УДК 519.2

Оценки плотности в пространствах произвольной природы

А.И. Орлов

Московский государственный технический университет им. Н.Э. Баумана, Институт высоких статистических технологий и эконометрики; Московский физико-технический институт; Москва, Россия;
prof-orlov@mail.ru; http://orlovs.pp.ru; +7(916)8305117; 123104, Москва, Сытинский пер., д.7/14, кв.14.

Аннотация. *Введены линейные оценки плотности распределения вероятностей в пространствах произвольной природы и их частные случаи – ядерные и гистограммные оценки, оценки типа Фикс-Ходжеса. Состоятельность и асимптотической нормальность линейных оценок доказана при выполнении естественных условий. Показано, что вероятность попадания в область может быть найдена с помощью линейных оценок плотности. Рассмотрен частный случай конечного множества, установлено, что выборочная мода сходится к теоретической.*
Ключевые слова: нечисловая статистика, плотность распределения вероятностей, пространства произвольной природы, линейные оценки плотности, предельные теоремы, состоятельные оценки, асимптотическая нормальность.

Оценки плотности распределения вероятностей в пространствах произвольной природы используют для решения различных задач нечисловой статистики [1], называемой также статистикой объектов нечисловой природы. Однако систематическое изложение теории таких оценок ранее не публиковалось. Настоящая статья посвящена частичному заполнению этого пробела.

1. Различные виды оценок плотности

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

$$q(C) = \int_C f(x)dp \qquad (1)$$

для любого множества C из сигма-алгебры измеримых множеств $\mathfrak{A}$. 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 [2, с.460].

Пусть $X_1, X_2, \ldots, X_n$ – независимые одинаково распределенные случайные элементы (величины), распределение которых задается вероятностной мерой q. В настоящей статье рассмотрим несколько видов непараметрических оценок плотности вероятности q по выборке $X_1, X_2, \ldots, X_n$. А именно, линейные оценки и их частные случаи – линейные и гистограммные, и оценки типа Фикс-Ходжеса, не являющиеся линейными.

Мера p предполагается заданной. В случае конечномерного евклидова пространства $Z = \mathbf{R}^k$ в качестве p обычно используют лебегову меру. Если пространство объектов нечисловой природы конечно, то в качестве p можно использовать меру, приписывающую каждому элементу x из Z единичный вес [1]. В качестве p можно применять распределение определенного случайного элемента со значениями в Z. В теории случайных процессов рассматривают плотности по гауссовским мерам [3, 4, 5].

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 [6, с.24]:

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

в которых действительнозначные функции g_n удовлетворяют некоторым условиям регулярности, обсуждаемым ниже.

Укажем несколько частных случаев оценок (2). Гистограммные оценки определяются с помощью последовательности T_n разбиений пространства Z на непересекающиеся области (элементы T_n) и функций

$$g_n(x, X_i) = \begin{cases} \dfrac{1}{p(A(x))}, & X_i \in A(x), \\ 0 & X_i \notin A(x), \end{cases} \tag{3}$$

где $A(x)$ – элемент разбиения T_n, которому принадлежит x. Первая работа по

непараметрическим оценкам плотности вероятности вида (2) принадлежит Н.В. Смирнову [7], изучившему оценки (2) – (3) с измельчающейся последовательностью разбиений T_n, для которых максимальный (по x) диаметр областей $A(x)$ стремится к 0.

Проекционные оценки получаются при разложении плотности в ряд по базисным функциям и рассмотрении в качестве оценки плотности конечного отрезка этого ряда с заменой коэффициентов на их оценки [8]. Теория проекционных оценок для пространств произвольной природы развита Н.Н. Ченцовым [9, разд.25]. Однако для построения таких оценок нужен ортонормальный базис в пространстве функций на Z, а для конкретных пространств объектов нечисловой природы методы построения подобных базисов, пригодные для проведения расчетов, обычно не разработаны. Поэтому мы вынуждены разрабатывать другие виды непараметрических оценок плотности.

Пусть d – показатель различия на Z [1] (в наиболее важных частных случаях – метрика на Z). В [10] введены ядерные оценки плотности – оценки вида (2) с

$$g_n(x, X_i) = \frac{1}{b(h_n, x)} K\left(\frac{d(x, X_i}{h_n}\right), \qquad K : [0, +\infty) \to \mathbf{R}^1, \tag{4}$$

где $K = K(u)$ – ядро (ядерная функция), h_n – последовательность положительных чисел (показателей размытости), $b(h_n, x)$ – нормировочный множитель. В [8] линейные оценки (2) с функциями g_n из (4) названы "обобщенными оценками типа Парзена-Розенблатта", т.к. в частном случае $Z = \mathbf{R}^1, d(x, X_i) = |x - X_i|, b(h_n, x) = h_n$ они переходят в известные оценки, введенные Розенблаттом [11] и Парзеном [12], которым посвящены сотни работ.

Естественный класс оценок плотности, не являющихся линейными, был предложен в частном случае конечномерного пространства Фикс и Ходжесом [13]. Эти оценки использовались прежде всего в задачах классификации (дискриминантного анализа, диагностики) и известны как оценки "методом k_n ближайших соседей" (см., например, [14, разд. 6.2], [15, разд. 4.4]). Выбирается шар с центром в точке x, имеющий минимальный радиус среди всех шаров, содержащих k_n элементов выборки. Пусть V_n – объем этого шара (ясно, что V_n – случайная величина). В качестве оценки плотности используют случайную величину $f_n(x) = k_n/V_n$.

Для произвольных пространств Z объектов нечисловой природы обобщенная оценка типа Фикс-Ходжеса определена нами в [8] с помощью связанных с точкой x пространства Z системы расширяющихся множеств $U(x, r), r \geq 0$, такой, что $U(x, r_1)$ является частью $U(x, r_2)$ при $r_1 < r_2$, а объединение всех $U(x, r)$ при $r \geq 0$ совпадает с Z. Пусть r^* – точная нижняя грань r таких, что $U(x, r)$ содержит не менее k_n элементов выборки, тогда обобщенной оценкой типа Фикс-Ходжеса называется $f_n(x) = k_n/p(U(x, r^*))$.

Если Z является метрическим пространством с метрикой d или же пространством с показателем различия d, то естественно использовать $U(x,r) = \{y : d(x,y) \le r\}$.

Есть и иные методы оценки плотности случайной величины. Так, в [16] предложено находить оценку как решение экстремальной статистической задачи. По существу речь идет о том, чтобы оптимально оценить число слагаемых в частном случае проекционных оценок Ченцова, однако, ссылки на работы Н.Н. Ченцова отсутствуют. Оценки находятся лишь численно. В [17, 18] предложено использовать аналог проекционных оценок для квадратного корня из плотности вероятности.

Рассмотрим частный случай $Z = \mathbf{R}^1, d(x, X_i) = |x - X_i|, b(h_n, x) = h_n$. Известно, что среди ядерных оценок вида (4) можно найти сходящиеся с наилучшей возможной по порядку величины скоростью [19, с.321]. Аналогичный результат верен и для проекционных оценок Ченцова [9]. В [8] нами найдены главные члены среднего квадрата ошибки $\mathbf{M}\left[f_n(x) - f(x)\right]^2$ для оценки (4) типа Парзена-Розенблатта с ядерной функцией $K(u) = 0,5$ при $|u| \le 1$ и $K(u) = 0$ при $|u| > 1$ (согласно [20, с.96]) и для оценки Фикс-Ходжеса, вычисленные нами на основе [21]. Оптимальный порядок скорости сходимости для обеих оценок одинаков и достигается при $k_n = nh_n = n^{4/5}$ (отметим, что вопреки мнению [14, с.188] следует выбирать k_n достаточно большим). При этом множители перед степенями k_n и h_n в формулах для средних квадратов ошибок являются функциями от плотности и ее второй производной, причем сравнить эти множители в общем случае не представляется возможным: результат сравнения зависит от конкретного вида указанных функций.

Из сказанного с учетом результатов работ [22, 23] вытекает, что в классическом случае $Z = \mathbf{R}^k$ нет оснований установить, какими из различных видов непараметрических оценок плотности следует пользоваться. Поэтому в статистике объектов нечисловой природы целесообразно проработать возможность использования оценок плотности различных типов. При этом выделяются линейные оценки, поскольку они согласно (2) являются суммами случайных функций, независимых и одинаково распределенных в силу того, что $X_1, X_2, \ldots, X_n$ – выборка. Их легко реализовать численно. Среди конкретных видов линейных оценок выделяются ядерные оценки [24], поскольку разработаны аксиоматические подходы к выбору метрики в пространствах объектов нечисловой природы [1]. Ядерные оценки выгодно отличаются от гистограммных отсутствием произвола при выборе разбиений T_n. Ядерные оценки при фиксированной метрике (показателя различия) d имеют конкретный вид с точностью до ядерной функции $K(u)$ и последо-

вательности h_n показателей размытости, как и в классическом случае.

Будем рассматривать сходимость по вероятности. Перенос результатов на случай сходимости с вероятностью 1 обычно не вызывает трудностей.

2. Линейные оценки

Положим $Y_{in} = Y_{in}(x) = g_n(x, X_i)$, тогда согласно (2)

$$f_n(x) = \frac{1}{n} \sum_{1 \le i \le n} Y_{in}. \tag{5}$$

Поскольку случайные величины Y_{in} независимы и одинаково распределены, то согласно (5) для состоятельности и асимптотической нормальности $f_n(x)$ необходимо и достаточно, чтобы при безграничном росте объема выборки n были выполнены предельные соотношения

$$\mathbf{M} f_n(x) = \mathbf{M} Y_{in} \to f(x), \qquad \mathbf{D} f_n(x) = \frac{\mathbf{D} Y_{in}}{n} \to 0. \tag{6}$$

Укажем естественные условия, при которых справедливы соотношения (6). Поскольку

$$\mathbf{M} f_n(x) = \mathbf{M} g_n(x, X_1) = \int\limits_Z g_n(x, y) f(y) p(dy), \tag{7}$$

то для существования математического ожидания $\mathbf{M} f_n(x)$ достаточно, чтобы были выполнены следующие условия:

$$(I) \qquad c_{1n} = \int\limits_Z |g_n(x, y)| p(dy) < \infty,$$

$$(II) \qquad c_2 = \sup_{x \in Z} f(x) < \infty.$$

Выполнение условия (I) можно обеспечить путем выбора g_n, в то время как условие (II) наложено на неизвестную плотность f.

Нам понадобится условие нормировки

$$(III) \qquad \int\limits_Z g_n(x, y) p(dy) = 1.$$

Если $g_n(x, y) = g_n(y, x)$, то условие (III) вытекает из естественного требования того, чтобы функция $f_n(x)$ была плотностью, т.е.

$$\int\limits_Z f_n(x) p(dx) = \int\limits_Z g_n(x, X_1) p(dx) = 1. \tag{8}$$

Из соотношения (7) и условия (III) следует, что

$$\mathbf{M}f_n(x) - f(x) = \int\limits_Z g_n(x,y)(f(y) - f(x))p(dy). \tag{9}$$

Для того, чтобы изучить интеграл в правой части (9), разобьем его на два – по окрестности $U(x)$ точки x и по ее внешности $Z \setminus U(x)$. Чтобы такое разбиение позволило получить полезные выводы, введем условие (IV).

(IV). Функция f непрерывна в точке x.

Возьмем произвольное число $a > 0$. В силу условия (IV) существует окрестность $U(x)$ точки x такая, что

$$|f(x) - f(y)| < a \tag{10}$$

для всех точек y из окрестности $U(x)$ точки x.

Замечание. Вплоть до условия (IV) пространство $(Z, \mathfrak{A})$ рассматривалось как измеримое. В условии (IV) появилось понятие непрерывности, т.е. предположение, что Z – топологическое пространство. Будем считать, что измеримая и топологическая структуры пространства Z согласованы между собой, т.е. открытые множества измеримы. Для Z из конечного числа элементов, представляющих основной интерес в нечисловой статистике [1], это условие выполнено тривиально. Согласно (9) имеем

$$\mathbf{M}f_n(x) - f(x) = \tag{11}$$
$$= \int\limits_{U(x)} g_n(x,y)(f(y) - f(x))p(dy) + \int\limits_{Z\setminus U(x)} g_n(x,y)(f(y) - f(x))p(dy).$$

Каждое из слагаемых в правой части (11) рассмотрим по отдельности. Для первого из них справедлива цепочка неравенств:

$$\left| \int\limits_{U(x)} g_n(x,y)[f(y) - f(x)]p(dy) \right| \le \int\limits_{U(x)} |g_n(x,y)||f(y) - f(x)|p(dy) \le$$
$$\le a \int\limits_{U(x)} |g_n(x,y)|p(dy) \le a \int\limits_Z |g_n(x,y)|p(dy) = ac_{1n}. \tag{12}$$

Чтобы гарантировать, что первое слагаемое в (11) стремится к 0, когда a стремится к 0, добавим новое условие:

$$(V) \qquad \sup_n c_{1n} = c_1(x) < \infty$$

(отметим, что в условии (I) $c_{1n} = c_{1n}(x)$). Тогда

$$\left| \int\limits_{U(x)} g_n(x,y)(f(y) - f(x))p(dy) \right| \le a\, c_1(x). \tag{13}$$

Для второго неравенства в (11) с учетом неравенства

$$|f(y) - f(x)| \le \sup\{f(z), z \in Z\}$$

справедлива оценка

$$\left| \int\limits_{Z\setminus U(x)} g_n(x,y)(f(y) - f(x))p(dy) \right| \le c_2 \int\limits_{Z\setminus U(x)} |g_n(x,y)|p(dy). \tag{14}$$

Для того, чтобы правая часть неравенства (14) стремилась к 0 при безграничном n, введем условие (VI).

(VI). Для любой окрестности $U(x)$ точки x

$$\lim_{n\to\infty} \int\limits_{Z\setminus U(x)} |g_n(x,y)|p(dy) = 0.$$

Теорема 1. *Если условия (I) – (VI) выполнены, то*

$$\lim_{n\to\infty} \mathbf{M} f_n(x) = f(x). \tag{15}$$

Доказательство. Рассмотрим малое число $b > 0$. Положим $a = b/(2c_1(x))$. Рассмотрим окрестность $U(x) = U(x,b)$ такую, что неравенство (10) выполнено для этого a. Тогда правая часть неравенства (13) равна $b/2$. Из условия (VI) следует, что существует число $n_0 = n_0(x,b)$ такое, что

$$\int\limits_{Z\setminus U(x)} |g_n(x,y)|p(dy) < \frac{b}{2c_2} \tag{16}$$

при $n > n_0$, следовательно, правая часть неравенства (14) меньше $b/2$. Из равенства (11) следует, что $|\mathbf{M} f_n(x) - f(x)| < b$ при $n > n_0$, следовательно, соотношение (15) выполнено.

Замечание. При доказательстве теоремы 1 использовалось только равенство (7), т.е. одинаковая распределенность элементов выборки X_i – их независимость не требовалась.

Если случайные величины $Y_{in} = g_n(x, X_i)$ некоррелированы и имеют дисперсию, то

$$\mathbf{D}f_n(x) = \frac{\mathbf{D}g_n(x, X_1)}{n} = \frac{1}{n}\left\{\int\limits_Z g_n^2(x,y)f(y)p(dy) - [\mathbf{M}f_n(x)]^2\right\}. \qquad (17)$$

Для существования дисперсии $\mathbf{D}f_n(x)$ достаточно предположить, что выполнено условие

$$(VII) \qquad d_n = d_n(x) = \int\limits_Z g_n^2(x,y)p(dy) < \infty,$$

учитывая условия (I), (II) и равенство (17). Напрашивающееся условие ограниченности последовательности d_n является слишком жестким – ему не удовлетворяют ядерные оценки (4).

Теорема 2. *Пусть случайные величины $Y_{in} = g_n(x, X_i), i = 1, 2, , n$, независимы и одинаково распределены, выполнены условия (I) – (VII) и*

$$\lim_{n\to\infty} \frac{\mathbf{D}g_n(x, X_1)}{n} = 0, \qquad \mathbf{D}g_n(x, X_1) \neq 0. \qquad (18)$$

Тогда $f_n(x)$ – состоятельная и асимптотически нормальная оценка плотности f в точке x.

Доказательство. Из теоремы 1, соотношений (17) и (18) следует, что средний квадрат ошибки $\mathbf{M}\left[f_n(x) - f(x)\right]^2$ стремится к 0 при безграничном росте объема выборки n, и с помощью неравенства Чебышёва получаем состоятельность. Асимптотическая нормальность следует из Центральной предельной теоремы (следствие на с.255 [25]), поскольку случайные величины Y_{in} независимы, одинаково распределены и имеют ненулевую дисперсию.

Замечание 1. Для проекционных оценок условие (VI) отражают плотность "в целом", а оценки, удовлетворяющие условие (VI), – локально.

Замечание 2. Условия (I) – (VII) проверяют для конкретных видов оценок.

Получим аналог равенства (1), определяющего понятие плотности, для оценок плотности $f_n(x)$. Для любого события A, любого малого числа $\varepsilon > 0$ и любого натурального числа $n = 1, 2, \ldots$, положим

$$\Gamma p(A|\varepsilon, n) = \qquad (19)$$
$$= \{x \in A : \int\limits_{Z\setminus A} |g_n(y,x)|\, p(dy) > \varepsilon\} \cup \{x \in Z \setminus A : \int\limits_A |g_n(y,x)|\, p(dy) > \varepsilon\}.$$

Содержательный смысл $\Gamma p(A|\varepsilon, n)$ – окрестность границы множества A, заданная в терминах g_n.

Теорема 3. *Пусть выполнены условия (III), (V) равномерно для всех x из Z и*

$$\lim_{n\to\infty} \mathbf{P}\left\{X_1 \in \Gamma p(A|\varepsilon, n)\right\} = 0 \tag{20}$$

для любого $\varepsilon > 0$. Тогда по вероятности

$$\lim_{n\to\infty} \int\limits_A f_n(x)p(dx) = \mathbf{P}\{X_1 \in A\}. \tag{21}$$

Доказательство. Выборку $X_1, X_2, \ldots, X_n$ разобьем на три части: H_1 – совокупность тех элементов выборки, которые входят во внутреннюю часть A, т.е. в $A \setminus \Gamma p(A|\varepsilon, n)$, H_2 – множество результатов наблюдений, попавших в $\Gamma p(A|\varepsilon, n)$, и H_3 – множество результатов наблюдений, лежащих в остальной части Z, т.е. в дополнении к объединению A и $\Gamma p(A|\varepsilon, n)$. Тогда сумма, задающая линейную оценку плотности согласно (2), разбивается на три суммы в соответствии с делением выборки на три части H_1, H_2, H_3:

$$f_n(x) = \frac{1}{n}\left(\sum_{X_i\in H_1} g_n(x, X_i) + \sum_{X_i\in H_2} g_n(x, X_i) + \sum_{X_i\in H_3} g_n(x, X_i)\right). \tag{22}$$

Для X_i из H_1 в силу (19) и условия (III)

$$\left|\int\limits_A g_n(x, X_i)p(dx) - 1\right| < \varepsilon. \tag{23}$$

Аналогично для X_i из H_3 по тем же причинам

$$\left|\int\limits_A g_n(x, X_i)p(dx)\right| < \varepsilon. \tag{24}$$

Наконец, для X_i из H_2 в силу условия (V) (а также условия (I))

$$\left|\int\limits_A g_n(x, X_i)p(dx)\right| \le \left|\int\limits_Z |g_n(x, X_i)|p(dx)\right| \le c_1. \tag{25}$$

Из последних четырех формул (21) – (25) следует, что

$$\left|\int\limits_A f_n(x)p(dx) - \frac{|H_1|}{n}\right| \le \varepsilon\left(\frac{|H_1| + |H_3|}{n}\right) + c_1\frac{|H_2|}{n}, \tag{26}$$

где $|H_i|$ обозначает число элементов множества H_i, $i = 1, 2, 3$.

Первое слагаемое в правой части неравенства (26) не превосходит ε. Рассмотрим второе слагаемое. Случайная величина $|H_2|$ является числом успехов в n испытаниях Бернулли с вероятностью успеха p в каждом испытании, где p есть вероятность попадания случайной величины (элемента) X_1 в $\Gamma p(A|\varepsilon, n)$. Из соотношения (20) и неравенства Чебышёва следует, что второе слагаемое в правой части неравенства (26) стремится к 0 при безграничном росте объема выборки n.

Рассмотрим левую часть неравенства (26). Случайная величина $|H_1|$ является числом успехов в n испытаниях Бернулли с вероятностью успеха p в каждом испытании, где p есть вероятность попадания случайной величины (элемента) X_1 во внутренность множества A, т.е. в $A \setminus \Gamma p(A|\varepsilon, n)$. В силу соотношения (20) эта вероятность успеха при безграничном росте объема выборки n стремится к вероятности попадания случайной величины X_1 в множество A. Из неравенства (26) и последних утверждений вытекает соотношение (21). Теорема 3 доказана.

Обсудим сходимость выборочной моды к теоретической. Поскольку выборочная мода есть $\arg\max f_n(x)$, где максимум берется по всем x из Z, а теоретическая мода есть $\arg\max f(x)$, где максимум берется по тем же x, то для доказательства сходимости выборочной моды к теоретической кажется естественным применить методы изучения асимптотики решений экстремальных статистических задач (см. [1], [26]). Однако возникают сложности, связанные с тем, что не являются ограниченными случайные функции $g_n(x, X_i)$ и их дисперсии. Кроме того, эти функции не являются асимптотически равномерно разбиваемыми [26]. В общей теории асимптотики решений экстремальных статистических задач показано, что асимптотическая равномерная разбиваемость тесно связана с равномерной сходимостью, в то время как для линейных оценок плотности на прямой, как известно, [20, с.68-70], требуется выполнение ряда условий. Поэтому нельзя ожидать простоты формулировок аналогичных результатов для пространств общей природы. Приведем один результат о сходимости выборочной моды к теоретической.

Теорема 4. *Пусть Z состоит из конечного числа элементов, условия теоремы 2 выполнены для всех x из Z. Тогда выборочная мода сходится к теоретической по вероятности при росте объема выборки.*

Доказательство вытекает из теоремы 2.2.2 [1] и теоремы 2 выше.

Пусть множество Z конечно, сигма-алгебра $\mathfrak{A}$ измеримых подмножеств совпадает с множеством всех подмножеств Z, мера p – считающая, т.е. $p(\{x\}) = 1$ для любого x из Z. Тогда $f(x) = \mathbf{P}(X_1 = x)$ для любого x из Z,

условия (I), (II), (IV) и (VII) всегда выполнены, условия (III), (V) и (VI) переходят в условия

$$\sum_{y\in Z} g_n(x,y) = 1, \tag{27}$$

$$\max_n \sum_{y\in Z} |g_n(x,y)| < \infty, \tag{28}$$

$$\lim_{n\to\infty} \sum_{y:y\neq x} |g_n(x,y)| = 0 \tag{29}$$

соответственно. Условие (29) можно заменить на более простое: для любого x из Z

$$\lim_{n\to\infty} g_n(x,x) = 1. \tag{30}$$

Обычная оценка вероятности $\mathbf{P}(X_1 = x)$ – частота (число совпадений элементов выборки с точкой x, деленное на объем выборки) – является частным случаем гистограммной оценки (3), если области разбиения T_n есть одноэлементные множества $\{x\}$. Переход к использованию g_n позволяет "сглаживать" частотную оценку. Замечание. Поскольку плотность вероятности $f \geq 0$, то представляется естественным потребовать, чтобы выполнялось условие $f_n \geq 0$, а потому и условие $g_n \geq 0$, что делает ненужным условие (I). Однако при этом увеличивается смещение и уменьшается скорость сходимости ядерных оценок (4). Оказывается целесообразным использование знакопеременных ядерных функций (см. [1], [24]). Поэтому мы не считаем целесообразным принимать условие $f_n \geq 0$. Асимптотической теории конкретных видов линейных оценок (ядерных, гистограммных, типа Фикс-Ходжеса), а также применению линейных оценок и их частных видов для решения различных задач нечисловой статистики (построения оценок условной плотности, условного среднего, т.е. регрессионной зависимости, правил принятия решений в дискриминантном анализе, при проверке гипотезы однородности двух выборок и др.) должны быть посвящены отдельные публикации.

Библиографический список

1. *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.

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

3. *Ибрагимов И.А., Розанов Ю.А.* Гауссовские случайные процессы. - М.: її Медиа, 2012. - 385 с.

4. *Липцер Р.Ш., Ширяев А.Н.* Статистика случайных процессов. - М.: Наука, 1974. - 696 с.

5. *Скороход А.В.* Интегрирование в гильбертовом пространстве. - М.: Наука, 1975. - 232 с.

6. *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.

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

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

9. *Ченцов Н.Н.* Статистические решающие правила и оптимальные выводы. - М.: її Медиа, 2012. - 524 с.

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

11. *Rosenblatt M.* Remarks on some nonparametric estimates of a density function // Ann. Math. Statist. - 1956. - V.27. - №5. - P. 832 - 837.

12. *Parzen E.* On estimation of a probability density function and mode // Ann. Math. Statist. - 1962. - V.33. - №6. - P.1065-1076.

13. *Fix E., Hodges J.L.* Discriminatory analysis: nonparametric discrimination: consistency properties. - Rep. N 4. - USAF school of Aviation Medicine. - Texas. - February 1951. - Project 21-49-004. - Contract AF-41-(128)-31.

14. *Фукунага К.* Введение в статистическую теорию распознавания образов. - М.: Наука, 1979. - 368 с.

15. *Дуда Р., Харт П.* Распознавание образов и анализ сцен. - М.: Мир, 1976. – 511 с.

16. *Вапник В.Н., Стефанюк А.Р.* Непараметрические методы восстановления плотности вероятности // Автоматика и телемеханика. – 1978. – №8. – С.38 - 52.

17. *Богданов Ю.И.* Информация Фишера и непараметрическая аппроксимация плотности распределения // Заводская лаборатория. – 1998. – №7. – С.56-61.

18. *Богданов Ю.И.* Метод максимального правдоподобия и корневая оценка плотности распределения // Заводская лаборатория. – 2004. – №3. – С.51- 59.

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

20. *Мания Г.М.* Статистическое оценивание распределения вероятностей. - Тбилиси: Изд-во Тбилисского ун-та, – 1974. – 240 с.

21. *Мешалкин Л.Д.* Локальные методы классификации // Статистические методы классификации: - М.: Изд-во МГУ им. М.В. Ломоносова, – 1969. – Вып.1. – С.58-78.

22. *Деврой Л., Дьерди Л.* Непараметрическое оценивание плотности (L1 -подход). - М.: Мир, 1988. – 408 с.

23. *Лапко А.В., Лапко В.А.* Непараметрическая оценка плотности вероятности независимых случайных величин // Стохастические системы. – 2011. – №3(29). – С.118-124.

24. *Орлов А.И.* Ядерные оценки плотности в пространствах произвольной природы // Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского гос. университета, – 1996. – С.68-75.

25. *Гнеденко Б.В.* Курс теории вероятностей. Изд. 6-е, перераб. и доп. - М.: Наука, 1988. – 448 с.

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

Density estimates in spaces of arbitrary nature

A.I. Orlov

Bauman Moscow State Technical University, Institute of high statistical technologies and econometrics; Moscow Physics-Technical Institute; Moscow, Russia; prof-orlov@mail.ru; http://orlovs.pp.ru; +7(916)8305117; 123104, Moscow, the Sytinsky lane, house 7/14, apartment 14.

Abstract. Linear estimators the probability density in the spaces of an arbitrary nature and particular cases - nuclear, histogram, the Fix-Hodges type estimates are introduced. Consistency and asymptotic normality of linear estimates are proved under natural conditions. It is shown that the probability of the area can be found by linear density estimates. A special case of a finite set are discussed, it was found that sample mode converges to the theoretical one.

Key words: non-numeric statistics, probability density function, the space of an arbitrary nature, linear density estimators, limit theorems, consistent estimators, asymptotic normality.