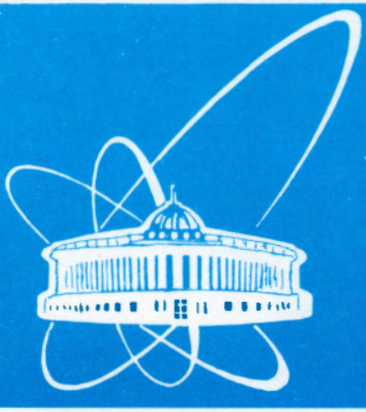

ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ

Дубна

P11-99-329

Е.П.Жидков[1], А.Г.Соловьев[2], А.Н.Соснин[3]

ПОВЫШЕНИЕ ТОЧНОСТИ ОЦЕНКИ НЕИЗВЕСТНОЙ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПО ЭМПИРИЧЕСКИМ ДАННЫМ

Направлено в журнал «Nuclear Instruments and Methods»

[1]E-mail: zhidkov@lcta41.jinr.ru
[2]E-mail: solovjev@decimal.jinr.ru
[3]E-mail: sosnin@decimal.jinr.ru

1999

ВВЕДЕНИЕ

В настоящей работе[1] изучается проблема оценки неизвестной плотности распределения случайной величины по результатам ее наблюдений. Существующие различные способы такой оценки (о некоторых из них будет сказано ниже) позволяют приблизить искомую плотность с той или иной погрешностью. Как правило, эта погрешность тем меньше, чем больше число наблюдений, по которым делается оценка. Заметим, однако, что скорость ее убывания невысока: при оценке неизвестной плотности может иметь место сходимость со скоростью не выше чем $n^{-1/2}$, где n — число наблюдений. Это естественный факт, так как в оценке значения искомой плотности в некоторой точке принимает участие не вся выборка, а лишь те наблюдения, которые сосредоточились в некоторой окрестности этой точки.

В связи с этим актуальным является вопрос о повышении скорости сходимости оценки к искомой плотности. Для решения этой задачи в настоящей работе предлагается подход, идея которого аналогична той, что используется в [1, 2] для уточнения приближенных решений сингулярных краевых задач на бесконечном интервале. Применительно к изучаемому здесь вопросу об оценке неизвестной плотности по наблюдениям предлагаемый метод заключается в экстраполяции по числу наблюдений n. А именно, если имеются две независимые серии наблюдений, то по ним можно, во-первых, построить две оценки неизвестной плотности — по каждой выборке отдельно, и, во-вторых, построить оценку, объединив все наблюдения в одну выборку. Определенная же линейная комбинация всех трех оценок при $n \to \infty$ будет сходиться к искомой плотности быстрее, чем каждая из этих оценок по отдельности.

1. СПОСОБЫ ОЦЕНКИ НЕИЗВЕСТНОЙ ПЛОТНОСТИ

Пусть действительная непрерывная случайная величина ξ имеет плотность распределения $f(x)$, которая заранее неизвестна. Эту неизвестную плотность $f(x)$ требуется оценить на основе выборки

$$x_1\,, \quad x_2\,, \quad \ldots\,, \quad x_n\,, \tag{1.1}$$

являющейся результатом n последовательных независимых наблюдений ξ в неизменных условиях. Опишем несколько способов решения этой задачи.

Часто практикуемый прием приближенной оценки неизвестной функции $f(x)$ по данным выборки заключается в построении *гистограммы* (см., например, [3], с. 361–362). Для этого задают отрезок $[a, b]$, содержащий выборку (1.1), разбивают его на маленькие интервалы Δ_i длины h_n и полагают $f_n(x) = \nu_i/(nh_n)$ при $x \in \Delta_i$, где ν_i — число элементов выборки, попавших в Δ_i (вне отрезка $[a, b]$ полагают $f_n(x) = 0$). Если $f(x)$ непрерывна, и $h_n \to 0$, $nh_n \to \infty$ при $n \to \infty$, то

[1]Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (коды проектов 97 – 01 – 00746, 98 – 01 – 00190).

гистограмма $f_n(x)$ сходится по вероятности к искомой плотности $f(x)$. Более точные утверждения о сходимости были получены В.И.Гливенко (см. [4], с. 179–180). При определенной скорости убывания последовательности h_n погрешность оценки $f_n(x)$ имеет порядок $n^{-1/3}$ (см. [5]).

Другой способ оценки неизвестной плотности состоит в построении *полигона частот*. Это делается следующим образом (см., например, [6], с. 23–24). В серединах определенных выше интервалов Δ_i строятся перпендикуляры длины $\nu_i/(nh_n)$, где ν_i — число выборочных значений, попавших в Δ_i, и верхние концы этих перпендикуляров соединяются. Полученная ломаная линия $f_n(x)$ считается оценкой плотности $f(x)$. Если $f(x)$ имеет ограниченную вторую производную, то при определенной скорости убывания h_n погрешность $f_n(x)$ имеет порядок $n^{-2/5}$ (см. [5]). Повышение точности при таком подходе по сравнению с гистограммой обусловлено тем, что кусочно-линейные функции, к которым относится полигон частот, дают возможность приблизить гладкую функцию гораздо лучше, чем кусочно-постоянные (типа гистограммы).

В настоящей работе для оценки неизвестной плотности $f(x)$ выбран класс оценок, который впервые был введен Е.Парзеном и М.Розенблаттом (см., например, [7], с. 71–72; [8], с. 312–314). Оценка искомой плотности $f(x)$ — *эмпирическая плотность* — строится по выборке (1.1) следующим образом:

$$f_n(x) = \frac{1}{nh_n} \sum_{k=1}^{n} q\left(\frac{x - x_k}{h_n}\right) , \tag{1.2}$$

где последовательность h_n такова, что

$$h_n \to 0 , \quad nh_n \to \infty , \tag{1.3}$$

при $n \to \infty$, а функция $q(x)$ гладкая, ограниченная и интегрируемая на интервале $-\infty < x < \infty$, причем

$$\int_{-\infty}^{\infty} q(x)dx = 1 . \tag{1.4}$$

Другие требования к последовательности h_n и функции $q(x)$ будут сформулированы ниже. (Заметим, что эмпирическая плотность (1.2) есть не что иное, как "средняя сумма" плотностей $q(x)$, сжатых до размеров h_n и "посаженных" в точки x_k). При определенных условиях гладкости искомой плотности $f(x)$ можно построить для нее оценки вида (1.2), обладающие более высокой по сравнению с гистограммой или полигоном частот скоростью сходимости.

2. СХОДИМОСТЬ ЭМПИРИЧЕСКОЙ ПЛОТНОСТИ

Исследуем вопрос о сходимости оценки (1.2) к искомой плотности $f(x)$. Исходным для нас является следующее утверждение (см. [7], с. 73–74):

Теорема 1. *Пусть функция $f(x)$ непрерывна и ограничена, последовательность h_n удовлетворяет условиям* (1.3), *а функция $q(x)$ — условию* (1.4), *а также следующему условию:*

$$\int_{-\infty}^{\infty} q^2(x)dx = d^2 < \infty . \tag{2.1}$$

Тогда для оценки (1.2) *имеет место разложение:*

$$f_n(x) = \mathsf{M}f_n(x) + \zeta_n(x)/\sqrt{nh_n} , \tag{2.2}$$

где $\mathsf{M}f_n(x)$ — математическое ожидание случайной величины $f_n(x)$, причем $\mathsf{M}f_n(x) \to f(x)$ при $h_n \to 0$, а распределение случайной величины $\zeta_n(x)$ при $n \to \infty$ сходится к нормальному со средним значением 0 и дисперсией $f(x)d^2$.

Таким образом, при условиях теоремы 1 оценка $f_n(x)$ сближается с искомой плотностью $f(x)$ с ростом числа наблюдений n, причем скорость этой сходимости не выше чем $n^{-1/2}$ и зависит от скорости убывания последовательности h_n. Действительно, запишем (2.2) в виде:

$$f_n(x) = f(x) + [\mathsf{M}f_n(x) - f(x)] + \zeta_n(x)/\sqrt{nh_n} ,$$

откуда следует, что погрешность оценки $f_n(x)$ имеет систематическую и случайную составляющие, первая убывает по абсолютной величине при $h_n \to 0$, вторая — при $nh_n \to \infty$. Поэтому если с ростом n последовательность h_n убывает слишком медленно, то в погрешности преобладает систематическая ошибка, а если слишком быстро — случайная. Для того чтобы выбрать последовательность h_n оптимальным образом, необходимо, прежде всего, сделать более определенные выводы о функции $\mathsf{M}f_n(x)$, входящей в разложение (2.2). Это можно сделать при некоторых дополнительных предположениях.

Относительно искомой функции $f(x)$ мы будем предполагать существование у нее ограниченной четвертой производной. А от функции $q(x)$, дополнительно к уже сделанным относительно нее предположениям (1.4) и (2.1), потребуем следующее:

1. Плотность $q(x)$ имеет ограниченные моменты до четвертого порядка включительно:

$$\int_{-\infty}^{\infty} x^l q(x)dx = D_l < \infty , \quad l = 1, 2, 3, 4 . \tag{2.3}$$

2. Функция $q(x)$ четная, то есть $q(x) = q(-x)$, из чего следует, что

$$D_1 = D_3 = 0 . \tag{2.4}$$

3. Функция $q(|x|)$ монотонно не возрастает.[2]

Получим выражение для $\mathsf{M}f_n(x)$ при этих условиях. Исходя из (1.2) имеем:

$$\begin{aligned}\mathsf{M}f_n(x) &= \mathsf{M}h_n^{-1}q\left[(x-x_1)/h_n\right] = \\ &= \frac{1}{h_n}\int\limits_{-\infty}^{\infty} q\left(\frac{x-t}{h_n}\right) f(t)dt = \int\limits_{-\infty}^{\infty} q(z)f(x-zh_n)dz\ . \end{aligned} \tag{2.5}$$

В силу нашего предположения о функции $f(x)$ для нее справедливо разложение Тейлора:

$$f(x-zh_n) = f(x) - f'(x)zh_n + f''(x)z^2h_n^2/2 - f'''(x)z^3h_n^3/6 + R_4(x)\ ,$$

в котором остаточный член $R_4(x)$ имеет вид

$$R_4(x) = f^{(4)}\left(x-\theta zh_n\right)z^4h_n^4/24\ , \quad 0<\theta<1\ .$$

Из ограниченности четвертой производной функции $f(x)$ следует, что

$$R_4(x) = O\left[(zh_n)^4\right]\ .$$

Следовательно, подставив разложение Тейлора в интеграл (2.5) и воспользовавшись условиями (2.3), получим

$$\mathsf{M}f_n(x) = f(x) - f'(x)D_1h_n + f''(x)D_2h_n^2/2 - f'''(x)D_3h_n^3/6 + O\left(h_n^4\right)\ .$$

Учитывая теперь условия (2.4), получаем следующее выражение:

$$\mathsf{M}f_n(x) = f(x) + f''(x)D_2h_n^2/2 + O\left(h_n^4\right)\ . \tag{2.6}$$

Последовательность h_n, как это часто делается, мы определим с учетом условий (1.3) следующим образом:

$$h_n = cn^{-\alpha}\ , \quad 0<\alpha<1\ , \tag{2.7}$$

(см., например, [8], с. 315, 317), где c — положительная постоянная, о выборе которой будет сказано ниже. В этом случае разложение (2.6) принимает вид

$$\mathsf{M}f_n(x) = f(x) + \left(f''(x)D_2c^2/2\right)n^{-2\alpha} + O\left(n^{-4\alpha}\right)\ .$$

Подставив это выражение в (2.2), получим

$$f_n(x) = f(x) + \left(f''(x)D_2c^2/2\right)n^{-2\alpha} + \left(\zeta_n(x)/\sqrt{c}\right)n^{(\alpha-1)/2} + O\left(n^{-4\alpha}\right)\ . \tag{2.8}$$

Из этого разложения нетрудно установить, каким образом надо выбирать значение α из интервала $0<\alpha<1$, чтобы получить максимальный порядок сходимости оценки $f_n(x)$ к точной плотности $f(x)$.

[2]Последнее условие нам потребуется ниже для оценки погрешности эмпирической плотности. При исследовании функции $\mathsf{M}f_n(x)$ оно не используется.

Как правило, максимальной скорости сходимости $f_n(x)$ к $f(x)$ добиваются следующим образом (см. [7], с. 75; [8], с. 315, 317). Из (2.8) видно, что систематическая ошибка оценки $f_n(x)$ имеет порядок $O\left(n^{-2\alpha}\right)$, и скорость ее убывания при $n\to\infty$ тем выше, чем больше α. С другой стороны, случайная погрешность в (2.8) имеет порядок $O\left(n^{(\alpha-1)/2}\right)$ и убывает при $n\to\infty$ тем быстрее, чем α меньше. Очевидно поэтому, что наилучшую по порядку величины скорость сходимости мы получим, если $-2\alpha = (\alpha-1)/2$, то есть при $\alpha = 1/5$. В этом случае $f_n(x) = f(x) + O\left(n^{-2/5}\right)$.

В настоящей работе для улучшения сходимости оценки к искомой плотности предлагается другой подход, позволяющий достичь более высокой, чем $O\left(n^{-2/5}\right)$, скорости сходимости. Заметим, что коэффициент при $n^{-2\alpha}$ в разложении (2.8) не зависит от n, что позволяет сократить это слагаемое путем экстраполяции по n (подробно о том, как это делается, речь пойдет ниже). Поэтому, следуя предыдущим рассуждениям, выберем значение α таким образом, чтобы $-4\alpha = (\alpha-1)/2$, то есть положим $\alpha = 1/9$. Тогда после экстраполяции получим оценку с погрешностью порядка $O\left(n^{-4/9}\right)$.

Замечания. 1. Отметим, что при сделанных предположениях скорость сходимости порядка $O\left(n^{-4/9}\right)$ можно обеспечить также выбором такой функции $q(x)$, для которой $D_2 = 0$ (см. [8], с. 316–317). При этом, очевидно, она должна принимать значения обоих знаков. В результате с положительной вероятностью в качестве оценки неотрицательной плотности появляется знакопеременная функция. Иногда это может быть неудобно, например, когда делается предварительная оценка неизвестной плотности с использованием выборки небольшого объема. В настоящей работе мы будем выбирать только неотрицательные функции $q(x)$, а скорость сходимости оценки к искомой плотности при $n\to\infty$ будем повышать, применяя экстраполяцию по n.

2. Скорость убывания последовательности h_n (в нашем случае эта скорость определяется параметром α) иногда предпочитают выбирать таким образом, чтобы систематическая погрешность построенной оценки при $n\to\infty$ имела более высокий порядок малости, чем случайная (так, например, делается при оценке неизвестной плотности гистограммой или полигоном частот в [9], с. 205–223). Из (2.8) видно, что для этого следует выбирать α из интервала $1/5<\alpha<1$. А при том подходе, в котором предполагается последующая экстраполяция по n, — из интервала $1/9<\alpha<1$. Отметим, что при таком выборе α скорость сходимости будет несколько ниже, чем максимально возможная, однако в этом случае появляется возможность построить доверительную область, которая с заданной вероятностью будет покрывать график искомой плотности. Это часто бывает необходимо в прикладных задачах.

3. УТОЧНЕНИЕ ЭМПИРИЧЕСКОЙ ПЛОТНОСТИ

Опишем теперь метод повышения скорости сходимости оценки (1.2) к искомой плотности. При сделанных предположениях справедлива следующая

Теорема 2. *Пусть имеются две независимые выборки*

$$x_1^{(1)}\ ,\quad x_2^{(1)}\ ,\quad \ldots\ ,\quad x_n^{(1)}\ ,$$

$$x_1^{(2)}\ ,\quad x_2^{(2)}\ ,\quad \ldots\ ,\quad x_n^{(2)}\ ,$$

и $f_n^{(1)}(x)$, $f_n^{(2)}(x)$ — отвечающие им эмпирические плотности, построенные по формуле (1.2), причем $q(x)$ удовлетворяет условиям 1 — 3, а h_n определяется по формуле (2.7). Пусть, кроме того, $f_{2n}^{(1+2)}(x)$ — эмпирическая плотность, полученная аналогичным образом после объединения этих двух выборок в одну.

Тогда для линейной комбинации

$$f_{2n}^{(*)}(x) = \gamma_1 f_n^{(1)}(x) + \gamma_2 f_n^{(2)}(x) + \gamma_{1+2} f_n^{(1+2)}(x)\ , \tag{3.1}$$

в которой

$$\gamma_1 = \gamma_2 = -2^{-1}/\left(2^{2\alpha}-1\right)\ ,\quad \gamma_{1+2} = 2^{2\alpha}/\left(2^{2\alpha}-1\right)\ , \tag{3.2}$$

имеет место разложение:

$$f_{2n}^{(*)}(x) \;=\; f(x) + \eta_{2n}^{(*)}(x) n^{(\alpha-1)/2} + O\left(n^{-4\alpha}\right)\ , \tag{3.3}$$

распределение случайной величины $\eta_{2n}^{(*)}(x)$ в котором при $n\to\infty$ сходится к нормальному со средним значением 0 и дисперсией, ограниченной величиной

$$\left[f(x)d^2\left(2^{5\alpha}+1-2^{1+2\alpha}\right)\right]/\left[4c\left(2^{2\alpha}-1\right)^2\right]\ .$$

Доказательство. Запишем разложение (2.8) для всех имеющихся эмпирических плотностей:

$$\begin{aligned}
f_n^{(1)}(x) = f(x) &+ \left(f''(x)D_2c^2/2\right)n^{-2\alpha} + \\
&+ \left(\zeta_n^{(1)}(x)/\sqrt{c}\right)n^{(\alpha-1)/2} + O\left(n^{-4\alpha}\right)\ , \\
f_n^{(2)}(x) = f(x) &+ \left(f''(x)D_2c^2/2\right)n^{-2\alpha} + \\
&+ \left(\zeta_n^{(2)}(x)/\sqrt{c}\right)n^{(\alpha-1)/2} + O\left(n^{-4\alpha}\right)\ , \\
f_n^{(1+2)}(x) = f(x) &+ 2^{-2\alpha}\left(f''(x)D_2c^2/2\right)n^{-2\alpha} + \\
&+ 2^{(\alpha-1)/2}\left(\zeta_{2n}^{(1+2)}(x)/\sqrt{c}\right)n^{(\alpha-1)/2} + O\left(n^{-4\alpha}\right)\ .
\end{aligned}$$

Подставив эти разложения в линейную комбинацию (3.1), получим

$$\begin{aligned}
f_{2n}^{(*)}(x) \;=\;& \left(\gamma_1+\gamma_2+\gamma_{1+2}\right)f(x) + \\
&+ \left(\gamma_1+\gamma_2+2^{-2\alpha}\gamma_{1+2}\right)\left(f''(x)D_2c^2/2\right)n^{-2\alpha} + \\
&+ \left[\left(\gamma_1\zeta_n^{(1)}(x)+\gamma_2\zeta_n^{(2)}(x)+2^{(\alpha-1)/2}\gamma_{1+2}\zeta_{2n}^{(1+2)}(x)\right)/\sqrt{c}\right]n^{(\alpha-1)/2} + \\
&+ O\left(n^{-4\alpha}\right)\ .
\end{aligned}$$

Выберем теперь коэффициенты γ_1, γ_2 и γ_{1+2} таким образом, чтобы

$$\gamma_1+\gamma_2+\gamma_{1+2} = 1\ ,\quad \gamma_1+\gamma_2+2^{-2\alpha}\gamma_{1+2} = 0\ .$$

Из этой системы уравнений следует, что

$$\gamma_1+\gamma_2 = -1/\left(2^{2\alpha}-1\right)\ ,\quad \gamma_{1+2} = 2^{2\alpha}/\left(2^{2\alpha}-1\right)\ .$$

Поскольку оценки $f_n^{(1)}(x)$ и $f_n^{(2)}(x)$ построены по выборкам равного объема, ни одной из этих оценок нельзя отдать предпочтение. Поэтому естественно положить $\gamma_1=\gamma_2$. В результате коэффициенты γ_1, γ_2 и γ_{1+2} определяются выражениями (3.2), а для линейной комбинации (3.1) имеет место разложение (3.3), где

$$\eta_{2n}^{(*)}(x) = \left[2^{(5\alpha-1)/2}\zeta_{2n}^{(1+2)}(x) - \left(\zeta_n^{(1)}(x)+\zeta_n^{(2)}(x)\right)/2\right]/\left[\sqrt{c}\left(2^{2\alpha}-1\right)\right]\ .$$

Нам, таким образом, остается доказать утверждения теоремы, касающиеся случайной величины $\eta_{2n}^{(*)}(x)$.

Согласно [7], с. 73, имеем

$$\begin{aligned}
\zeta_n^{(1)}(x) &= \sum_{k=1}^{n}\left\{q\left[(x-x_k^{(1)})/h_n\right] - h_n\mathsf{M}f_n^{(1)}(x)\right\}/\sqrt{nh_n}\ , \\
\zeta_n^{(2)}(x) &= \sum_{k=1}^{n}\left\{q\left[(x-x_k^{(2)})/h_n\right] - h_n\mathsf{M}f_n^{(2)}(x)\right\}/\sqrt{nh_n}\ , \\
\zeta_{2n}^{(1+2)}(x) &= \sum_{k=1}^{2n}\left\{q\left[(x-z_k)/h_{2n}\right] - h_{2n}\mathsf{M}f_{2n}^{(1+2)}(x)\right\}/\sqrt{2nh_{2n}}\ ,
\end{aligned}$$

где z_k есть элемент объединенной выборки: $z_k = x_k^{(1)}$, $z_{n+k} = x_k^{(2)}$, $k=1,\ldots,n$, а $\mathsf{M}f_n^{(1)}(x) = \mathsf{M}f_n^{(2)}(x)$. Поэтому, как нетрудно убедиться, случайную величину $\eta_{2n}^{(*)}(x)$ можно представить в виде

$$\eta_{2n}^{(*)}(x) = \sum_{k=1}^{2n}\nu_k(x)\ ,$$

где

$$\begin{aligned}
\nu_k(x) \;=\;& \left\{\left[2^{5\alpha/2}q\left[(x-z_k)h_{2n}\right]/\sqrt{h_{2n}} - q\left[(x-z_k)/h_n\right]/\sqrt{h_n}\right] - \right. \\
&\left. - \left[2^{5\alpha/2}\sqrt{h_{2n}}\mathsf{M}f_{2n}^{(1+2)}(x) - \sqrt{h_n}\mathsf{M}f_n^{(1)}(x)\right]\right\}/\left[2\sqrt{c}\left(2^{2\alpha}-1\right)\sqrt{n}\right]\ .
\end{aligned}$$

Следовательно, $\eta_{2n}^{(*)}(x)$ есть сумма независимых одинаково распределенных случайных величин. Применимость к ней центральной предельной теоремы можно доказать, пользуясь рассуждениями [7], с. 74. Действительно, условие Линдберга в нашем случае имеет вид:

$$n\int\limits_{\{x:|\nu_1(x)|>\epsilon\}}\nu_1^2(x)f(x)dx \to 0$$

при $n\to\infty$, и можно проверить, что оно выполнено. Таким образом, случайная величина $\eta_{2n}^{(*)}(x)$ асимптотически нормальна (доказательство теоремы

Линдберга см., например, в [10], с. 251–255). Оценим предельные значения ее математического ожидания и дисперсии при $n \to \infty$.

Поскольку $\mathsf{M}\nu_k(x) = 0$, отюда сразу получаем, что $\mathsf{M}\eta_{2n}^{(*)}(x) = 0$. (Предельное значение математического ожидания $\mathsf{M}\eta_{2n}^{(*)}(x)$, очевидно, также равно нулю). Для оценки дисперсии $\eta_{2n}^{(*)}(x)$ будем исходить из выражения

$$\begin{aligned}\mathsf{M}\nu_k^2(x) &= \left\{\left[2^{5\alpha}\mathsf{M}\left(q^2\left[(x-z_k)/h_{2n}\right]\right)/h_{2n} + \mathsf{M}\left(q^2\left[(x-z_k)/h_n\right]\right)/h_n\right. - \\ &- 2^{1+5\alpha/2}\mathsf{M}\left(q\left[(x-z_k)/h_{2n}\right]q\left[(x-z_k)/h_n\right]\right)/\sqrt{h_{2n}h_n}\Big] - \\ &- \left[2^{5\alpha}h_{2n}\left(\mathsf{M}f_{2n}^{(1+2)}(x)\right)^2 + h_n\left(\mathsf{M}f_n^{(1)}(x)\right)^2 - \right. \\ &- \left.2^{1+5\alpha/2}\sqrt{h_{2n}h_n}\mathsf{M}f_{2n}^{(1+2)}(x)\mathsf{M}f_n^{(1)}(x)\right]\Big\}\Big/\left[4c\left(2^{2\alpha}-1\right)^2 n\right] . \quad (3.4)\end{aligned}$$

Оценим это выражение при $n \to \infty$. Учитывая, что при этом $h_n \to 0$ и $h_{2n} \to 0$, получаем

$$\begin{aligned}\frac{1}{h_{2n}}\mathsf{M}q^2\left(\frac{x-z_k}{h_{2n}}\right) &= \frac{1}{h_{2n}}\int_{-\infty}^{\infty} q^2\left(\frac{x-t}{h_{2n}}\right)f(t)dt = \\ &= \int_{-\infty}^{\infty} q^2(z)f(x-zh_{2n})dz \to f(x)d^2 , \\ \frac{1}{h_n}\mathsf{M}q^2\left(\frac{x-z_k}{h_n}\right) &= \frac{1}{h_n}\int_{-\infty}^{\infty} q^2\left(\frac{x-t}{h_n}\right)f(t)dt = \\ &= \int_{-\infty}^{\infty} q^2(z)f(x-zh_n)dz \to f(x)d^2 .\end{aligned}$$

Кроме того, $h_{2n}/h_n = 2^{-\alpha}$ в силу (2.7), поэтому

$$\begin{aligned}&\frac{1}{\sqrt{h_{2n}h_n}}\mathsf{M}\left[q\left(\frac{x-z_k}{h_{2n}}\right)q\left(\frac{x-z_k}{h_n}\right)\right] = \\ &= \frac{1}{\sqrt{h_{2n}h_n}}\int_{-\infty}^{\infty} q\left(\frac{x-t}{h_{2n}}\right)q\left(\frac{x-t}{h_n}\right)f(t)dt = \\ &= \sqrt{\frac{h_{2n}}{h_n}}\int_{-\infty}^{\infty} q(z)q\left(\frac{h_{2n}}{h_n}z\right)f(x-zh_{2n})dz \to \\ &\to f(x)\sqrt{\frac{h_{2n}}{h_n}}\int_{-\infty}^{\infty} q(z)q\left(\frac{h_{2n}}{h_n}z\right)dz = 2^{-\alpha/2}f(x)\int_{-\infty}^{\infty} q(z)q\left(2^{-\alpha}z\right)dz .\end{aligned}$$

Так как функция $q(x)$ удовлетворяет условиям 2 и 3, и $2^{-\alpha} < 1$ при $0 < \alpha < 1$, для интеграла в последнем выражении получим оценку

$$\int_{-\infty}^{\infty} q(z)q\left(2^{-\alpha}z\right)dz = 2\int_0^{\infty} q(z)q\left(2^{-\alpha}z\right)dz \geq 2\int_0^{\infty} q^2(z)dz = d^2 .$$

Остальные слагаемые в фигурных скобках в (3.4) стремятся к нулю при $n \to \infty$. В результате приходим к следующей оценке:

$$\lim_{n\to\infty} n\mathsf{M}\nu_k^2(x) \leq \left[f(x)d^2\left(2^{5\alpha}+1-2^{1+2\alpha}\right)\right]/\left[4c\left(2^{2\alpha}-1\right)^2\right] .$$

Заметим, что $\mathsf{D}\nu_k(x) = \mathsf{M}\nu_k^2(x)$ (здесь $\mathsf{D}\nu_k(x)$ — дисперсия случайной величины $\nu_k(x)$). Отсюда получаем оценку для предельного значения дисперсии случайной величины $\eta_{2n}^{(*)}(x)$:

$$\lim_{n\to\infty} \mathsf{D}\eta_{2n}^{(*)}(x) \leq \left[f(x)d^2\left(2^{5\alpha}+1-2^{1+2\alpha}\right)\right]/\left[4c\left(2^{2\alpha}-1\right)^2\right] .$$

Теорема доказана.

Замечание 3. Отметим, что если $\alpha < 1/9$, то, как видно из (3.3), систематическая погрешность оценки $f_{2n}^{(*)}(x)$ при $n \to \infty$ убывает по абсолютной величине медленнее, чем случайная, что нежелательно. Если же $\alpha > 1/5$, то из (2.8) видим, что систематическая погрешность оценки $f_n(x)$ при $n \to \infty$ имеет более высокий порядок малости, чем случайная, поэтому экстраполяция по n в этом случае нецелесообразна.

4. О ПОСТРОЕНИИ ДОВЕРИТЕЛЬНОЙ ОБЛАСТИ

Для оценки погрешности эмпирической плотности, построенной по формуле (3.1), а также для построения доверительной области, с заданной вероятностью покрывающей график искомой плотности $f(x)$, докажем следствие из теоремы 2:

Теорема 3. *Пусть выполнены условия теоремы* 2, *причем параметр* α *в формуле* (2.7) *удовлетворяет условиям* $1/9 < \alpha \leq 1/5$, *а искомая плотность* $f(x)$ *— условию* $\min\limits_{a\leq x\leq b} f(x) > 0$.

Тогда для любого $\lambda \geq 0$ *имеет место неравенство*

$$\lim_{n\to\infty} \mathsf{P}\left\{\max_{a\leq x\leq b}\left[\left|f_{2n}^{(*)}(x) - f(x)\right|/\sqrt{f(x)}\right] \leq \lambda r n^{(\alpha-1)/2}\right\} \geq 2\Phi(\lambda) , \quad (4.1)$$

где P *обозначает вероятность,*

$$r = d\sqrt{2^{5\alpha}+1-2^{1+2\alpha}}/\left[2\sqrt{c}\left(2^{2\alpha}-1\right)\right] , \quad (4.2)$$

а функция $\Phi(\lambda)$ *определена*[3] *следующим образом:*

$$\Phi(\lambda) = \frac{1}{\sqrt{2\pi}}\int_0^{\lambda} e^{-t^2/2}dt .$$

Доказательство. В теореме 2 было установлено, что

$$f_{2n}^{(*)}(x) - f(x) = \eta_{2n}^{(*)}(x)n^{(\alpha-1)/2} + s_{2n}^{(*)}(x) ,$$

[3]Таблица значений функции $\Phi(x)$ приводится, например, в [10], с. 442–443.

где $s_{2n}^{(*)}(x) = O\left(n^{-4\alpha}\right)$ при $n \to \infty$. Кроме того, из этой теоремы следует, что распределение случайной величины $\eta_{2n}^{(*)}(x)/\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)}$ при $n \to \infty$ сходится к нормальному со средним значением 0 и дисперсией 1, то есть для любого $\lambda \geq 0$ имеем:

$$\lim_{n\to\infty} \mathsf{P}\left\{\left|\eta_{2n}^{(*)}(x)\right| / \sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} \leq \lambda\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{\lambda} \mathrm{e}^{-t^2/2} dt = 2\Phi(\lambda) \; .$$

Отсюда получаем

$$\lim_{n\to\infty} \mathsf{P}\left\{\left|\eta_{2n}^{(*)}(x)\right| n^{(\alpha-1)/2} \leq \lambda\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} = 2\Phi(\lambda) \; . \tag{4.3}$$

Докажем теперь, что для любого $\epsilon > 0$ имеют место следующие неравенства:

$$\begin{aligned}
\mathsf{P}\left\{\left|f_{2n}^{(*)}(x) - f(x)\right| \;\leq\; & \lambda\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} \geq \\
\geq\; & \mathsf{P}\left\{\left|\eta_{2n}^{(*)}(x)\right| n^{(\alpha-1)/2} \leq (\lambda-\epsilon)\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} - \\
-\; & \mathsf{P}\left\{\left|s_{2n}^{(*)}(x)\right| > \epsilon\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} \; , \qquad (4.4) \\
\mathsf{P}\left\{\left|f_{2n}^{(*)}(x) - f(x)\right| \;\leq\; & \lambda\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} \leq \\
\leq\; & \mathsf{P}\left\{\left|\eta_{2n}^{(*)}(x)\right| n^{(\alpha-1)/2} \leq (\lambda+\epsilon)\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} + \\
+\; & \mathsf{P}\left\{\left|s_{2n}^{(*)}(x)\right| > \epsilon\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} \; , \qquad (4.5)
\end{aligned}$$

Действительно, из одновременного выполнения неравенств

$$\begin{aligned}
\left|\eta_{2n}^{(*)}(x)\right| n^{(\alpha-1)/2} \;&\leq\; (\lambda-\epsilon)\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2} \; , \\
\left|s_{2n}^{(*)}(x)\right| \;&\leq\; \epsilon\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2} \; ,
\end{aligned}$$

следует, что

$$\left|f_{2n}^{(*)}(x) - f(x)\right| \leq \lambda\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2} \; ,$$

и, кроме того, всегда

$$\mathsf{P}(AB) \geq \mathsf{P}(A) - \mathsf{P}(\bar{B}) \; . \tag{4.6}$$

Отсюда получаем неравенство (4.4). В то же время, из одновременного выполнения неравенств

$$\begin{aligned}
\left|f_{2n}^{(*)}(x) - f(x)\right| \;&\leq\; \lambda\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2} \; , \\
\left|s_{2n}^{(*)}(x)\right| \;&\leq\; \epsilon\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2} \; ,
\end{aligned}$$

следует, что

$$\left|\eta_{2n}^{(*)}(x)\right| n^{(\alpha-1)/2} \leq (\lambda+\epsilon)\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2} \; ,$$

поэтому, на основании (4.6), имеет место неравенство (4.5).

Заметим теперь, что при любом $\epsilon > 0$

$$\lim_{n\to\infty} \mathsf{P}\left\{\left|s_{2n}^{(*)}(x)\right| > \epsilon\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} = 0 \; ,$$

поскольку $s_{2n}^{(*)}(x) = O\left(n^{-4\alpha}\right)$ при $n \to \infty$, и $\alpha < 1/5$ (то есть $s_{2n}^{(*)}(x) = o\left(n^{(\alpha-1)/2}\right)$ при $n \to \infty$). Поэтому из соотношений (4.3), (4.4) и (4.4) получаем

$$\lim_{n\to\infty} \mathsf{P}\left\{\left|f_{2n}^{(*)}(x) - f(x)\right| \leq \lambda\sqrt{\mathsf{D}\eta_{2n}^{(*)}(x)} n^{(\alpha-1)/2}\right\} = 2\Phi(\lambda) \; . \tag{4.7}$$

Как было доказано в теореме 2, предельное значение дисперсии $\mathsf{D}\eta_{2n}^{(*)}(x)$ ограничено сверху величиной $\left[\left(2^{5\alpha}+1-2^{1+2\alpha}\right) f(x) d^2\right] / \left[4c\left(2^{2\alpha}-1\right)^2\right]$. Поэтому вероятность в (4.7) не уменьшится, если заменить $\mathsf{D}\eta_{2n}^{(*)}(x)$ этой величиной. В результате имеем:

$$\lim_{n\to\infty} \mathsf{P}\left\{\left[\left|f_{2n}^{(*)}(x) - f(x)\right| / \sqrt{f(x)}\right] \leq \lambda r n^{(\alpha-1)/2}\right\} \geq 2\Phi(\lambda) \; ,$$

где r определяется формулой (4.2). Поскольку последнее неравенство справедливо для всех значений x, для которых $f(x)$ отлична от нуля, отсюда следует (4.1). Теорема доказана.

С помощью теоремы 3 легко решается вопрос о построении доверительной области для теоретической плотности $f(x)$, отвечающей заданному коэффициенту доверия $0 < \beta < 1$. Прежде всего, по данному значению β из уравнения $2\Phi(\lambda) = \beta$ определяем величину λ_β. Неравенство

$$\max_{a\leq x\leq b}\left[\left|f_{2n}^{(*)}(x) - f(x)\right| / \sqrt{f(x)}\right] \leq t_\beta = \lambda_\beta r n^{(\alpha-1)/2}$$

означает, как легко показать элементарными рассуждениями, что график функции $f(x)$ на всем отрезке $[a,b]$ не выходит из полосы, ограниченной следующими двумя кривыми:

$$\begin{aligned}
y_1(x) \;&=\; f_{2n}^{(*)}(x) + t_\beta^2/2 - t_\beta\sqrt{f_{2n}^{(*)}(x) + t_\beta^2/4} \; , \\
y_2(x) \;&=\; f_{2n}^{(*)}(x) + t_\beta^2/2 + t_\beta\sqrt{f_{2n}^{(*)}(x) + t_\beta^2/4} \; .
\end{aligned}$$

Из теоремы 3 следует, что при достаточно большом n эта полоса накрывает график неизвестной плотности $f(x)$ с вероятностью, не меньшей β.

5. ВОПРОСЫ ОПТИМИЗАЦИИ

Из предыдущих рассуждений ясно, что погрешность оценки (1.2), а также оценки (3.1) при данном n будет тем меньше, чем меньше величина

$$d^2 = \int_{-\infty}^{\infty} q^2(x)dx \ ,$$

определенная формулой (2.1). Поэтому при выборе функции $q(x)$ следует руководствоваться требованием минимизировать эту величину при условиях (1.4), (2.3), (2.4):

$$\int_{-\infty}^{\infty} q(x)dx = 1 \ , \quad \int_{-\infty}^{\infty} x^2 q(x)dx = D_2 \ , \quad \int_{-\infty}^{\infty} x^4 q(x)dx = D_4 \ ,$$
$$\int_{-\infty}^{\infty} xq(x)dx = \int_{-\infty}^{\infty} x^3 q(x)dx = 0 \ .$$

Поскольку постоянные D_2 и D_4, входящие в эти условия, вообще говоря, неизвестны, необходимо поставить некоторые дополнительные условия. Мы потребуем, во-первых, чтобы функция $q(x)$ была всюду непрерывной вместе со своей производной, и во-вторых, чтобы она была отлична от нуля лишь на конечном отрезке. Без ограничения общности можно считать, что $q(x) = 0$ при $|x| > 1$.

При этих предположениях, используя метод множителей Лагранжа, найдем, что условный минимум функционала (2.1) при условиях (1.4), (2.3), (2.4) достигается на функциях вида

$$q(x) = \begin{cases} (\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4)/2 \ , & |x| \le 1 \ , \\ 0 \ , & |x| > 1 \ , \end{cases} \tag{5.1}$$

где α_0, α_1, α_2, α_3 и α_4 — множители Лагранжа. Из (2.4) следует, что $\alpha_1 = \alpha_3 = 0$, а из (2.3) находим, что α_0, α_2 и α_4 удовлетворяют следующей системе уравнений:

$$\begin{aligned} &\alpha_0 + \alpha_2/3 + \alpha_4/5 = 1 \ , \\ &\alpha_0/3 + \alpha_2/5 + \alpha_4/7 = D_2 \ , \\ &\alpha_0/5 + \alpha_2/7 + \alpha_4/9 = D_4 \ . \end{aligned} \tag{5.2}$$

Из условия непрерывности функции (5.1) и ее производной в точках $x = \pm 1$ получаются еще два уравнения:

$$\begin{aligned} &\alpha_0 + \alpha_2 + \alpha_4 = 0 \ , \\ &\alpha_2 + 2\alpha_4 = 0 \ . \end{aligned} \tag{5.3}$$

Решая совместно системы уравнений (5.2) и (5.3), находим $\alpha_0 = \alpha_4 = 15/8$ и $\alpha_2 = -15/4$. В результате получаем окончательный вид функции (5.1):

$$q(x) = \begin{cases} (15/16)\left[1 - x^2\right]^2 \ , & |x| \le 1 \ , \\ 0 \ , & |x| > 1 \ . \end{cases} \tag{5.4}$$

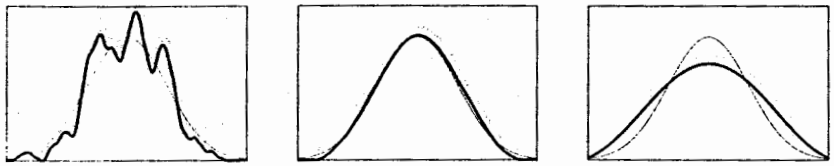

Рис. 1: Результаты численных экспериментов по выбору постоянной c в формуле (2.7). График эмпирической плотности $f_{2n}^{(*)}(x)$ (жирная линия) при изменении c меняется следующим образом. При $c = 1$ график эмпирической плотности случайным образом колеблется около графика теоретической плотности (слева). При $c = 7$ график эмпирической плотности чрезмерно сглажен (справа). Выбор $c = 4$ (в центре), таким образом, является оптимальным.

При этом $D_2 = 1/7$, $D_4 = 1/21$ и $d^2 = 5/7$.

Рассмотрим теперь вопрос о выборе постоянной c в формуле (2.7). Заметим, что при фиксированном значении n систематическая составляющая погрешности эмпирической плотности (как построенной по формуле (1.2), так и уточненной по формуле (3.1)) убывает по абсолютной величине при уменьшении величины c. Случайная же составляющая этой погрешности убывает при увеличении c. В результате при малом значении c в погрешности преобладает случайное слагаемое (при этом график эмпирической плотности имеет множество случайных максимумов и минимумов), а при большом c доминирует систематическая ошибка (график эмпирической плотности при этом получается чрезмерно сглаженным). В связи с этим относительно выбора c можно дать следующую рекомендацию. Построив эмпирические плотности для нескольких различных значений c, следует выбрать минимальное значение c, при котором на графике соответствующей эмпирической плотности отсутствуют случайные колебания. Ниже мы проиллюстрируем процедуру выбора c на конкретном примере.

6. ПРИМЕР

Рассмотрим здесь задачу об оценке плотности по выборке из нормального распределения. То есть пусть искомая плотность имеет вид

$$f(x) = \left(1/\sqrt{2\pi}\right) \mathrm{e}^{-x^2/2} \ . \tag{6.1}$$

Выборку из этого распределения нетрудно получить, воспользовавшись одним из многочисленных генераторов случайных чисел на ЭВМ (о получении случайных величин на ЭВМ см., например, [11], с. 20–32).

Учитывая приведенные выше рекомендации, для оценки плотности (6.1) по формуле (1.2) мы выберем функцию $q(x)$ в виде (5.4). Положим $\alpha = 1/7$ (ниже мы представим результаты, полученные и с другими значениями α). Для выбора параметра c в формуле (2.7) построим эмпирические плотности $f_{2n}^{(*)}(x)$ при нескольких различных значениях этого параметра (для таких предварительных расчетов можно использовать выборки небольшого объема). Некоторые из

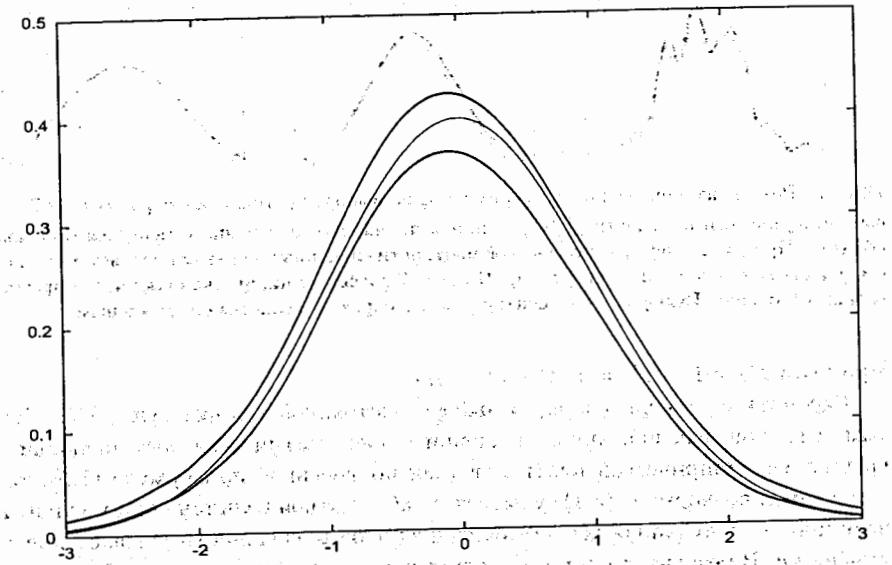

Рис. 2: Доверительная область (ее границы показаны жирными линиями), отвечающая коэффициенту доверия $\beta \approx 0.95$ и полному числу независимых наблюдений $2n = 2000$. График теоретической плотности практически не выходит за ее пределы.

этих численных экспериментов представлены на рисунке 1 (при этом $n = 50$). В результате выбрано $c = 4$. Подчеркнем здесь, что при других (бо́льших или меньших) значениях c сходимость эмпирической плотности к точной при $n \to \infty$ тоже имеет место, однако в этом случае для достижения заданной точности потребуется значительно большее n.

При таких $q(x)$, c и α положим теперь $n = 1000$ (полное число независимых наблюдений $2n = 2000$) и зададим коэффициент доверия $\beta \approx 0.95$. Отвечающая этим параметрам доверительная область показана на рисунке 2 (видно, что график теоретической плотности практически не выходит за ее пределы). Погрешности оценок при этом оказываются следующими:

$$\max_{-3\le x\le 3}\left|f_n^{(1)}(x) - f(x)\right| = 0.056\,, \qquad \max_{-3\le x\le 3}\left|f_n^{(2)}(x) - f(x)\right| = 0.053\,,$$
$$\max_{-3\le x\le 3}\left|f_{2n}^{(1+2)}(x) - f(x)\right| = 0.046\,, \qquad \max_{-3\le x\le 3}\left|f_{2n}^{(*)}(x) - f(x)\right| = 0.016\,.$$

Таким образом, оценка $f_{2n}^{(*)}(x)$, полученная в результате экстраполяции по n, точнее, чем каждая из оценок $f_n^{(1)}(x)$, $f_n^{(2)}(x)$ и $f_{2n}^{(1+2)}(x)$ по отдельности.

Рассмотрим теперь другие значения параметра α. Выше отмечалось, что имеет смысл выбирать α из интервала $1/9 \le \alpha \le 1/5$. Построим и сравним оценки $f_{2n}^{(*)}(x)$ и $f_{2n}^{(1+2)}(x)$ для крайних значений α из этого интервала ($q(x)$ и c при этом оставим прежними и положим $n = 1000$). Рисунок 3 иллюстрирует

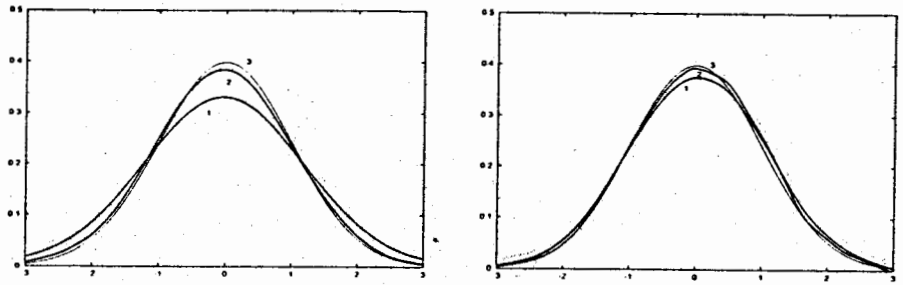

Рис. 3: Сравнение оценок $f_{2n}^{(1+2)}(x)$ (линия 1) и $f_{2n}^{(*)}(x)$ (линия 2), построенных для различных значений α (теоретическая плотность показана линией 3). При $\alpha = 1/9$ (слева) оценка $f_{2n}^{(*)}(x)$ обладает более высокой, чем $f_{2n}^{(1+2)}(x)$, скоростью сходимости, и в результате экстраполяции наблюдается заметное повышение точности. При $\alpha = 1/5$ (справа) скорости сходимости этих оценок одинаковы, экстраполяция не дает существенного уточнения, а лишь несколько уменьшает систематическую погрешность (но применение экстраполяции и в этом случае может быть целесообразным, поскольку дает обоснованную возможность построения доверительной области).

результаты этих численных экспериментов. Погрешности при этом, соответственно, следующие:

$$\max_{-3\le x\le 3}\left|f_n^{(1)}(x) - f(x)\right| = 0.079\,, \qquad \max_{-3\le x\le 3}\left|f_n^{(2)}(x) - f(x)\right| = 0.077\,,$$
$$\max_{-3\le x\le 3}\left|f_{2n}^{(1+2)}(x) - f(x)\right| = 0.069\,, \qquad \max_{-3\le x\le 3}\left|f_{2n}^{(*)}(x) - f(x)\right| = 0.017\,,$$

— для $\alpha = 1/9$, и

$$\max_{-3\le x\le 3}\left|f_n^{(1)}(x) - f(x)\right| = 0.027\,, \qquad \max_{-3\le x\le 3}\left|f_n^{(2)}(x) - f(x)\right| = 0.034\,,$$
$$\max_{-3\le x\le 3}\left|f_{2n}^{(1+2)}(x) - f(x)\right| = 0.024\,, \qquad \max_{-3\le x\le 3}\left|f_{2n}^{(*)}(x) - f(x)\right| = 0.019\,,$$

— для $\alpha = 1/5$.

В заключение покажем зависимость погрешностей оценок $f_{2n}^{(*)}(x)$ и $f_{2n}^{(1+2)}(x)$ от полного числа независимых наблюдений $2n$ при $\alpha = 1/9$ (как отмечалось, при таком α оценка $f_{2n}^{(*)}(x)$ имеет максимальный порядок сходимости). Эта зависимость приводится на рисунке 4.[4]

Рассмотренный здесь пример наглядно иллюстрирует эффективность предложенного метода для повышения точности оценки плотности распределения случайной величины по ее наблюдениям.

[4] Для построения графиков на рисунке 4 сначала была получена выборка из 12000 элементов, а затем из нее формировались выборки нужного объема: первые n элементов составляли первую выборку, следующие за ними n элементов — вторую. Если вместо этого для каждого n генерировать требующиеся выборки заново, то получаемые графики будут иметь бо́льшие случайные колебания, но в среднем будут убывать так же, как показано на рисунке 4.

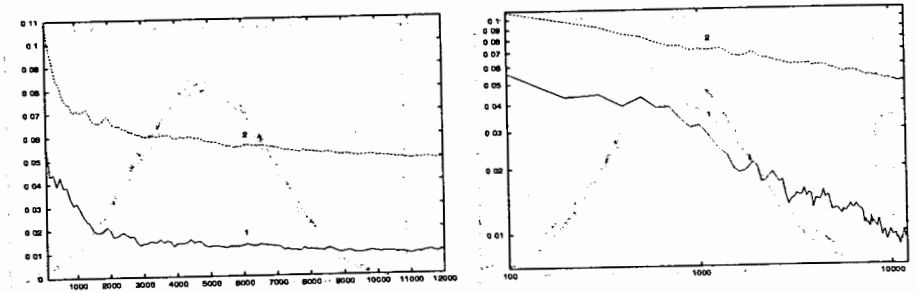

Рис. 4: Поведение погрешностей оценок $f_{2n}^{(*)}(x)$ (линия 1) и $f_{2n}^{(1+2)}(x)$ (линия 2) при увеличении полного числа независимых наблюдений $2n$. Они убывают, соответственно, как $n^{-4/9}$ и $n^{-2/9}$. Это хорошо видно в логарифмическом масштабе (справа).

ЗАКЛЮЧЕНИЕ

В настоящей работе для оценки неизвестной плотности распределения случайной величины по эмпирическим данным выбран класс оценок Парзена—Розенблатта. Получение этих оценок легко реализуется на ЭВМ и не занимает много счетного времени (как правило, затраты счетного времени не больше, чем при гистограммировании). Предложен метод, основанный на экстраполяции по числу независимых наблюдений и позволяющий повысить точность получаемой эмпирической плотности. Описано построение доверительной области, с заданной вероятностью покрывающей график теоретической плотности. Эффективность предложенного метода проиллюстрирована на модельной задаче.

СПИСОК ЛИТЕРАТУРЫ

1. *Жидков Е.П., Соловьев А.Г.* Повышение точности определения собственных значений и собственных функций краевой задачи на полуоси. // Ж. вычисл. матем. и матем. физ. 1999. Т. 39. N 7. С. 1098 - 1118.
2. *Жидков Е.П., Соловьев А.Г.* Уточнение приближенных решений краевой задачи на полупрямой. // Ж. вычисл. матем. и матем. физ. 1997. Т. 37. N 11. С. 1340 - 1344.
3. *Крамер Г.* Математические методы статистики. М.: Мир, 1975.
4. *Гливенко Г.И.* Курс теории вероятностей. М.: ГОНТИ, 1939.
5. *Смирнов Н.В.* О посторении доверительной области для плотности распределения случайной величины. Доклады АН СССР, 1950, **74**, N 2, 189–191.
6. *Гаскаров Д.В., Шаповалов В.И.* Малая выборка. М.: Статистика, 1978.
7. *Боровков А.А.* Математическая статистика. Новосибирск: Наука, 1997.
8. *Ибрагимов И.А., Хасьминский Р.З.* Асимптотическая теория оценивания. М.: Наука, 1979.
9. *Смирнов Н.В.* Теория вероятностей и математическая статистика. Избранные труды. М.: Наука, 1970.
10. *Гнеденко Б.В.* Курс теории вероятностей. М.: Наука, 1988.
11. *Соболь И.М.* Метод Монте-Карло. М.: Наука, 1968.

Рукопись поступила в издательский отдел
15 декабря 1999 года.

Жидков Е.П., Соловьев А.Г., Соснин А.Н.
P11-99-329

Повышение точности оценки неизвестной плотности
распределения случайной величины по эмпирическим данным

В работе изучается задача об оценке неизвестной плотности распределения случайной величины на основе результатов ее наблюдений. Для решения этой задачи выбран класс оценок Парзена—Розенблатта. Предлагается метод повышения точности эмпирической плотности, который основан на экстраполяции по числу независимых наблюдений. Эффективность метода демонстрируется на примере его применения в модельной задаче. Разработанная методика может применяться для оценок спектров вторичных частиц при моделировании процессов взаимодействия высокоэнергетических частиц с веществом, в особенности при определении характеристик нейтронных полей в моделируемых электроядерных установках.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна, 1999

Zhidkov E.P., Soloviev A.G., Sosnin A.N.
P11-99-329

Improving the Accuracy of Estimation of Unknown Random
Variable Probability Density over Empirical Data

A problem under study is to estimate the unknown probability density of a random variable basing on results of observations. Parzen—Rozenblatt estimates have been selected to solve the problem. A method improving the accuracy of the empirical density is suggested. It is based on extrapolation over a number of independent observations. Effectiveness of the method is demonstrated by an example of its application in a model problem. The method developed can be applied to estimate spectra of secondary particles while modelling processes of interactions of high energy particles in matter, in particular, while defining characteristics of neutron fields in electronuclear installations under study.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna, 1999