непараметрические оценки регрессии и их применение для решения задач дискриминантного анализа в пространстве общей природы

Ключевые слова: МАТЕМАТИЧЕСКАЯ СТАТИСТИКА, НЕЧИСЛОВАЯ СТАТИСТИКА, ПЛОТНОСТЬ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ, ПРОСТРАНСТВО ОБЩЕЙ ПРИРОДЫ, СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ, ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ, ОБЩАЯ ТОПОЛОГИЯ, БИКОМПАКТНОСТЬ

**Doi: 10.21515/1990-4665-131-070**

UDC 519:2

Physics and mathematical sciences

**ASYMPTOTICS OF ESTIMATES OF PROBABILITY DISTRIBUTION DENSITY**

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor
*Bauman Moscow State Technical University, Moscow, Russia*

Nonparametric estimates of the probability distribution density in spaces of arbitrary nature are one of the main tools of non-numerical statistics. Their particular cases are considered - kernel density estimates in spaces of arbitrary nature, histogram estimations and Fix-Hodges-type estimates. The purpose of this article is the completion of a series of papers devoted to the mathematical study of the asymptotic properties of various types of nonparametric estimates of the probability distribution density in spaces of general nature. Thus, a mathematical foundation is applied to the application of such estimates in non-numerical statistics. We begin by considering the mean square error of the kernel density estimate and, in order to maximize the order of its decrease, the choice of the kernel function and the sequence of the blur indicators. The basic concepts are the circular distribution function and the circular density. The order of convergence in the general case is the same as in estimating the density of a numerical random variable, but the main conditions are imposed not on the density of a random variable, but on the circular density. Next, we consider other types of nonparametric density estimates - histogram estimates and Fix-Hodges-type estimates. Then we study nonparametric regression estimates and their application to solve discriminant analysis problems in a general nature space

Keywords*:* MATHEMATICAL STATISTICS, NON-NUMERICAL STATISTICS, DENSITY OF PROBABILITY DISTRIBUTION, SPACE OF GENERAL NATURE, STATISTICAL ESTIMATION, LIMIT THEOREMS, GENERAL TOPOLOGY, BICOMPACTNESS

## 1. Введение

Согласно новой парадигме прикладной математической статистики [1], сердцевиной этой научной области является статистическая теория в пространствах произвольной природы. Эти пространства не предполагаются линейными. Подходы и результаты статистики в пространствах произвольной природы могут применяться могут применяться как при анализе числовых данных, так и нечисловых (бинарных отношений, множеств и др.). Статистическая теория в пространствах произвольной природы и статистические методы анализа конкретных нечисловых данных выделены в 1979 г. как самостоятельная область прикладной математической статистики [2, 3]. Она была названа статистикой объектов нечисловой природы. Позже ее стали называть статистикой нечисловых данных или нечисловой статистикой [4, 5].

Непараметрические оценки плотности распределения вероятностей в пространствах произвольной природы - один из основных инструментов нечисловой статистики. Систематическое изложение теории таких оценок начато в статьях [6 - 11], непосредственным продолжением которых является настоящая статья. Регулярно используются ссылки на условия и утверждения из статей [7, 9, 10].

Введем обозначения. Пусть $(Z, A)$ – измеримое пространство, $p$ и $q$ – сигма-конечные меры на $(Z, A)$, причем $p$ абсолютно непрерывна относительно $q$, т.е. из $q(B) = 0$ следует $p(B) = 0$ для любого множества $B$ из сигма-алгебры $A$. В этом случае на $(Z, A)$ существует неотрицательная измеримая функция $f(x)$ такая, что

$$q(C) = \int_C f(x)\, p(dx) \qquad (1)$$

для любого множества $C$ из сигма-алгебры измеримых множеств $A$. Функция $f(x)$ называется производной Радона - Никодима меры $q$ по мере

$p$, а в случае, когда $q$ - вероятностная мера, также плотностью вероятности $q$ по отношению к мере $p$ [12, с.460].

Пусть $X_1$, $X_2$ ,…, $X_n$ – независимые одинаково распределенные случайные элементы (величины), распределение которых задается вероятностной мерой $q$. В статьях [6, 7] введено несколько видов непараметрических оценок плотности вероятности $q$ по выборке $X_1$, $X_2$ ,…, $X_n$. Подробнее изучены линейные оценки. В статьях [8, 9] рассмотрены их частные случаи – ядерные оценки плотности в пространствах произвольной природы. В статьях [10, 11] асимптотическая теория ядерных оценок плотности развита, прежде всего, для нужд статистики конкретных видов объектов нечисловой природы, в которой основной интерес представляют конечные пространства $Z$. Мера $p$ при этом не непрерывная, а дискретная, например, считающая. Таким образом, в рамках единого подхода удается рассмотреть оценки плотностей и оценки вероятностей.

В предположении непрерывности неизвестной плотности $f(x)$ представляется целесообразным «размазать» каждый атом эмпирической меры, т.е. рассмотреть линейные оценки, введенные в нашей первой работе по нечисловой статистике [3, с.24]:

$$f_n(x) = \frac{1}{n} \sum_{1 \le i \le n} g_n(x, X_i), \quad g_n : Z^2 \to R^1, \qquad (2)$$

в которых действительнозначные функции $g_n$ удовлетворяют некоторым условиям регулярности.

Пусть $d$ – показатель различия (синоним - мера близости) на $Z$ [4] (в наиболее важных частных случаях – метрика на $Z$). В [13] нами введены ядерные оценки плотности – оценки вида (2) с

$$g_n(x, X_i) = \frac{1}{b(h_n, x)} K\left( \frac{d(x, X_i)}{h_n} \right), \quad K : [0, +\infty) \to R^1, \qquad (3)$$

где $K = K(u)$ – ядро (ядерная функция), $h_n$ – последовательность положительных чисел (показателей размытости), $b(h_n, x)$ – нормировочный множитель. В [14] линейные оценки (2) с функциями $g_n$ из (3) названы нами «обобщенными оценками типа Парзена-Розенблатта», т.к. в частном случае $Z = R^1$, $d(x, X_i) = | x - X_i |$, $b(h_n, x) = h_n$ они переходят в известные оценки, введенные Розенблаттом [15] и Парзеном [16].

Цель настоящей статьи - завершение цикла работ, начатого статьями [6 - 11], посвященного математическому изучению асимптотических свойств различных видов непараметрических оценок плотности распределения вероятности в пространствах общей природы. Тем самым подводится математический фундамент под применения таких оценок в нечисловой статистике (см. [3 - 5, 13, 14] и другие работы, в которых доказательства предельных свойств рассматриваемых оценок были опущены).

Начинаем с рассмотрения среднего квадрата ошибки ядерной оценки плотности

$$\alpha_n = M(f_n(x) - f(x))^2 = (Mf_n(x) - f(x))^2 + Df_n(x) \qquad (4)$$

и выбора последовательности $h_n$ с целью максимизации порядка убывания $\alpha_n$ при $n \to \infty$ (здесь и далее $M$ - символ математического ожидания, $D$ - символ дисперсии). Затем рассматриваем другие виды непараметрических оценок плотности - гистограммные оценки и оценки типа Фикс-Ходжеса. Затем изучаем непараметрические оценки регрессии и их применение для решения задач дискриминантного анализа в пространстве общей природы.

## 2. Круговая функция распределения

Пусть справедливы следующие условия регулярности:

R1) плотность $f(x)$ непрерывна в точке $x$, в которой оцениваем плотность;

R2) мера $p$ согласована с показателем различия $d$, т.е. мера шара радиуса $t$ равна $t$,

$$p\{y: d(x, y) < t\} = t, \qquad 0 \le t \le t_1 = p(Z), \qquad (5)$$

другими словами, показатель различия является предпочтительным [7],

R3) ядро $K(u)$ - непрерывная финитная функция, $K(u) = 0$ при $u > E$, такая, что

$$\int_0^E K(u)du = \int_0^\infty K(u)du = 1. \qquad (6)$$

Введем в рассмотрение шары $L_t(x) = \{y: d(x, y) < t\}$ радиуса $t$ и аналог функции распределения случайной величины $X$ со значениями в $Z$ с плотностью $f(x)$:

$$G(x,t) = P\{X \in L_t(x)\} = \int_{L_t(x)} f(y)\,p(dy). \qquad (7)$$

Назовем $G(x, t)$ круговой функцией распределения в точке $x$. Все нужные в дальнейшем свойства вероятностной модели, как будет показано, выражаются с помощью круговой функции распределения.

Из непрерывности плотности в точке $x$ и равенства (5) следует, что круговая функция распределения $G(x, t)$ дифференцируема по $t$ при $t = 0$ и

$$G_t'(x,0) = f(x). \qquad (8)$$

Изучим смещение оценки $f_n(x)$. Имеем цепочку равенств

$$Mf_n(x) = \frac{1}{h_n}\int_Z K\left(\frac{d(x,y)}{h_n}\right) f(y)\,p(dy) = \frac{1}{h_n}\int_0^\infty K\left(\frac{t}{h_n}\right) dG(x,t) = \int_0^E K(u)\frac{dG(x,h_n u)}{h_n}. \qquad (9)$$

Пусть справедливо следующее условие.

$R4$) В некоторой окрестности точки $t = 0$ (т.е. при $0 \le t \le t_0$ при некотором $t_0$) существует производная по $t$ круговой функцией распределения $G(x, t)$, т.е. $G_t'(x,t) = g(x,t)$.

Тогда при $h_n E < t_0$ имеем

$$Mf_n(x) = \int_0^E K(u)g(x,h_n u)du. \qquad (10)$$

Основная идея дальнейших рассуждений состоит в том, чтобы для изучения скорости сходимости ядерных оценок применить разложение $g(x, t)$ в ряд по степеням $t$ в окрестности $t = 0$ (в предположении существования указанного разложения).

R5) Пусть справедливо разложение

$$g(x,t) = g(x,0) + tg'_t(x,0) + \frac{t^2}{2} g''_{tt}(x,0) + o(h_n^2),\ 0 \le t \le h_n E \qquad (11)$$

(согласно (8) $g(x, 0) = f(x)$). Подставим это разложение в (10), получим с учетом (6) и непрерывности функции $K(u)$:

$$Mf_n(x) = f(x) + h_n g'_t(x,0)\int_0^E uK(u)du + h_n^2 g''_{tt}(x,0)\int_0^E u^2 K(u)du + o(h_n^2)\,. \qquad (12)$$

### 3. Первые оценки скорости сходимости

Согласно теореме 7 статьи [9] при справедливости рассматриваемых условий

$$Df_n(x) = \frac{f(x)}{nh_n}\int_0^E K^2(u)du + o\left(\frac{1}{nh_n}\right). \qquad (13)$$

Если

$$a = g'_t(x,0)\int_0^E uK(u)du \neq 0\,, \qquad (14)$$

то средний квадрат ошибки ядерной оценки плотности (4) согласно (12) и (13) равен

$$\alpha_n = h_n^2 a^2 + \frac{f(x)}{nh_n}\int_0^E K^2(u)du + o\left(h_n^2 + \frac{1}{nh_n}\right). \qquad (15)$$

Следовательно, при $f(x) \neq 0$ оптимальное по порядку скорости сходимости значение $h_n$ определяется из условия "уравнивания погрешностей" [2]

$$h_n^2 = \frac{1}{nh_n},\ h_n = n^{-1/3}. \qquad (16)$$

Тогда, как легко видеть, средний квадрат ошибки ядерной оценки плотности имеет порядок "$n$ в степени (-2/3)":

$$\alpha_n \cong Cn^{-2/3} \qquad (17)$$

при некоторой константе $C$.

Обоснуем сказанное более подробно. С точностью до бесконечно малых более высокого порядка средний квадрат ошибки ядерной оценки плотности равен

$$B(h) = a^2h^2 + \frac{F}{nh},\ F = f(x)\int_0^E K^2(u)du,\ h = h_n. \qquad (18)$$

С целью решения задачи оптимизации

$$B(h) \to \max$$

вычислим производную $B(h)$ по $h$ и приравняем ее 0:

$$B'(h) = 2a^2h - \frac{F}{nh^2} = 0. \qquad (19)$$

Решая уравнение (19) относительно $h$, получаем, что

$$h = h_n = \left(\frac{F}{2a^2n}\right)^{\frac{1}{3}} = \left(\frac{F}{2a^2}\right)^{\frac{1}{3}} n^{-1/3}. \qquad (20)$$

Соотношение (20) уточняет ранее полученное соотношение (17).

Для частного случая $Z = R^1$, т.е. для оценок Парзена-Розенблатта, соотношения (16) - (17) известны (см. [17, с.315]).

Если соотношение (14) не выполнено, т.е. $a = 0$, то согласно (12) заключаем, что

$$\alpha_n = [g''_{tt}(x,0)]^2\left[\int_0^E u^2K(u)du\right]^2 + \frac{f(x)}{nh_n}\int_0^E K^2(u)du + o\left(h_n^4 + \frac{1}{nh_n}\right). \qquad (21)$$

Оптимальное по порядку сходимости $h_n$ определяется из условия

$$h_n^4 = \frac{1}{nh_n},\ h_n = n^{-1/5}, \qquad (22)$$

и при этом

$$\alpha_n \cong C_1n^{-4/5}. \qquad (23)$$

при некоторой константе $C_1$.

Соотношения (22) - (23) совпадают с известными результатами для весьма частного случая $Z = R^1$, т.е. для оценок Парзена-Розенблатта (см. [17, с.316]).

**4. Примеры ядерных оценок**

Из сравнения формул (17) и (23) ясно, что сходимость убыстряется при $a = 0$, где $a$ определено в (14). Поскольку $a$ - произведение двух сомножителей, то $a = 0$ тогда и только тогда, когда $g'_t(x,0) = 0$ или

$$\int_0^E uK(u)du = 0. \qquad (24)$$

Роль сомножителей разная: первый определяется свойствами пространства с мерой, показателя различия (другими словами, меры близости) и плотности распределения случайной величины, а второй - свойствами ядра.

С целью выявления свойств первого сомножителя рассмотрим два примера.

*Пример 1.* Рассмотрим множество действительных чисел $Z = R^1$. Пусть $p$ - мера Лебега, $d$ - расстояние Евклида (показатель различия), $F(x)$ - функция распределения случайной величины $X$, причем ее плотность $f$ (по мере Лебега, т.е. в обычном смысле) дважды непрерывно дифференцируема. Тогда

$$G(x,t) = P\left\{X \in \left(x - \frac{t}{2}, x + \frac{t}{2}\right)\right\} = F\left(x + \frac{t}{2}\right) - F\left(x - \frac{t}{2}\right), \qquad (25)$$

а потому

$$g(x,t) = \frac{1}{2}\left\{f\left(x + \frac{t}{2}\right) + f\left(x - \frac{t}{2}\right)\right\}. \qquad (26)$$

Продифференцируем (26) по $t$:

$$g'_t(x,t) = \frac{1}{4}\left\{f'\left(x + \frac{t}{2}\right) - f'\left(x - \frac{t}{2}\right)\right\}. \qquad (27)$$

При $t = 0$ при всех $x$

$$g'_t(x,0) = 0 . \qquad (28)$$

*Пример 2.* Рассмотрим Z = [0, +∞). Пусть мера $p$ и расстояние $d$ получены из рассмотренных в примере 1 сужением на [0, +∞). Пусть функция распределения и плотность обладают теми же свойствами, что и в примере 1. Тогда

$$G(0,t) = F(t),\ g(0,t) = f(t),\ g'_t(0,t) = f'(t) . \qquad (29)$$

Следовательно, первый сомножитель в (14), вообще говоря, отличен от 0.

Можно показать, что для конечномерного пространства $Z = R^k$, меры Лебега $p$, евклидова расстояния $d$ и дважды дифференцируемой плотности $f$ первый сомножитель в (14) обращается в 0, а для мер $p$, отличных от Лебеговой, вообще говоря, не обращается (при сохранении прочих перечисленных в примерах 1 и 2 условий).

Из сказанного следует, что для введенных нами оценок (2) - (3) в случае конечномерного пространства $Z = R^k$, меры Лебега $p$, евклидова расстояния $d$ и дважды дифференцируемой плотности $f$ скорость сходимости задается формулой (23), а для классических оценок Парзена-Розенблатта - формулой (17) (см. также [17, с.315-316]), т.е. введенные нами оценки сходятся гораздо быстрее, чем оценки Парзена-Розенблатта.

## 5. Улучшение скорости сходимости ядерных оценок

Поскольку статистик может сам выбирать ядро, то для повышения скорости сходимости целесообразно принять условие (24). Однако скорость сходимости можно еще более повысить за счет выбора ядра из более узкого класса.

При более высокой гладкости круговой плотности $g(x, t)$ можно получить более высокую скорость сходимости среднего квадрата ошибки

ядерной оценки плотности $\alpha_n$, соответствующим образом выбирая ядро $K(u)$. Предположим, что круговая плотность $g(x,t)$ допускает разложение

$$g(x,t) = f(x) + tg'_t(x,0) + \frac{t^2}{2} g''_{tt}(x,0) + \frac{t^3}{3!} g'''_{ttt}(x,0) + ... + \frac{t^k}{k!} g^{(k)}_{t^{(k)}}(x,0) + o(h_n^k), \qquad (30)$$

причем остаточный член равномерно ограничен на отрезке $[0, h_nE]$. Тогда

$$Mf_n(x) = f(x) + h_n g'_t(x,0)\int_0^E uK(u)du + \frac{h_n^2}{2} g''_{tt}(x,0)\int_0^E u^2K(u)du + \\ + \frac{h_n^3}{3!} g'''_{ttt}(x,0)\int_0^E u^3K(u)du + ... + \frac{h_n^k}{k!} g^{(k)}_{t^{(k)}}(x,0)\int_0^E u^kK(u)du + \theta_n h_n^k, \qquad (31)$$

где $\theta_n \to 0$ при $n \to \infty$.

Пусть теперь

$$\int_0^E u^iK(u)du = 0, \quad i = 1,2,...,k-1. \qquad (32)$$

Тогда

$$Mf_n(x) - f(x) = \frac{h_n^k}{k!} g^{(k)}_{t^{(k)}}(x,0)\int_0^E u^kK(u)du + o(h_n^k) \qquad (33)$$

Следовательно,

$$\alpha_n = h_n^{2k}\left(\frac{1}{k!} g^{(k)}_{t^{(k)}}(x,0)\int_0^E u^kK(u)du\right)^2 + \frac{f(x)}{nh_n}\int_0^E K^2(u)du + o\left(h_n^{2k} + \frac{1}{nh_n}\right) \cong Ah_n^{2k} + \frac{B}{nh_n} \qquad (34)$$

при соответствующих $A$ и $B$. Оптимальная по порядку скорость сходимости будет при

$$h_n^{2k} = \frac{1}{nh_n}, \quad h_n = n^{-1/2k+1} \qquad (35)$$

(в предположении $A \neq 0,\ B \neq 0$). При этом

$$\alpha_n \cong n^{-2k/2k+1} = n^{\left(-1+1/2k+1\right)}. \qquad (36)$$

Этот результат - продвинутое обобщение теоремы 4.1 в книге И.А. Ибрагимова и Р.З. Хасьминского [17, с.316], относящейся к оцениванию плотности одномерной случайной величины. Порядок сходимости в общем случае тот же, что и в указанной теореме, но условия наложены не на

плотность $f(x)$, а на круговую плотность $g(x, t)$. Это существенно в случае $Z \neq R^k$, т.к. в $R^k$ имеются традиционно выделенные мера (Лебега) и расстояние (Евклида).

С прикладной точки зрения предположение (30) о гладкости круговой плотности представляется достаточно естественным. Напомним, что вплоть до XIX в. математики практически не делали различия между непрерывными, дифференцируемыми и аналитическими функциями. Инженеры не делают такого различия и сейчас. Отсюда методологическое предложение: в прикладных задачах допустимо использовать математические модели с той степенью гладкости рассматриваемых функций, которая позволяет наиболее легко обосновать алгоритмы расчетов, разумеется, если эта степень гладкости не противоречит фактам соответствующей предметной области. Другой пример подобного методологического подхода: шкала любого прибора конечна, поэтому результаты первичных измерений целесообразно моделировать с помощью финитных случайных величин; такие величины имеют все моменты, а это существенно облегчает получение для них предельных теорем. Методологическим вопросам посвящены наши работы [18, 19].

Из соотношения (36) следует, что для любого $\varepsilon > 0$ существует ядро $K(u)$ такое, что для соответствующей оценки

$$\alpha_n = O\left(n^{-1+\varepsilon}\right), \qquad (37)$$

а также

$$\lim_{n\to\infty} n^{+1/2-\varepsilon}\left(f_n(x) = f(x)\right) = 0 \qquad (38)$$

по вероятности. Хотя при любом $k$ множество ядер $K(u)$, удовлетворяющих соотношениям (32), бесконечно, не существует ядра, удовлетворяющего этим соотношениям сразу при всех $k$.

Перенос полученных результатов на случай конечных пространств объектов нечисловой природы осуществляется тем же путем, что и в статьях [10, 11]. Грубо говоря, необходимо, чтобы

$$\alpha_{mn}(g) = o(h_n^k), \qquad (39)$$

где $m$ - параметр дискретности (см. [10]). Поскольку принципиальных трудностей, как ясно из рассуждений статьи [10], в указанном переносе нет, мы его здесь не приводим.

## 6. Гистограммные оценки

Развитие теории гистограммных оценок облегчается тем, что гистограмме в произвольном пространстве можно поставить в соответствие гистограмму одномерной случайной величины, соотнеся область в произвольном пространстве $Z$ и отрезок той же меры и той же вероятности попадания в него. Так, теорема 1 из основополагающей статьи Н.В. Смирнова [20] может быть перенесена на случай произвольного пространства $Z$ следующим образом (теоремы 1 - 2).

***Теорема 1.*** *Пусть $\alpha$, $\beta$ и $\gamma$ - положительные константы. При каждом $n$ = 1, 2, .... рассмотрим $k(n)$ положительных чисел $p_1(n)$, $p_2(n)$, ..., $p_{k(n)}(n)$ таких, что*

$$p_1(n) + p_2(n) + ... + p_{k(n)}(n) = 1 - \alpha < 1, \qquad (40)$$

$$\frac{\beta}{k(n)} \le p_i(n) \le \frac{\gamma}{k(n)}, \ i = 1,2,...,k(n) \qquad (41)$$

*Пусть проводится $n$ независимых мультиномиальных испытаний с $k(n)+1$ исходами, имеющими вероятности $p_1(n)$, $p_2(n)$, ..., $p_{k(n)}(n)$, $p_{k(n)+1}(n) = \alpha$ соответственно. Обозначим $m_1(n)$, $m_2(n)$, ..., $m_{k(n)}(n)$, $m_{k(n)+1}(n)$ количество осуществлений каждого из исходов. Положим*

$$M_n = \max\left( \frac{|m_i(n) - np_i(n)}{\sqrt{np_i(n)}}, \ i = 1,2,...,k(n) \right). \qquad (42)$$

*Пусть $k(n) \to \infty$ и*

$$\overline{\lim_{n\to\infty}}\left(\frac{k^3(n)(\ln k(n))^3}{n}\right)<\infty. \qquad (43)$$

*Тогда при любом действительном* $\lambda$

$$\lim_{n\leftarrow\infty} P\left\{M_n < m(k(n))+\frac{\lambda}{m(k(n))}\right\}=\exp\{-2\exp(-\lambda)\}, \qquad (44)$$

*где* $m(k(n))$ - *решение уравнения*

$$\frac{1}{\sqrt{2\pi}}\int_{m(k(n))}^{\infty}\exp\left\{-\frac{1}{2}x^2\right\}dx=\frac{1}{k(n)}. \qquad (45)$$

В постановке Н.В. Смирнова каждый из упомянутых в теореме 1 исходов состоит в попадании в один из интервалов равной длины, на которые разбит отрезок для построения гистограммы. Условие (40) - это условие (В) Н.В. Смирнова, левое ограничение в (41) - условие (А) Н.В. Смирнова, правое - вытекает из непрерывности оцениваемой плотности. Конечно, сказанное не дает гарантии, что любой последовательности мультиномиальных распределений, удовлетворяющей перечисленным в теореме 1 условиям, можно поставить в соответствие задачу оценивания плотности с помощью гистограммы, т.к. последняя предполагает вполне определенную согласованность указанных мультиномиальных распределений между собой. Однако анализ рассуждений Н.В. Смирнова показывает, что им фактически доказана именно сформулированная выше теорема 1, а результаты об оценивании одномерной плотности с помощью гистограммы можно рассматривать как следствия из этой теоремы.

В начале 70-х годов Э.А. Надарая показал [21, 22], что условие (40) можно отбросить (для дальнейшего удобнее принять, что в (40) можно положить $\alpha$ = 0), а условие (43) заменить на более слабое

$$\frac{k(n)(\ln k(n))^3}{n}\to 0 \qquad (46)$$

(см. также монографию Г.М. Мания [23, с.88]). Известно асимптотическое разложение левой части (44) по степеням $ln(k(n))$, найдены другие

аппроксимирующие выражения, более быстро сходящиеся, оценена скорость сходимости (см. статью В.Д. Конакова [24]).

***Теорема 2.*** *Пусть* $p(Z) = 1$. *Пусть существуют* $0 < \chi < \nu$ *такие, что для плотности* $f(x)$ *случайного элемента имеем при всех* $x \in Z$

$$\chi \le f(x) \le \nu . \qquad (47)$$

*Пусть для построения гистограммных оценок используются* $k(n)$ *областей равной меры* $X_1^n, X_2^n, ..., X_{k(n)}^n$, *т.е.*

$$X_1^n \cup X_2^n \cup ... \cup X_{k(n)}^n = Z,\ X_i^n \cap X_j^n = \varnothing,\ i \ne j,\ p(X_i^n) = \frac{1}{k(n)}. \qquad (48)$$

*Положим*

$$f^*(x) = \frac{1}{k(n)} \int_{X_i^n} f(y)\, p(dy),\ x \in X_i^n\ i = 1,2,...,k(n). \qquad (49)$$

*Пусть* $k(n) \to \infty$ *и выполнено* (46). *Тогда при любом* $\lambda$

$$P\left\{ \max_x \left| \frac{f_n(x) - f^*(x)}{\sqrt{f^*(x)}} \right| < \sqrt{\frac{k(n)}{n}} \left( m(k(n)) + \frac{\lambda}{m(k(n))} \right) \right\} \to \exp\left(-2e^{-\lambda}\right) \qquad (50)$$

*при* $n \to \infty$, *где гистограммная оценка определяется как частный случай линейной оценки*

$$f_n(x) = \frac{1}{n} \sum_{1 \le i \le n} g_n(x, X_i).$$

(*Гистограммные оценки определяются с помощью последовательности* $T_n$ *разбиений* $Z$ *и функций*

$$g_n(x, X) = \begin{cases} \dfrac{1}{p(A(x))}, & x \in A(x), \\ 0, & x \notin A(x), \end{cases}$$

*где* $A(x)$ - *элемент разбиения* $T_n$, *которому принадлежит* $x$.)

Теорема 2 непосредственно вытекает из теоремы 1 и цитированных выше результатов Э.А. Надарая.

***Теорема 3.*** *В условиях теоремы 2 для любого* $\varepsilon > 0$

$$\lim_{n \to \infty} P\left\{ \sup_{x \in Z} \left| f_n(x) - f^*(x) \right| > \varepsilon \right\} = 0. \qquad (51)$$

*Доказательство*. Для корня $m(k) = m(k(n))$, $k = k(n)$ уравнения (45) справедливо асимптотическое равенство (см., например, [20, ф-ла (95)])

$$m(k) = \sqrt{2\ln k} - \frac{\ln\ln k + \ln(4\pi)}{2\sqrt{2\ln k}} + O\left(\frac{1}{\ln k}\right). \qquad (52)$$

Из (46) и (52) следует, что

$$\sqrt{\frac{k(n)}{n}}\left(m(k(n)) + \frac{\lambda}{m(k(n))}\right) < (\ln k + \lambda)\sqrt{\frac{k}{n}} < \frac{1}{\sqrt{\ln k}} + \frac{\lambda}{\ln^{3/2} k} \qquad (53)$$

при достаточно большом $n$. Тогда из (47), (49), (50) и (53) следует, что при достаточно большом $n$ и любом фиксированном $\lambda$

$$P\left\{\sup_x |f_n(x) - f^*(x)| > \sqrt{\nu}\left[(\ln k)^{-1/2} + \lambda(\ln k)^{-3/2}\right]\right\} < 1 - \exp\{-2\exp(-\lambda)\}, \qquad (54)$$

откуда и следует (51).

*Замечание.* Ясно, что можно отказаться от рассмотрения областей равной меры, однако при этом аналоги теорем 2 и 3 будут формулироваться несколько более громоздко. Поскольку принципиально нового при этом не появляется, мы ограничивается сказанным выше.

Чтобы установить равномерную сходимость гистограммных оценок, в силу теоремы 3 достаточно указать условия, при которых

$$\sup_{x \in Z} |f(x) - f^*(x)| \to 0 \qquad (55)$$

при измельчении разбиения. Понадобятся некоторые дополнительные результаты.

Гистограммной в общем случае называют функцию $f^*: Z \to R^1$ такую, что

$$f^*(x) = f(x_i^n),\ x \in X_i^n,\ i = 1,2,...,k(n), \qquad (56)$$

для некоторых $x_i^n \in X_i^n,\ i = 1,2,...,k(n)$. Если плотность $f$ непрерывна, а $X_i^n$ - связные бикомпакты, то формула (49) дает частный случай (56).

***Теорема 4.*** *Пусть Z счетно компактно, топология в Z порождена естественной мерой близости* [9]. *Тогда Z бикомпактно*.

Пусть $f$ - непрерывная функция, $Z$ - счетно-компактно, топология в $Z$ порождена естественной мерой близости, Тогда условие (55) выполнено.

***Теорема 5.*** *Пусть $Z$ - счетно-компактно, топология в $Z$ порождена естественной мерой близости, мера $p$ безатомная, плотность $f$ непрерывна и положительна. Тогда существует последовательность разбиений такая, что для любого $\varepsilon > 0$*

$$\lim_{n\to\infty} P\left\{\sup_{x\in Z} | f(x) - f^*(x) | > \varepsilon\right\} = 0, \qquad (57)$$

*где $f_n(x)$ - гистограммная оценка плотности, причем указанная последовательность разбиений не зависит от плотности $f$ (одна и та же для всех положительных непрерывных плотностей $f$).*

При этом мера $p$ называется безатомной, если для любого измеримого подмножества $A$ с $p(A) > 0$ и любого числа $0 < \delta < p(A)$ найдется измеримое подмножество $B \subset A$ такое, что $p(B) = \delta$.

*Замечание.* Теорема 5 относится к случаю бикомпактного $Z$. Если $Z$ не таково, то равномерная сходимость имеет место для бикомпактных подмножеств $Z$, что устанавливается с помощью теоремы 1 (без модификации Э.А. Надарая).

*Доказательство* теоремы 5. Их теоремы 4 следует, что выполнено следующее условие:

*Условие А.* Существует последовательность разбиений, измельчающаяся и такая, что для любой точки $x \in Z$ и любой ее окрестности $U(x)$ найдется такое разбиение из этой последовательности, что его область, содержащая $x$, полностью содержится в $U(x)$.

Будем рассматривать последовательность разбиений, указанную в этом условии. Ясно, что в силу безатомности $p$ можно считать, что отношения мер одного разбиения равномерно отделены от 0 и $\infty$. В силу теоремы 4 $Z$ бикомпактно, а потому плотность $f$ достигает своих минимального и максимального значений. Поскольку $f$ всюду

положительна, то минимальное значение также положительно, т.е. выполнены неравенства (8). Число областей в разбиениях искомой последовательности будем регулировать так, чтобы выполнялось (46) (для этого достаточно нужное число раз повторить одно и то же разбиение из исходной последовательности, рассмотренной выше в связи с условием А). Тогда по теореме 3 справедливо (51) (в связи с заменой условия равной меры областей разбиений на условие отделенности отношений мер областей от 0 и ∞ необходимо использовать замечание после теоремы 3). По теореме 4 справедливо (55). Из (51) и (55) вытекает (57). Теорема 5 доказана.

**7. Оценки типа Фикс-Ходжеса**

Обобщенные оценки типа Фикс-Ходжеса [7] определяются с помощью расширяющейся последовательности множеств $U(x, r)$. Ограничимся частным случаем

$$U(x, r) = L_r(x),\ p(L_r(x)) = r,\ 0 \le r \le r_0. \qquad (58)$$

Некоторое обобщение дают ядерные оценки со случайными $h_n = h_n(x, \omega)$, имеющие вид

$$f_n(x) = \frac{1}{n h_n} \sum_{1 \le i \le n} K\left( \frac{d(x, X_i)}{h_n(x; X_1, X_2, ..., X_n)} \right), \qquad (59)$$

где

$$h_n = h_n(x; X_1, X_2, ..., X_n) = \inf\left\{ r : \sum_{1 \le i \le n} \chi(X_i \in L_r(x)) \ge k_n \right\}, \qquad (60)$$

где $\chi(C) = 1$, если условие С выполнено, и $\chi(C) = 0$ в противном случае.

Если

$$K(u) = \begin{cases} 1, & 0 \le u \le 1, \\ 0, & u > 1, \end{cases} \qquad (61)$$

то оценка, задаваемая соотношениями (2) - (3), является обобщенной оценкой типа Фикс-Ходжеса [7]. Таким образом, ядерные оценки и оценки

типа Фикс-Ходжеса и оценки типа Фикс-Ходжеса можно рассматривать в рамках одной и той же схемы. Однако в силу (60) оценка (59) не является суммой независимых одинаково распределенных случайных величин, что затрудняет ее изучение.

Рассмотрим распределение случайной величины $h_n(\omega)$ из (60). Ясно, что $h_n$ является $k_n$-ой порядковой статистикой совокупности $\{d(x, X_i), i = 1, 2, ..., n\}$, а потому при естественных предположениях имеет асимптотически нормальное распределение. Укажем эти предположения.

Функцией распределения случайных величин $\eta_i = d(x, X_i),\ i = 1,2,...$ является круговая функция распределения

$$G(x,t) = P\{X \in L_t(x)\} = \int_{L_t(x)} f(y)\,p(dy).$$

Предположим, что она непрерывна и строго возрастает. Имеем

$$P\{G(x,h_n) \le y\} = P\{h_n \le G^{-1}(x,y)\} = P\left\{\sum_{1\le i\le n} \chi(\eta_i \in [0;G^{-1}(x,e)] \ge k_n\right\}. \qquad (62)$$

Справа в (5) стоит вероятность того, что не менее $k_n$ успехов имело быть в $n$ испытаниях Бернулли с вероятностью успеха $p = G(x, G^{-1}(x, y)) = y$ в каждом.

Рассмотрим последовательность $y = y_n$, $n = 1, 2, ...,$ такую, что

$$ny_n \to \infty. \qquad (63)$$

Тогда к правой части (62) можно применить центральную предельную теорему (см., например, [25, с.121]), т.е.

$$\lim_{n\to\infty} \max_k \left| P\left\{\sum_{1\le i\le n} \chi(\eta_i \in [0;G^{-1}(x,y_n)] \ge k\right\} - \Phi\left(\frac{ny_n - k}{\sqrt{ny_n(1-y_n)}}\right)\right| = 0. \qquad (64)$$

Дополнительно к (63) предположим, что $y_n$ меняется так, что при некоторых (произвольных, но фиксированных) $a$ и $b$

$$a < \frac{ny_n - k_n}{\sqrt{ny_n(1-y_n)}} < b. \qquad (65)$$

Ясно, что с учетом (63) для этого необходимо выполнения условия

(I) $k_n \to \infty$ при $n \to \infty$.

Это условие соответствует условию "$nh_n \to \infty$ при $n \to \infty$" для ядерных оценок [9].

Поскольку из (65) вытекает, что

$$\left| y_n - \frac{k_n}{n} \right| \le \frac{\max(|a|,|b|)}{2\sqrt{n}}, \qquad (66)$$

то в (64) можно заменить $ny_n(1 - y_n)$ на $k_n(1 - k_n/n)$, т.е. можно переписать (64) в виде

$$\lim_{n\to\infty} \sup_{w\in[a;b]} \left| P\left\{ \frac{nG(x,h_n) - k_n}{\sqrt{k_n\left(1 - \frac{k_n}{n}\right)}} \le w \right\} - \Phi(w) \right| = 0. \qquad (67)$$

Таким образом. величина $G(x, h_n(\omega))$ является асимптотически нормальной случайной величиной с параметрами

$$MG(x,h_n(\omega)) \approx \frac{k_n}{n}, \quad DG(x,h_n(\omega)) \approx \frac{1}{n}\frac{k_n}{n}\left(1 - \frac{k_n}{n}\right). \qquad (68)$$

***Теорема 6.*** *Пусть мера p и метрика (мера близости, показатель различия) d связаны соотношением* (58). *Пусть плотность f(x) непрерывна и ограничена на Z. Пусть ядро K(u) таково, что*

$$\int_0^\infty K(u)du = 1, \quad \int_0^\infty |K(u)|\,du < \infty. \qquad (69)$$

*Пусть последовательность* $k_n$ *удовлетворяет условиям*

$$k_n \to \infty, \ \frac{k_n}{n} \to 0 \qquad (70)$$

*при* $n \to \infty$. *Тогда обобщенная оценка типа Фикс-Ходжеса* (59) - (60) *является асимптотически несмещенной оценкой плотности f(x).*

*Пусть, кроме того,*

$$\lim_{a\to\infty} \int_0^a K^2(u)du < \infty. \qquad (71)$$

*Тогда*

$$\lim_{n\to\infty} Df_n(x) = 0 \qquad (72)$$

*и $f_n(x)$ из* (*59*) - (*60*) *является состоятельной оценкой плотности $f(x)$.*

*Доказательство* вытекает фактически из того, что в силу приведенных выше рассуждений $h_n(x,\omega) \to 0$ (по вероятности) при $n \to \infty$. Можно фиксировать последовательность $\{h_n,\ n = 1, 2, ...\}$ и при этом условии повторить рассуждения, проведенные при доказательстве соответствующих теорем для ядерных оценок [9], поскольку эти рассуждения - аналитические, а не вероятностные. Для получения (72) необходимо учесть зависимость слагаемых в (59), что делается стандартным образом.

## 8. Непараметрические оценки регрессии

Начнем с рассмотрения условных плотностей. Пусть пространство $Z$ есть прямое произведение двух пространств $Z_1$ и $Z_2$, т.е. $Z = Z_1 \times Z_2$, а мера $p$ в $Z$ есть прямое произведение мер $p_1$ в $Z_1$ и $p_2$ в $Z_2$, т.е. $p = p_1 \times p_2$. Тогда элемент $x \in Z$ будем записывать в виде $x = (x_1, x_2)$, где $x_1 \in Z_1$ и $x_2 \in Z_2$. Пусть $X = (X_1, X_2)$ - случайная величина со значениями в Z, где $X_1 \in Z_1$ и $X_2 \in Z_2$, а $(X_{11}, X_{12})$, $(X_{21}, X_{22})$, ..., $(X_{n1}, X_{n2})$ - выборка из распределения, соответствующего $X = (X_1, X_2)$.

Как известно [26, с.145-146], условная плотность распределения $X_1$ при фиксированном значении $X_2 = x_2$ имеет вид

$$f(x_1 \mid X_2 = x_2) = f(x_1 \mid x_2) = \frac{f(x_1, x_2)}{\int\limits_{Z_1} f(x_1, x_2)\, p_1(dx_1)}, \quad x_1 \in Z_1, x_2 \in Z_2, \qquad (73)$$

где $f(x_1, x_2)$ - плотность случайного элемента $(X_1, X_2)$ в пространстве $Z = Z_1 \times Z_2$, а знаменатель в (73) отличен от 0.

Для оценки условной плотности (73) представляется естественным заменить совместную плотность $f(x_1, x_2)$ в (73) на ее непараметрическую оценку, т.е. в качестве оценки условной плотности $f(x_1|x_2)$ применять

$$f_n(x_1 \mid x_2) = \frac{f_n(x_1, x_2)}{\int\limits_{Z_1} f_n(x_1, x_2)\, p_1(dx_1)}, \qquad (74)$$

где $f_n(x_1, x_2)$ - оценка совместной плотности $f(x_1, x_2)$.

В [7, 9, 10] и выше в настоящей статье приведен ряд условий, при которых различные оценки плотности вероятности являются состоятельными, т.е. при $n \to \infty$

$$f_n(x_1, x_2) \to f(x_1, x_2) \qquad (75)$$

(сходимость по вероятности). Когда оценка условной плотности является состоятельной? Из (73) и (74) ясно, что при справедливости (75) для состоятельности $f_n(x_1|x_2)$ достаточно, чтобы при $n \to \infty$ по вероятности

$$\alpha = \int\limits_{Z_1} (f_n(x_1, x_2) - f(x_1, x_2))\, p_1(dx_1) \to 0. \qquad (76)$$

***Теорема 7.*** Пусть $p_1(Z_1) < \infty$,

$$\limsup_{n\to\infty} \sup_{x_1 \in Z_1} | Mf_n(x_1, x_2) - f(x_1, x_2) | = 0, \qquad (77)$$

$$\limsup_{n\to\infty} \sup_{x_1 \in Z_1} Df_n(x_1, x_2) = 0. \qquad (78)$$

*Тогда выполнено* (*76*).

*Доказательство.* Имеем

$$\alpha = \int\limits_{Z_1} (f_n(x_1, x_2) - Mf_n(x_1, x_2))\, p_1(dx_1) + \int\limits_{Z_1} (Mf_n(x_1, x_2) - f(x_1, x_2))\, p_1(dx_1). \qquad (79)$$

В силу (77) и условия $p_1(Z_1) < \infty$ второе слагаемое в (79) стремится к 0 при $n \to \infty$. Вычислим дисперсию $\alpha$. Положим

$$t(x_1) = f_n(x_1, x_2) - M(f_n(x_1, x_2)). \qquad (80)$$

Воспользуемся соотношением

$$\left[\int\limits_{Z_1} t(x_1)\, p_1(dx_1)\right]^2 = \int\limits_{Z_1} t(x_1)\, p_1(dx_1) \int\limits_{Z_1} t(z)\, p_1(dz) = \int\limits_{Z_1 \times Z_1} t(x_1) t(z)\, p_1 \times p_1(dx_1 \times dz). \qquad (81)$$

Имеем в силу теоремы Фубини

$$D\alpha = M\left(\int\limits_{Z_1} t(x_1)\, p_1(dx_1)\right)^2 = \int\limits_{Z_1 \times Z_1} Mt(x_1) t(z)\, p_1 \times p_1(dx_1 \times dz). \qquad (82)$$

По неравенству Коши-Буняковского

$$|Mt(x_1)t(z)| \le \left\{Df_n(x_1,x_2)Df_n(z,x_2)^{1/2}\right\}. \quad (83)$$

Из (76) и (83) следует, что при $n \to \infty$

$$D\alpha \to 0. \quad (84)$$

Из неравенства Чебышёва вытекает, что первое слагаемое в (79) также стремится к 0 при $n \to \infty$ (по вероятности). Следовательно, имеет место соотношение (76), теорема 7 доказана.

*Замечание.* Соотношения (77) и (78) доказаны для ядерных оценок в [9, 10]. Справедливость этих соотношений для гистограммных оценок вытекает из теорем 2 - 4 раздела 6 "Гистограммные оценки" настоящей статьи. Общие формулировки для линейных оценок вытекают из рассмотрений [7] в предположении, что используемые в неравенствах оценки являются равномерными.

Перейдем к рассмотрению регрессии, т.е. условного математического ожидания. Пусть $h(x, a)$ - мера близости на $Z_1$. В соответствии с оптимизационным подходом к определению средних величин [4, 27, 28] условным математическим ожиданием (регрессией $X_1$ на $X_2$) называется решение оптимизационной задачи

$$M\left(X_1 \mid X_2 = x_2,\ h\right) = Arg\min_{a \in Z_1} \int_{Z_1} h(x_1,a)\, f(x_1 \mid x_2)\, p_1(dx_1). \quad (85)$$

Подставив в (85) вместо плотности ее непараметрическую оценку, получим эмпирическую регрессию (непараметрическую оценку регрессии)

$$M_n\left(X_1 \mid X_2 = x_2,\ h\right) = Arg\min_{a \in Z_1} \int_{Z_1} h(x_1,a)\, f_n(x_1 \mid x_2)\, p_1(dx_1). \quad (86)$$

(В (85) и (86) предполагаем, что знаменатель в (73) отличен от 0.)

Установим, что при $n \to \infty$ эмпирическая регрессия (86) сходится к теоретической регрессии (85):

$$M_n\left(X_1 \mid X_2 = x_2,\ h\right) \to M\left(X_1 \mid X_2 = x_2,\ h\right) \quad (87)$$

(сходимость понимается так, как при формулировке законов больших чисел [28] и изучении асимптотического поведения решения экстремальных статистических задач [4, 29], поскольку в (85) и (86) определены, вообще говоря, не элементы, а множества).

Как следует из предельной теории решений экстремальных статистических задач, для доказательства состоятельности эмпирической регрессии как оценки теоретической, т.е. для доказательства (87), достаточно установить равностепенную непрерывность на $Z_1$ функций

$$q_n(a) = \int_{Z_1} h(x_1, a)\, f_n(x_1 \mid x_2)\, p_1(dx_1) \qquad (88)$$

и то, что при любом $a \in Z_1$

$$q_n(a) \to q(a) = \int_{Z_1} h(x_1, a)\, f(x_1 \mid x_2)\, p_1(dx_1) \qquad (89)$$

по вероятности при $n \to \infty$.

***Теорема 8.*** *Пусть* $Z_1$ - *бикомпакт*, $h(x, a)$ - *непрерывная функция на* $Z_1 \times Z_1$ *и* $f_n(x_1, x_2) \geq 0$ *с вероятностью* 1. *Тогда последовательность функций* $q_n(a)$ *равностепенно непрерывна на* $Z_1$.

*Доказательство*. Из условия теоремы 8 следует, что для любого $\varepsilon > 0$ существует разбиение пространства $Z_1$ такое, что для любых точек $a$ и $a'$ из одного и того же элемента этого разбиения имеем

$$\sup_{x_1 \in Z_1} | h(x_1, a) - h(x_1, a') | < \varepsilon, \qquad (90)$$

а тогда

$$| q_n(a) - q_n(a') | = \left| \int_{Z_1} (h(x_1, a) - h(x_1, a'))\, f_n(x_1 \mid x_2)\, p_1(dx_1) \right| \leq \varepsilon \int_{Z_1} f_n(x_1 \mid x_2)\, p_1(dx_1) = \varepsilon, \qquad (91)$$

что и доказывает теорему 8.

**Теорема 9.** *Пусть выполнены условия теоремы* 7, *измеримая мера близости* $h(x, a)$ *ограничена, знаменатель в* (73) *отличен от* 0. *Тогда справедливо* (89).

*Доказательство*. Достаточно получить (89) для

$$s_n(a) = q_n(a) \int_{Z_1} f_n(x_1, x_2)\, p_1(dx_1) = \int_{Z_1} h(x_1, a)\, f_n(x_1, x_2)\, p_1(dx_1)\,. \qquad (92)$$

Доказательство проводится как в теореме 7, отличие только в том, что в аналогах (79) и (82) добавляются множители $h(x_1, a)$ и $h(x_1, a)\, h(z, a)$ соответственно, являющиеся неотрицательными (т.к. $h(x_1, a)$ - мера близости) и ограниченными (в силу условия теоремы).

Соединив условия теорем 8 и 9, получим условия сходимости эмпирической регрессии к теоретической.

***Теорема 10.*** *Пусть* $Z_1$ - *бикомпакт,* $h(x, a)$ - *непрерывная неотрицательная функция на* $Z_1 \times Z_1$, *для неотрицательной с вероятностью* 1 *оценки плотности* $f_n(x_1, x_2)$ *выполнены условия теоремы* 7 *для всех* $x_2 \in Z_2$ *и знаменатель в* (73) *отличен от* 0 *для всех* $x_2 \in Z_2$. *Тогда для всех* $x_2 \in Z_2$ *справедливо* (87) (*в смысле указанных ранее работ* [4, 28, 29]).

*Замечание* 1. Условия теоремы 10 могут быть ослаблены. Так, из доказательства теоремы 8 видно, что условие неотрицательности $f_n(x_1, x_2)$ может быть заменено на условие

$$\overline{\lim_{n \to \infty}} \int_{Z_1} | f_n(x_1 \mid x_2) |\, p_1(dx_1) < \infty\,. \qquad (93)$$

Это существенно, в частности, для ядерных оценок, поскольку согласно результатам раздела 1 настоящей статьи отказ от неотрицательности ядерных оценок плотности позволяет ускорить их сходимость.

*Замечание* 2. Теоремы 7 - 10 основаны на равномерной сходимости моментов (77) - (78). Другой ряд теорем с аналогичными заключениями может быть получен в предположении равномерной сходимости оценок плотности. Из равномерной сходимости сходимость моментов, вообще говоря, не следует, хотя для гистограммных оценок имеет место и то, и другое.

**9. Дискриминантный анализ в пространстве общей природы**

Рассмотрим постановку с двумя классами, заданными плотностями $f(x)$ и $g(x)$ соответственно, $x \in Z$. Если $f$ и $g$ известны, то по лемме Неймана-Пирсона область отнесения к первому классу задается неравенством

$$\frac{f(x)}{g(x)} > A \qquad (94)$$

где $A$ - некоторая константа (ср. [30]). Если плотности $f$ и $g$ неизвестны, но имеются их состоятельные оценки $f_n(x)$ и $g_m(x)$ по обучающим выборкам объемов $n$ и $m$ соответственно, то область

$$\frac{f_n(x)}{g_m(x)} > A \quad (n \to \infty, m \to \infty) \qquad (95)$$

является состоятельной (в соответствующем смысле) оценкой теоретической области (94).

В случае $k$ классов известные постановки ([30], [31]) задачи минимизации средних потерь от принятия ошибочных решений приводят к решениям в терминах плотностей, описывающих классы, априорных вероятностей классов и функции потерь. При этом результаты наблюдений могут лежать в произвольном пространстве (последнее обстоятельство обычно не осознается авторами публикаций по этой тематике), поэтому нам нет необходимости развивать самостоятельную теорию (впрочем, в [2, с.221-223] теория расписана для конечных случайных множеств). При использовании обучающих выборок теоретические плотности заменяются их непараметрическими оценками, рассмотренными выше. Из-за отсутствия принципиально новых моментов развернутую теорию дискриминантного анализа в пространствах общей природы здесь не приводим, ограничившись данными выше замечаниями.

## Литература

1. Орлов А.И. О новой парадигме прикладной математической статистики // Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. / Перм. гос. нац. иссл. ун-т. – Пермь, 2013. Вып. 25. С.162-176.

2. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.

3. Орлов А.И. Статистика объектов нечисловой природы и экспертные оценки // Экспертные оценки / Вопросы кибернетики. Вып.58. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1979. С.17-33.

4. Орлов А.И. Организационно-экономическое моделирование: учебник : в 3 ч. Часть 1: Нечисловая статистика. – М.: Изд-во МГТУ им. Н.Э. Баумана. 2009. – 541 с.

5. Орлов А.И. О развитии статистики объектов нечисловой природы // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2013. № 93. С. 41-50.

6. Орлов А.И. Оценки плотности в пространствах произвольной природы // Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. / Перм. гос. нац. иссл. ун-т. – Пермь, 2013. Вып. 25. С.21-33.

7. Орлов А.И. Оценки плотности распределения вероятностей в пространствах произвольной природы // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 99. С. 15-32.

8. Орлов А.И. Ядерные оценки плотности в пространствах произвольной природы // Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. / Перм. гос. нац. иссл. ун-т. – Пермь, 2015. Вып. 26. С. 43-57.

9. Орлов А.И. Предельные теоремы для ядерных оценок плотности в пространствах произвольной природы // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 108. С. 316 – 333.

10. Орлов А.И. Непараметрические ядерные оценки плотности вероятности в дискретных пространствах // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2016. № 122. С. 833 –855.

11. Орлов А.И. Ядерные оценки плотности в конечных пространствах // Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. / Перм. гос. нац. иссл. ун-т. – Пермь, 2016. – Вып. 27. – С. 24-37.

12. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. – М.: Большая Российская Энциклопедия, 1999. – 910 с.

13. Орлов А.И. Статистика объектов нечисловой природы // Теория вероятностей и ее применения. 1980. Т.XXV. № 3. С. 655-656.

14. Орлов А.И. Непараметрические оценки плотности в топологических пространствах // Прикладная статистика. Ученые записки по статистике, т.45. – М.: Наука, 1983. – С. 12-40.

15. Rosenblatt M. Remarks on some nonparametric estimates of a density function // Ann. Math. Statist. 1956. V.27. N 5. P. 832 – 837.

16. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statist. 1962. V.33. N 6. P. 1065-1076.

17. Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. – М.: Наука, 1979. – 528 с.

18. Орлов А.И. О развитии методологии статистических методов // Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. / Перм. гос. нац. иссл. ун-т. – Пермь, 2001. – Вып. 15. – С.118-131.

19. Орлов А.И. О влиянии методологии на последствия принятия решений // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2017. № 125. С. 319 – 345.

20. Смирнов Н.В. О приближении плотностей распределения случайных величин // Ученые записки МГПИ им. В.П. Потемкина. 1951. Т. XVI. Вып.3. С. 69-96.

21. Надарая Э.А. К построению доверительных областей для плотности вероятности // Сообщения АН ГрузССР. 1970. Т.59. № 1. С.33-36.

22. Надарая Э.А. О построению доверительных областей для плотности вероятности // Аннотации докладов семинара Института прикладной математики Тбилисского государственного университета. 1972. № 5. С. 27-32.

23. Мания Г.М. Статистическое оценивание распределения вероятностей. - Тбилиси: Издательство Тбилисского университета, 1974. - 240 с.

24. Конаков В.Д. Полные асимптотические разложения для максимального уклонения эмпирической функции плотности // Теория вероятностей и её применения. 1978. Т. XXIII. №3. С. 495-509.

25. Боровков А.А. Теория вероятностей. - М.: Наука, 1976. - 352 с.

26. Прохоров Ю,В., Розанов Ю.А. Теория вероятностей: Основные понятия. Предельные теоремы. Случайные процессы / Справочная математическая библиотека. - М.: Наука, 1973. - 496 с.

27. Орлов А.И. Прикладная статистика. Учебник для вузов. — М.: Экзамен, 2006. — 672 с.

28. Орлов А.И. Средние величины и законы больших чисел в пространствах произвольной природы // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2013. № 89. С. 556 – 586.

29. Орлов А.И. Асимптотика решений экстремальных статистических задач // Анализ нечисловых данных в системных исследованиях. Сборник трудов. Вып.10. - М.: Всесоюзный научно-исследовательский институт системных исследований, 1982. С. 4-12.

30. Орлов А.И. Математические методы теории классификации // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 95. С. 23 – 45.

31. Орлов А.И. Прогностическая сила – наилучший показатель качества алгоритма диагностики // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 99. С. 33-–49.

## References

1. Orlov A.I. O novoj paradigme prikladnoj matematicheskoj statistiki // Statisticheskie metody ocenivanija i proverki gipotez: mezhvuz. sb. nauch. tr. / Perm. gos. nac. issl. un-t. – Perm', 2013. Vyp. 25. S.162-176.

2. Orlov A.I. Ustojchivost' v social'no-jekonomicheskih modeljah. - M.: Nauka, 1979. - 296 s.

3. Orlov A.I. Statistika ob#ektov nechislovoj prirody i jekspertnye ocenki // Jekspertnye ocenki / Voprosy kibernetiki. Vyp.58. - M.: Nauchnyj Sovet AN SSSR po kompleksnoj probleme "Kibernetika", 1979. S.17-33.

4. Orlov A.I. Organizacionno-jekonomicheskoe modelirovanie: uchebnik : v 3 ch. Chast' 1: Nechislovaja statistika. – M.: Izd-vo MGTU im. N.Je. Baumana. 2009. – 541 s.

5. Orlov A.I. O razvitii statistiki ob#ektov nechislovoj prirody // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2013. № 93. S. 41-50.

6. Orlov A.I. Ocenki plotnosti v prostranstvah proizvol'noj prirody // Statisticheskie metody ocenivanija i proverki gipotez: mezhvuz. sb. nauch. tr. / Perm. gos. nac. issl. un-t. – Perm', 2013. Vyp. 25. S.21-33.

7. Orlov A.I. Ocenki plotnosti raspredelenija verojatnostej v prostranstvah proizvol'noj prirody // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2014. № 99. S. 15-32.

8. Orlov A.I. Jadernye ocenki plotnosti v prostranstvah proizvol'noj prirody // Statisticheskie metody ocenivanija i proverki gipotez: mezhvuz. sb. nauch. tr. / Perm. gos. nac. issl. un-t. – Perm', 2015. Vyp. 26. S. 43-57.

9. Orlov A.I. Predel'nye teoremy dlja jadernyh ocenok plotnosti v prostranstvah proizvol'noj prirody // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2015. № 108. S. 316 – 333.

10. Orlov A.I. Neparametricheskie jadernye ocenki plotnosti verojatnosti v diskretnyh prostranstvah // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2016. № 122. S. 833 –855.

11. Orlov A.I. Jadernye ocenki plotnosti v konechnyh prostranstvah // Statisticheskie metody ocenivanija i proverki gipotez: mezhvuz. sb. nauch. tr. / Perm. gos. nac. issl. un-t. – Perm', 2016. – Vyp. 27. – S. 24-37.

12. Verojatnost' i matematicheskaja statistika: Jenciklopedija / Gl. red. Ju.V. Prohorov. – M.: Bol'shaja Rossijskaja Jenciklopedija, 1999. – 910 s.

13. Orlov A.I. Statistika ob#ektov nechislovoj prirody // Teorija verojatnostej i ee primenenija. 1980. T.XXV. № 3. S. 655-656.

14. Orlov A.I. Neparametricheskie ocenki plotnosti v topologicheskih prostranstvah // Prikladnaja statistika. Uchenye zapiski po statistike, t.45. – M.: Nauka, 1983. – S. 12-40.

15. Rosenblatt M. Remarks on some nonparametric estimates of a density function // Ann. Math. Statist. 1956. V.27. N 5. P. 832 – 837.

16. Parzen E. On estimation of a probability density function and mode // Ann. Math. Statist. 1962. V.33. N 6. P. 1065-1076.

17. Ibragimov I.A., Has'minskij R.Z. Asimptoticheskaja teorija ocenivanija. – M.: Nauka, 1979. – 528 s.

18. Orlov A.I. O razvitii metodologii statisticheskih metodov // Statisticheskie metody ocenivanija i proverki gipotez: mezhvuz. sb. nauch. tr. / Perm. gos. nac. issl. un-t. – Perm', 2001. – Vyp. 15. – S.118-131.

19. Orlov A.I. O vlijanii metodologii na posledstvija prinjatija reshenij // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2017. № 125. S. 319 – 345.

20. Smirnov N.V. O priblizhenii plotnostej raspredelenija sluchajnyh velichin // Uchenye zapiski MGPI im. V.P. Potemkina. 1951. T. XVI. Vyp.3. S. 69-96.

21. Nadaraja Je.A. K postroeniju doveritel'nyh oblastej dlja plotnosti verojatnosti // Soobshhenija AN GruzSSR. 1970. T.59. № 1. S.33-36.

22. Nadaraja Je.A. O postroeniju doveritel'nyh oblastej dlja plotnosti verojatnosti // Annotacii dokladov seminara Instituta prikladnoj matematiki Tbilisskogo gosudarstvennogo universiteta. 1972. № 5. S. 27-32.

23. Manija G.M. Statisticheskoe ocenivanie raspredelenija verojatnostej. - Tbilisi: Izdatel'stvo Tbilisskogo universiteta, 1974. - 240 s.

24. Konakov V.D. Polnye asimptoticheskie razlozhenija dlja maksimal'nogo uklonenija jempiricheskoj funkcii plotnosti // Teorija verojatnostej i ejo primenenija. 1978. T. XXIII. №3. S. 495-509.

25. Borovkov A.A. Teorija verojatnostej. - M.: Nauka, 1976. - 352 s.

26. Prohorov Ju,V., Rozanov Ju.A. Teorija verojatnostej: Osnovnye ponjatija. Predel'nye teoremy. Sluchajnye processy / Spravochnaja matematicheskaja biblioteka. - M.: Nauka, 1973. - 496 s.

27. Orlov A.I. Prikladnaja statistika. Uchebnik dlja vuzov. — M.: Jekzamen, 2006. — 672 s.

28. Orlov A.I. Srednie velichiny i zakony bol'shih chisel v prostranstvah proizvol'noj prirody // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2013. № 89. S. 556 – 586.

29. Orlov A.I. Asimptotika reshenij jekstremal'nyh statisticheskih zadach // Analiz nechislovyh dannyh v sistemnyh issledovanijah. Sbornik trudov. Vyp.10. - M.: Vsesojuznyj nauchno-issledovatel'skij institut sistemnyh issledovanij, 1982. S. 4-12.

30. Orlov A.I. Matematicheskie metody teorii klassifikacii // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2014. № 95. S. 23 – 45.

31. Orlov A.I. Prognosticheskaja sila – nailuchshij pokazatel' kachestva algoritma diagnostiki // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2014. № 99. S. 33-–49.