№ 12 2006

РАСЧЕТ И КОНСТРУИРОВАНИЕ МАШИН

620.10

АДАПТИВНЫЕ МЕТОДЫ ВОССТАНОВЛЕНИЯ ФУНКЦИИ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ

Д-р техн. наук, проф. В.Н. СЫЗРАНЦЕВ, канд. техн. наук Я.П. НЕВЕЛЕВ, д-р техн. наук, проф. С.Л. ГОЛОФАСТ

Рассмотрен алгоритм восстановления плотности распределения – основной характеристики закона распределения вероятности, на основе имеющейся выборки случайной величины, полученной в результате проведенных экспериментальных исследований или компьютерного моделирования, что является базовой проблемой при решении задач прочностной надежности отдельных элементов и оборудования в целом в вероятностном аспекте. Приведен пример моделирования напряжений, возникающих в стенке трубопровода, нагруженного внутренним избыточным давлением, являющимся случайной величиной, а также результаты восстановления функции плотности распределения напряжений, при расчете которых толщина стенки трубопровода и его размер приняты случайными.

The algorithm of restoring a density distribution as the main concept of distribution probability law, on the basis of available samples of the aleatory variable received as a result of lead experimental researches or a computer simulation that is a base task of problem solving concerning strength reliability of separate elements and the equipment as a whole in probability aspect is examined. The given example is based on voltages simulation arising in a pipeline wall because of its internal overpressure which is an aleatory variable, and also results of restoring a frequency function of voltages distribution when the width of the wall and its size are random.

Решаемые в теориях вероятности и математической статистики задачи соотносятся между собой как прямые и обратные [1]. В рамках теории вероятности задачи формулируются следующим образом: для известного состава генеральной совокупности и известного закона распределения вероятности для заданной схемы проведения экспериментальных исследований оценить вероятность результатов эксперимента. Теория математической статистики направлена на решение задачи, обратной к рассмотренной выше: на основе результатов проведенных экспериментов восстановить закон распределения вероятности. Исчерпывающей характеристикой закона распределения вероятности является ее плотность. Знание плотности распределения вероятности позволяет решать все основные задачи статистического анализа. Восстановление плотности распределения вероятности на основе имеющейся выборки случайной величины, полученной в результате проведенных экспериментальных исследований или компьютерного моделирования, является базовой проблемой при решении задач прочностной надежности элементов и объектов нефтегазового оборудования.

Обозначим через x_i, $i = 1, N$ выборку случайной величины X, являющуюся исходной для решения задачи определения плотности распределения вероятности $P(X)$. Если искомая функция $P(X)$ известна с точностью до конечного числа параметров, то задача восстановления плотности в подавляющем большинстве случаев является корректно поставленной [1] и для ее решения используются методы параметрической статистики [1, 2]. В общем случае класс функций, к которому может принадлежать $P(X)$, может быть весьма широким. По физическому смыслу x_i, $i = 1, N$ единственным требованием к $P(X)$ является ее непрерывность. В этом случае для восстановления $P(X)$ используется следующий подход.

Из теории вероятности известно, что плотность распределения вероятности $P(X)$ связана с функцией распределения вероятности $F(y) = \Pr\{X \le y\}$ интегральным соотношением

$$\int_{-\infty}^{y} P(x)dx = F(y), \tag{1}$$

которое можно представить в форме

$$\int_{-\infty}^{\infty} \theta(y-x)P(x)dx = F(y), \tag{2}$$

где $\theta(s) = \begin{cases} 1, \text{ при } s \ge 0, \\ 0, \text{ при } s < 0. \end{cases}$ — функция единичного скачка (функция Хевисайда).

При условии непрерывности функции $P(X)$ решение (2) является единственным [3].

Рассмотрим эмпирическую кумулятивную функцию распределения вероятности $F_N(y)$. Если величина y превосходит k элементов выборки $x_i, i = \overline{1,N}$ объемом N, функция $F_N(y)$ имеет вид

$$F_N(y) = \frac{1}{N}\sum_{i=1}^{N} \theta(y - x_i). \tag{3}$$

Известно [3], что эмпирическая функция распределения $F_N(y)$ является оптимальной непараметрической оценкой в каждой точке y для теоретической функции распределения $F(y)$. Следуя центральной теореме математической статистики, с ростом объема выборки N функция $F_N(y)$ с вероятностью единица равномерно приближается к $F(y)$

$$\Pr\left\{\lim_{N\to\infty} \sup_{y} |F_N(y) - F(y)| = 0\right\} = 1. \tag{4}$$

В реальной ситуации обработки данных экспериментальных исследований правая часть уравнения (2), - функция распределения $F(y)$, заменяется эмпирической функцией распределения $F_N(y)$, найденной на основе выборки ограниченного объема. Поэтому решение уравнения (2) всегда будет приближенным. Для восстановления функции плотности распределения путем решения уравнения (2) в рамках теории непараметрической статистики разработаны специальные процедуры [3], обеспечивающие с ростом N сходимость последовательности решений уравнения (2) к искомой плотности вероятности $P(X)$ и учитывающие некорректность постановки задачи (2), связанную с необходимостью дифференцирования неточно заданной ее правой части уравнения (2).

Выше было отмечено, что требованием при решении уравнения (2) к функции $P(X)$ является ее непрерывность. Помимо этого, принимая во внимание физический смысл функции плотности распределения, она на всей оси изменения X : от ——∞ до +∞ имеет только положительные значения и удовлетворяет условию

$$\int_{-\infty}^{+\infty} P(x)dx = 1. \tag{5}$$

Для восстановления неизвестной функции плотности распределения в рамках теории непараметрической статистики разработан ряд методов и алгоритмов [3]: «гребенка», метод гистограмм, метод ближайших соседей, метод Парзена—Розенблатта, метод разложения по базисным функциям и другие. В то же время практика решения технических задач свидетельствует, что в подавляющем большинстве случаев для восстановления функции плотности используется метод гистограмм.

Пусть $F(y)$ - непрерывная функция, тогда функция плотности распределения $P(y) = F'(y)$. Если в качестве оценки $F(y)$ используется $F_N(y)$, то в качестве оценки $P(y)$ может быть принята функция $P_N(y) = F'_N(y)$. Естественно, в зависимости от вида функции $F_N(y)$ и ее наполнения априорной информацией получаемое представление $P_N(y)$ будет различным не только по форме, но и по содержанию.

При гистограммном методе оценки плотности распределения применяется разностная аппроксимация $P(y)$ в виде

$$P(y) = F'(y) = \lim_{h\to 0}\frac{F(y+h)-F(y)}{h} \approx \frac{F(y+h)-F(y)}{h},$$

а в качестве оценки функции $P(y)$ используется зависимость

$$P_N(y) = \frac{F_N(y+h)-F_N(y)}{h} = \frac{1}{Nh}\sum_{i=1}^{N}\left[\theta(y+h-x_i)-\theta(y-x_i)\right] = \frac{\nu_y}{Nh}, \qquad (6)$$

где ν_y — количество выборочных значений, попавших в интервал $(y; y + h]$.

Алгоритм восстановления плотности распределения на основе (6) заключается в следующем. Имея x_i, $i = \overline{1, N}$ выборку случайной величины X, устанавливаем интервал ее наблюдения $[a, b]$, где $a = \min_i(x_i)$, $b = \max_i(x_i)$, который разбивается на m непересекающихся интервалов $H_1, H_2, \ldots H_m$, каждый шириной h, и подсчитывается число выборочных значений ν_i, попавших в интервал H_i. На основе полученных величин ν_i оценка плотности распределения (в виде гистограммы) описывается следующим образом:

$$P_N(y) = \frac{\nu_i}{Nh}, \quad y \in H_i. \qquad (7)$$

Как показывает анализ [3], если ширина интервала h не стремится к нулю, оценка (7) является смещенной. Для уменьшения смещения необходимо увеличивать число интервалов m, что в реальной ситуации обработки экспериментальных данных по выборкам сравнительно небольшого объема далеко не всегда возможно. Более того, в работе [3] показано, что минимум вариации гистограммы достигается при вполне определенной величине h, зависящей как от объема выборки N, так и вида восстанавливаемой функции $P(y)$. То есть в практическом плане, при отсутствии дополнительной информации о виде функции $P(y)$, критерии рационального разбиения выборки на интервалы для реализации метода гистограмм не определены. Естественно, дальнейшее использование восстановленной по данному алгоритму плотности распределения возможно только с большой осторожностью. Примером изменения гистограмм плотности распределения при вариации интервалов разбиения является рис. 1, на котором показаны результаты обработки выборки случайной величины x_i, $i = \overline{1,46}$, а также гистограмма, соответствующая выборке этой же случайной величины длиной $N = 2000$. Нетрудно видеть, что для выборки длиной $N = 46$

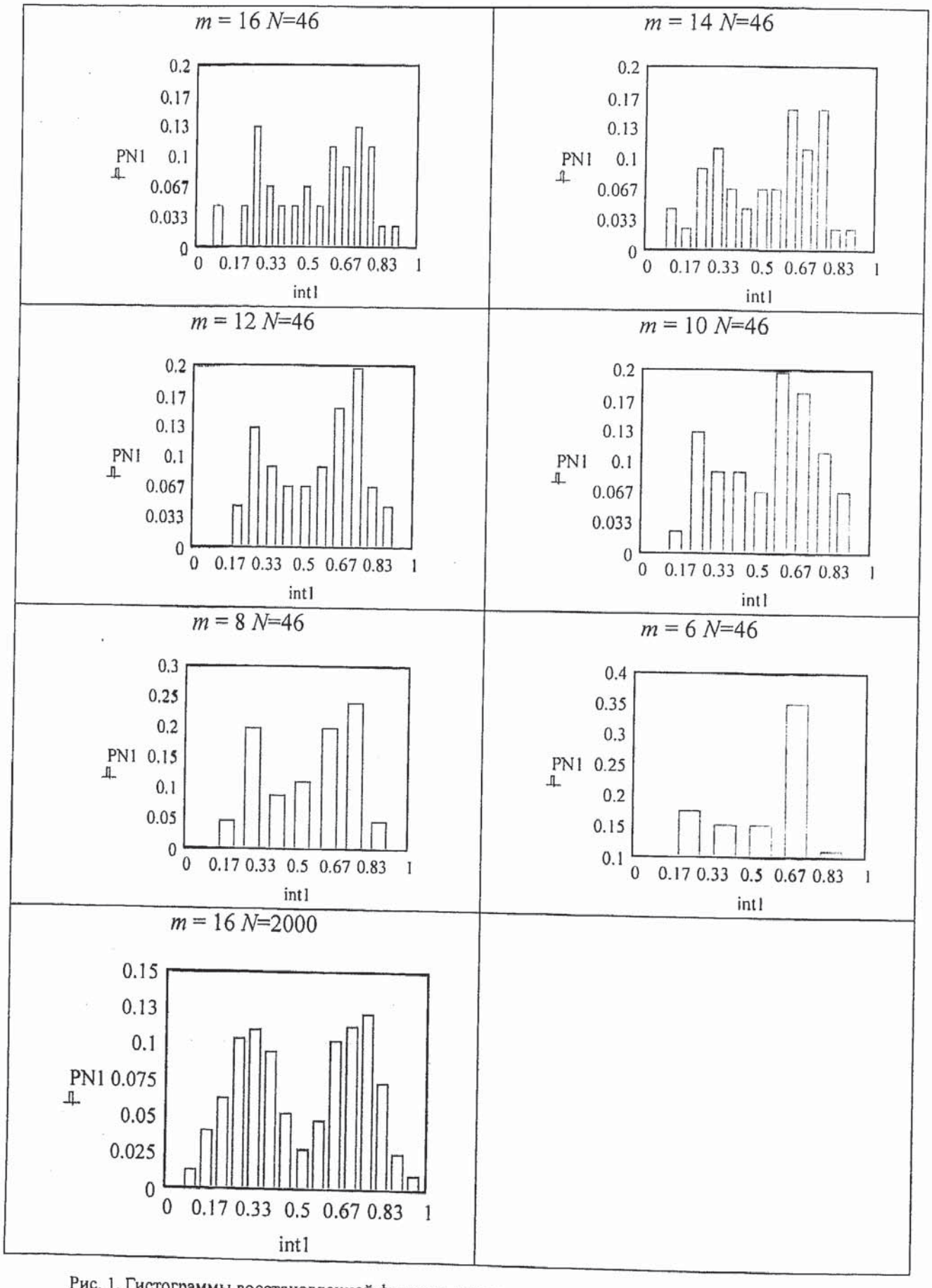


Рис. 1. Гистограммы восстановленной функции плотности распределения случайной величины

какие-либо выводы о виде функции $P(y)$ при изменении числа разбиений на интервалы сделать сложно, в то же время как на самом деле эта функция имеет вполне определенные закономерности. Оценка плотности распределения с использованием гистограмм является при решении технических задач весьма распространенным методом. Это, по-видимому, может быть объяснено лишь его простой реализацией. Применение этого метода в задачах прогнозирования надежности нефтегазового оборудования сопряжено с риском получения данных о вероятности безотказной работы, далеких от реальных.

Повысить степень гладкости получаемой оценки функции плотности распределения позволяют методы, предложенные Парзеном и Розенблаттом [1,3]. В них используется сглаженная эмпирическая функция распределения в виде

$$F_N(y) = \frac{1}{N}\sum_{i=1}^{N} G\left(\frac{y - x_i}{h_N}\right), \tag{8}$$

где $G(t)$ — монотонно неубывающая функция от 0 до 1 своего аргумента, при этом $G(t) = 1 - G(-t)$, т. е. $G(t)$ — функция, симметричная относительно нуля; h_N — параметр размытости.

После дифференцирования (8) имеем:

$$P_N(y) = F'_N(y) = \frac{1}{Nh_N}\sum_{i=1}^{N} G'\left(\frac{y - x_i}{h_N}\right) = \frac{1}{Nh_N}\sum_{i=1}^{N} K\left(\frac{y - x_i}{h_N}\right), \tag{9}$$

где $K(t) = G'(t)$ — плотность распределения $G(t)$ или ядерная функция (ядро).

Теоретические исследования функции (8) свидетельствуют [3], что смещение и вариация оценки (9) зависят от вида ядра $K(t)$ и значения параметра размытости h_N. В работе [3] предложены различные зависимости, которые можно использовать в качестве ядерных функций (табл. 1).

Таблица 1

Функции, используемые в качестве ядерных [3]

Нормальное $K_1(t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)$	Лапласа $K_2(t) = \frac{1}{2}\exp\left(-\lvert t\rvert\right)$	Фишера $K_3(t) = \frac{1}{2\pi}\left(\frac{\sin\left(\frac{t}{2}\right)}{\frac{t}{2}}\right)$
Коши $K_4(t) = \frac{1}{\pi}\left(\frac{1}{1+t^2}\right)$	Логистическое $K_5(t) = \frac{e^{-t}}{\left(1+e^{-t}\right)^2}$	Епанчикова $K_6(t) = \frac{3\left(1-\frac{t^2}{5}\right)}{4\sqrt{5}}, \lvert t\rvert \le \sqrt{5}$
Равномерное $K_7(t) = \frac{1}{2}, \lvert t\rvert \le 1$	Треугольное $K_8(t) = 1 - \lvert t\rvert, \lvert t\rvert \le 1$	Квадратичное $K_9(t) = \frac{3\left(1-t^2\right)}{4}, \lvert t\rvert \le 1$

Восстановление функции плотности распределения методом Парзена—Розенблатта на основе (9) связано с решением двух задач. Первая заключается в выборе ядерной функции $K(t)$ из числа известных (табл.1 или других). Вторая задача связана с определением значения параметра размытости h_N.

Для выделения среди конечного числа функций $K(t)$ наиболее подходящей, необходимо иметь критерий отбора. В качестве такого критерия может быть принят информационный функционал вида [3]

$$J = \int \ln[K(t)]P(t)dt = \int \ln[K(t)]dF(t), \tag{10}$$

максимальное значение которого соответствует условию $K(t) = P(t)$.

Тогда поиск оптимальных h_N и $K(t) \in K = \{K_1(t),...K_9(t)\}$ сводится к решению следующей задачи:

$$(h_N^*, K^*(t)) = \arg \max_{h_N, K(t)} J_N(h_N, K(t)) = \arg \max_{h_N, K(t)} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ln \left[\frac{1}{(N-1)h_N} \sum_{j \neq i}^{N-1} K\left(\frac{x_i - x_j}{h_N} \right) \right] \right\}. \tag{11}$$

Как показано в работе [3], задача оценивания оптимальной величины h_N является более сложной, нежели исходная задача восстановления плотности распределения, поскольку оптимальное значение h_N зависит от неизвестной плотности распределения и, тем более, неизвестных ее производных. В практических приложениях весьма часто необходимы оценки плотности во вполне определенных областях, например, при решении задач прогнозирования ресурса и надежности, в первую очередь, важна оценка плотности на хвостах распределения. Поэтому при решении данной задачи необходимы алгоритмы, обеспечивающие получение оптимальных значений параметра h_N на основе имеющейся выборки x_i, $i = \overline{1, N}$ случайной величины X.

В литературе по статистике для ряда функций $K(t)$ из табл.1 получены зависимости для расчета оценок оптимальной величины h_N^* на основе различных оценок выборки x_i, $i = 1, N$. Например, при использовании ядерной функции в виде нормального распределения (табл.1):

$$K\left(\frac{y - x_i}{h_N}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{y - x_i}{h_N}\right)^2}{2} \right) \tag{12}$$

оптимальное значение параметра h_N^* («ширины окна»), задается формулой

$$h_N^* = D_N N^{-\frac{1}{5}}, \tag{13}$$

где D_N — выборочная дисперсия, рассчитываемая на основе имеющейся выборки значений x_i, $i = \overline{1, N}$,

$$D_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \frac{1}{N} \sum_{i=1}^{N} x_i \right). \tag{14}$$

В результате для оценки плотности с ядром (12) и параметром размытости (13) на основе (9) имеем выражение

$$P_N(y) = \frac{1}{\sqrt{2\pi} N h_N^*} \sum_{i=1}^{N} \exp\left(-\frac{\left(\frac{y - x_i}{h_N^*} \right)}{2} \right). \qquad (15)$$

Анализируя (15), нетрудно видеть, что реализация метода Парзена — Розенблатта предполагает два этапа расчета. На первом этапе рассчитывается грубая характеристика выборки —выборочная дисперсия (14), которая в дальнейшем, через параметр размытости h_N^*, используется для уточнения оценки плотности распределения (15). Поскольку величина D_N чувствительна к выбросам и не отражает характер изменения функции плотности (одномодальный, многомодальный), то извлекаемая с помощью D_N информация о плотности распределения может оказаться недостаточной для корректного решения задачи рассматриваемым методом.

Для решения задачи (11) на основе представленных в табл. 1 ядерных функций разработан комплекс программ в системе Mathcad. Работу комплекса проиллюстрируем на примере моделирования напряжений, возникающих в трубопроводе, нагруженном внутренним избыточным давлением g, являющимся величиной случайной. Для генерирования выборки случайной величины g использована функция плотности в виде тригонометрического ряда

$$P(y) = \sum_{j}^{6} \alpha_j \cos\left[\left(\frac{2j-1}{2} \right) y\pi \right], \qquad (16)$$

при $\alpha_1 = 1{,}27027$; $\alpha_2 = -0{,}85566$; $\alpha_3 = 0{,}07521$; $\alpha_4 = -0{,}52205$; $\alpha_5 = -0{,}31440$; $\alpha_6 = 0{,}43318$ имеющая два явных экстремума (рис.1 при $N = 2000$).

Разработанный на основе (16) датчик случайных чисел обеспечивал получение выборки $g_i, i = \overline{1, N}$ величины g в пределах от $g_{min} = 5$ МПа до $g_{max} = 8$ МПа.

При расчете напряжений σ в трубопроводе под действием g

$$\sigma = g(d - 2\delta)/(2\delta)$$

толщина его стенки δ, как и диаметр d, приняты случайными, распределенными, соответственно, по равномерному ($\delta_{min} = 19$ мм; $\delta_{max} = 21$ мм) и нормальному (среднее значение d = 1020 мм, среднеквадратичное отклонение 1 мм) законам распределения.

Результаты восстановления функции плотности распределения напряжений (решение задачи (11)) при использовании различных (из табл. 1) ядерных функций представлены (для $N = 500$) на рис.2. Здесь же показаны полученные значения функционала (10), из анализа величин которых следует, что наилучшей оценкой для данной выборки является ядерная функция с равномерным ядром. Рис. 3 иллюстрирует результаты решения задачи восстановления функции плотности распределения σ на основе ядерной оценки с нормальным ядром при вариации объема выборки $N = 50, 100, 200, 500, 1000$. В правой части рис. 3 приведены рассчитанные значения σ при вероятности 95% и 99%, а также вероятность появления напряжения величиной σ = 170 МПа при различных объемах выборки. Представленные результаты расчета свидетельствуют об эффективности и достоверности разработанного программного комплекса при обработке статистических данных.

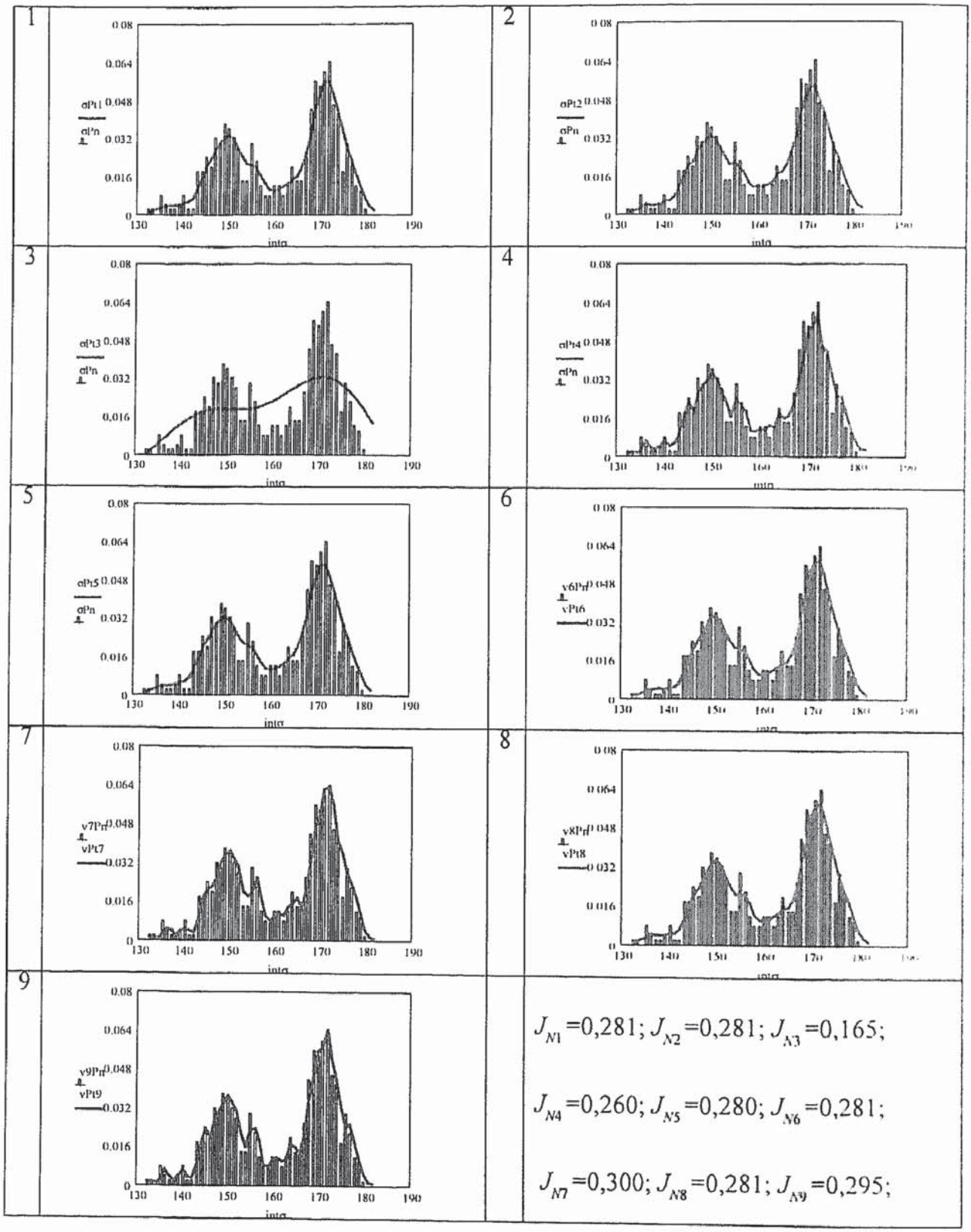


Рис. 2. Восстановленные функции плотности распределения случайной величины σ с использованием ядерных функций табл. 1

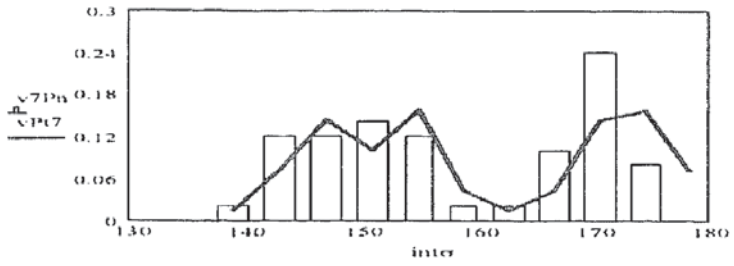
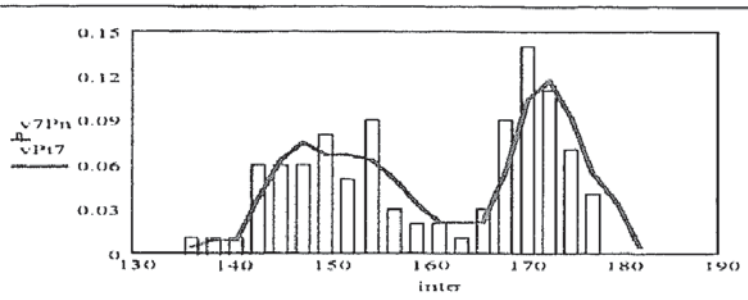
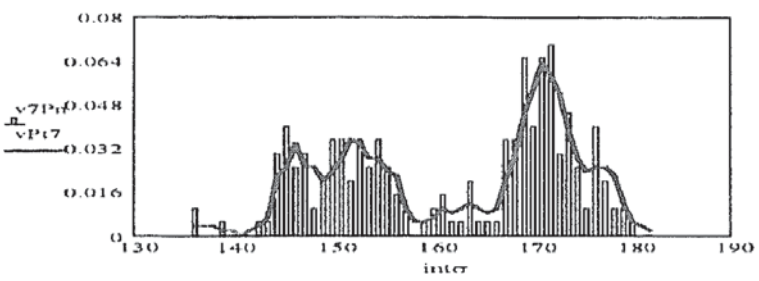
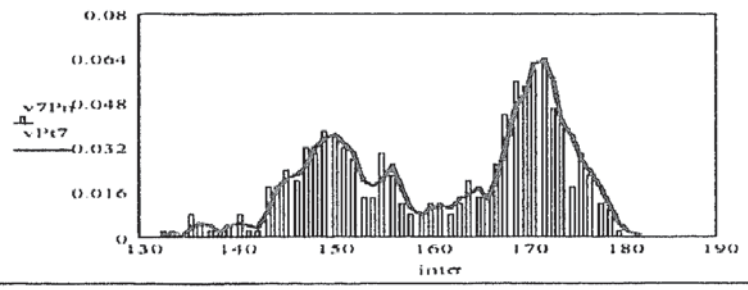
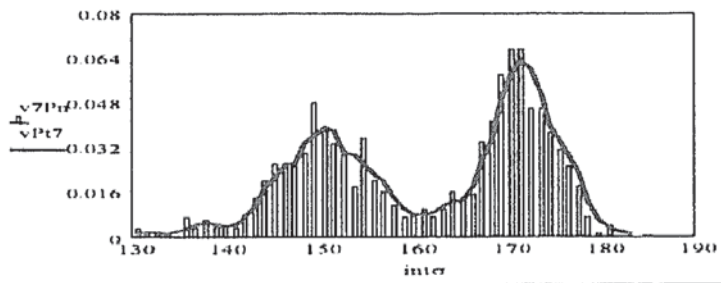
	N=50 m =10 h^*_{N7} =0,07 J_{N7}=0,114 $Pr(\sigma$ =170) =64,8 % $Pr(\sigma$ =177,8) = 95% $Pr(\sigma$ =181,0) = 99 %
	N=100 m =20 h^*_{N7} =0,06 J_{N7} = 0,236 $Pr(\sigma$ =170)=62,0 % $Pr(\sigma$ =177,9) = 95% $Pr(\sigma$ =181,8)= 99 %
	N=200 m =50 h^*_{N7} =0,03 J_{N7} =0,266 $Pr(\sigma$ =170) =62,6% $Pr(\sigma$ =177,4) = 95% $Pr(\sigma$ =179,4) = 99 %
	N=500 m =50 h^*_{N7} =0,025 J_{N7} =0,300 $Pr(\sigma$ =170) =64,2% $Pr(\sigma$ =176,8) = 95% $Pr(\sigma$ =179,1) = 99 %
	N=1000 m =50 h^*_{N7} =0,035 J_{N7}=0,350 $Pr(\sigma$ =170) =64,0% $Pr(\sigma$ =176,9)= 95% $Pr(\sigma$ =179,7) = 99 %

Рис. 3. Восстановление функции плотности распределения напряжений на основе ядерной оценки с равномерным ядром при различных объемах выборки

СПИСОК ЛИТЕРАТУРЫ

1. Деврой Л., Дьёрфи Л. Непараметрическое оценивание плотности. L_1 — подход: Пер. с англ. — М.: Мир, 1988. — 408 с.
2. Арасланов А. М. Расчет элементов конструкций заданной надежности при случайных воздействиях. — М.: Машиностроение, 1987. — 128 с.
3. Симахин В. А. Непараметрическая статистика. Ч.I. Теория оценок: Учебное пособие. — Курган: Изд-во Курганского гос. ун-та, 2004. — 207 с.