Применение численного вероятностного анализа в задачах интерполяции

О. А. Попова

Институт космических и информационных технологий СФУ, Красноярск, Россия

Контактный e-mail: `OlgaArc@yandex.ru`

Рассмотрено построение интерполяционных многочленов и сплайнов в условиях, когда входные данные носят случайный характер и могут быть представлены функциями, аппроксимирующими плотности вероятности этих случайных величин. Для решения поставленных задач мы развиваем численный метод, в основу которого положены компьютерная реализация арифметических операций над плотностями случайных величин и процедуры построения так называемых вероятностных расширений функций. Новый подход получил название *численного вероятностного анализа*. Исследовано, в частности, применение интерполяционных многочленов и сплайнов для построения гарантированных оценок функций распределения и функций плотности вероятности.

Ключевые слова: численный вероятностный анализ, функция случайных аргументов, интерполяция, гарантированные оценки.

Введение

Наличие неопределенностей и неоднозначностей во входных данных при решении многих практических задач приводит к необходимости создания методов, учитывающих эти неопределенности. Так, интервальная (или ограниченная) неопределенность, когда об интересующей величине известны лишь границы ее возможных значений, привела к развитию интервального анализа. Сами интервалы по определению не несут никакой дополнительной структуры (ни вероятностной, ни какой другой), и все их точки считаются равнозначными и имеющими одинаковый статус, или, как еще иногда говорят, *равновозможными*. Из-за этой концептуальной простоты методы интервального анализа получили наибольшее развитие среди дисциплин, занимающихся обработкой неопределенных и неточных данных. Дальнейшие обобщения интервального анализа направлены на получение более детальных характеристик получаемых множеств решений [1, 2].

С другой стороны, на практике часто оказывается, что не все точки полученных интервальных оценок (даже оптимальных по ширине) в самом деле равнозначны. Как правило, варианты решений, близкие к концам интервальных оценок, реализуются крайне редко или вообще никогда не случаются на практике. В последние годы развития интервального анализа появилось ощущение того, что более детальный учет структуры входных данных внутри интервалов их изменения позволит более точно оценивать интервалы оценок решений. В русле этих идей выполнена и настоящая работа. Мы используем информацию о плотности вероятности случайных величин внутри интервалов и развиваем численные методы для расчетов с такими величинами. Это дает возможность получать результаты в виде случайных величин, имеющих интервальные носители, с приближенно найденными плотностями вероятности.

Одним из подходов к решению задач со случайными входными данными является, как известно, метод Монте-Карло [3, 4]. При всех его положительных качествах этот метод обладает рядом недостатков, и один из самых существенных — низкая скорость сходимости. Так, в работе [5] сравнение трудоемкости метода Монте-Карло и численных операций с гистограммой арифметики для достижения одинаковой точности ответа показало преимущество гистограммной арифметики от 100 до 1000 раз. В целом при благоприятных обстоятельствах численные операции над плотностями вероятностей случайных величин позволяют существенно поднять точность расчетов при сравнительно небольшом объеме вычислений [6].

В работе рассматриваются вероятностные расширения интерполяционных полиномов и сплайнов на случай, когда входные данные представляют собой случайные величины, заданные своими плотностями вероятности. Далее случайные величины, по аналогии с интервальными переменными, будем обозначать $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$.

Задачу интерполяции можно сформулировать следующим образом: пусть для некоторой функции f ее значения $f_i = f(x_i)$ в точках $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ являются случайными величинами $\boldsymbol{f}_i$, $i = 0, 1, \ldots, n$, причем задана функция их совместной плотности вероятности $p(f_0, f_1, \ldots, f_n)$. Возникает задача приближенного восстановления всех реализаций функции f в произвольной точке x. Далее эта задача будет решена в рамках численного вероятностного анализа с использованием понятия вероятностного расширения. Для этих целей будут построены вероятностные расширения интерполяционных полиномов Лагранжа, кусочно-линейных функций, кубических сплайнов.

1. Элементы численного вероятностного анализа

Обозначим через $\boldsymbol{R}$ множество $\{\boldsymbol{x}\}$ случайных величин, заданных своими плотностями вероятности p_x, и соответственно через $\boldsymbol{R}^n$ пространство случайных n-векторов. Пусть имеется совокупность непрерывных случайных величин $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ с совместной плотностью распределения $p(x_1, x_2, \ldots, x_n)$, а случайная величина $\boldsymbol{y}$ связана с $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ функциональной зависимостью

$$\boldsymbol{y} = f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n).$$

Тогда функция распределения F_y для случайной величины $\boldsymbol{y}$ задается, как известно, выражением [7, 8]

$$F_y(z) = P(y < z) = \int\limits_{\Omega_z} p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n, \qquad (1)$$

где $\Omega_z = \{\, (x_1, x_2, \ldots, x_n) \mid f(x_1, x_2, \ldots x_n) < z \,\}$. В тех случаях, когда это возможно, дифференцируя F_y, можно получить плотность вероятности для $\boldsymbol{y}$:

$$\frac{dF_y(z)}{dz} = f_y(z).$$

Плотность вероятности f_y будем называть *вероятностным расширением* функции f.

Полагая $f(x_1, x_2) = x_1 * x_2$, где $* \in \{+, -, \cdot, /\}$, можно получить аналитические формулы для определения плотности вероятности результатов арифметических действий над случайными величинами. Например, для нахождения плотности вероятности $p_{x_1+x_2}$ суммы двух случайных величин $\boldsymbol{x}_1 + \boldsymbol{x}_2$ используется соотношение [8]

$$p_{x_1+x_2}(x) = \int_{-\infty}^{\infty} p(x - v, v)\, dv = \int_{-\infty}^{\infty} p(v, x - v)\, dv \tag{2}$$

(оно является сверткой плотностей вероятности слагаемых в случае их независимости). Для нахождения плотности вероятности p_{x_1/x_2} частного двух случайных величин $\boldsymbol{x}_1/\boldsymbol{x}_2$ имеем

$$p_{x_1/x_2} = \int_{0}^{\infty} vp(xv, v)\, dv - \int_{-\infty}^{0} vp(v, xv)\, dv. \tag{3}$$

Плотность вероятности $p_{x_1x_2}$ произведения двух случайных величин $\boldsymbol{x}_1\boldsymbol{x}_2$ определяется соотношением [7]

$$p_{x_1x_2} = \int_{0}^{\infty} (1/v)p(x/v, v)\, dv - \int_{-\infty}^{0} (1/v)p(v, x/v)\, dv. \tag{4}$$

Из этих выражений непосредственно вытекают коммутативность и ассоциативность операций сложения и умножения случайных величин.

Заметим, что для численного вычисления интегралов вида (1) использование кубатурных формул возможно только при небольших размерностях n, а в общем случае используются методы Монте-Карло [3]. В рамках численного вероятностного анализа рассматриваются численные арифметические операции над различными типами плотностей случайных величин. Это позволяет в некоторых случаях вычислять интегралы вида (1) с необходимой точностью.

В численном вероятностном анализе используются различные типы представления плотностей случайных величин: дискретные, гистограммы, полиграммы, полигоны, кусочно-полиномиальное и аналитическое представление. Далее подробно рассмотрим решение поставленных задач с помощью гистограмм.

Гистограммой называется случайная величина, плотность распределения которой представлена кусочно-постоянной функцией. Гистограмма P определяется некоторой сеткой $\{\, x_i \mid i = 0, \dots, n\}$, и на каждом подынтервале $[x_{i-1}, x_i]$, $i = 1, \dots, n$, гистограмма принимает постоянное значение p_i. Величины $h_i = (x_i - x_{i-1})$, $i = 1, 2, \dots, n$, обычно называют *шагами гистограммы* P, а $h = \max_{1 \le i \le n} h_i$ определяет ее максимальный шаг.

В символическом анализе данных и научном направлении, образно называемом Data Mining [9], гистограммы используются для исследования множества различных процессов и применяются для описания изменчивости количественных признаков. Использование гистограмм обусловлено прежде всего тем, что они позволяют достаточно точно представлять произвольные распределения.

Важно отметить, что, несмотря на свою простоту, гистограммы охватывают все возможные интервалы оценки плотности вероятности. Наиболее популярны из них гистограммы с постоянным шагом (т. е. с фиксированной шириной столбцов), которые широко используются на практике, полиграммы и дискретные плотности вероятности [6].

В случае неизвестных плотностей вероятности возможно использование *гистограмм второго порядка*, т. е. кусочно-гистограммных функций [10, 11].

1.1. Операции над плотностями вероятности случайных величин

В качестве примера рассмотрим арифметические операции над гистограммными переменными. Реализация арифметических операций основана на работе с $p(x,y)$ — совместной плотностью вероятности двух случайных величин x, y. Пусть p_z — гистограмма, приближающая плотность вероятности арифметической операции над двумя случайными величинами $x * y$, где $* \in \{+,-,\cdot,/\}$. Тогда значение гистограммы на интервале $[z_i, z_{i+1}]$ определяется по формуле [12]

$$p_i = \frac{1}{z_{i+1}-z_i} \int\limits_{\Omega_i} p(x,y)\, dx\, dy, \tag{5}$$

где $\Omega_i = \{(x,y) \mid z_i \le x * y \le z_{i+1}\}$.

Рассмотренный выше подход обобщается на случай большего числа переменных. Пусть требуется найти гистограмму p_z суммы

$$\boldsymbol{z} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \ldots + a_n\boldsymbol{x}_n$$

и пусть $p(x_1, x_2, \ldots, x_n)$ — плотность распределения вероятностей случайного вектора $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. Тогда вероятность попадания $\boldsymbol{z}$ в интервал (z_i, z_{i+1}) соответственно равна [8]

$$P(z_i < \boldsymbol{z} < z_{i+1}) = \int \cdots \int\limits_{\Omega_i} p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n,$$

где $\Omega_i = \{(x_1, x_2, \ldots, x_n) \mid z_i < a_1x_1 + a_2x_2 + \ldots + a_nx_n < z_{i+1}\}$, а p_{zi} имеет вид

$$p_{zi} = \frac{1}{z_{i+1}-z_i} \int \cdots \int\limits_{\Omega_i} p(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n.$$

Рассмотрим вопрос существования обратных элементов по сложению и умножению. Заметим, что для уравнения $\boldsymbol{a}+\boldsymbol{x} = 0$ существует решение, которое можно представить в виде совместной функции распределения $(\boldsymbol{a}, \boldsymbol{x})$:

$$p_{-a}(x_1, x_2) = \begin{cases} p_a(x_1), & \text{если } \ x_1 = -x_2, \\ 0, & \text{если } \ x_1 \ne -x_2, \end{cases}$$

где $p_a(x)$ — плотность вероятности случайной величины $\boldsymbol{a}$. Таким образом, $p_{-a}(x_1, x_2)$ определяет обратный элемент по сложению для $\boldsymbol{a}$. Аналогично несложно построить и обратный элемент по умножению.

1.2. Вероятностные расширения

Рассмотрим задачу нахождения закона распределения функции нескольких случайных аргументов, для которых распределения известны.

Пусть $f(x_1, \dots, x_n)$ — рациональная функция от случайных аргументов. Для вычисления гистограммы F распределения значений f заменим переменные x_1, x_2, …, x_n их гистограммными значениями, а арифметические операции — на соответствующие гистограммные. Полученную гистограмму F будем называть *естественным гистограммным расширением* функции f.

Теорема 1 ([13]). *Пусть $f(x_1, \dots, x_n)$ — рациональная функция, в выражении для которой каждая переменная встречается только один раз, и $x_1, \dots, x_n$ — независимые случайные величины. Тогда естественное гистограммное расширение аппроксимирует вероятностное расширение с точностью $O(h)$, где h — максимальный шаг гистограмм распределений аргументов.*

Рациональная функция

$$f(x, y, z) = x(y + z)$$

имеет равносильное представление

$$f(x, y, z) = xy + xz,$$

но только первое из этих представлений подпадает под условие теоремы 1. Как следствие, только для первого представления естественное гистограммное расширение будет аппроксимировать вероятностное с точностью $O(h)$. Таким образом, для численных операций сложения и умножения над гистограммами дистрибутивность не выполняется.

Теорема 1 легко обобщается на следующий случай.

Замечание 1 ([13]). *Пусть для функции $f(x_1, \dots, x_n)$ возможна замена переменных такая, что $f(z_1, \dots, z_k)$ — рациональная функция от переменных $z_1, \dots, z_k$, удовлетворяющая условиям теоремы 1, и z_i — функции от множества переменных $x_i, i \in Ind_i$, причем множества Ind_i попарно не пересекаются. Пусть для каждой z_i существуют вероятностные расширения. Тогда естественное расширение $f(z_1, \dots, z_k)$ будет аппроксимировать вероятностное с некоторой точностью.*

Пусть $f(x_1, x_2) = (-x_1^2 + x_1)\sin(x_2)$. Тогда, полагая $z_1 = (-x_1^2 + 1)$ и $z_2 = \sin(x_2)$, можно построить вероятностные расширения функций z_1 и z_2, а $f = z_1 z_2$ — рациональная функция, подпадающая под условия теоремы 1. Следовательно, ее естественное гистограммное расширение будет аппроксимировать вероятностное с некоторой точностью.

Рассмотрим случай, когда для $f(x_1, \dots, x_n)$ необходимо найти вероятностное расширение $\boldsymbol{f}$, но не удается построить замену переменных согласно замечанию 1. Пусть для определенности только x_1 встречается несколько раз. Заметим, что если подставить вместо случайной величины x_1 детерминированную t, то для функции $f(t, x_2, \dots, x_n)$ можно построить естественное вероятностное расширение.

Пусть t — дискретная случайная величина, аппроксимирующая x_1 следующим образом: t принимает значения t_i с вероятностью P_i, и пусть для каждой $f(t_i, x_2, \dots, x_n)$ можно построить естественное вероятностное расширение φ_i. Тогда вероятностное расширение $\boldsymbol{f}$ функции $f(x_1, \dots, x_n)$ можно аппроксимировать плотностью вероятности φ следующим образом [13]:

$$\varphi(\xi) = \sum_{i=1}^{n} P_i \varphi_i(\xi).$$

Для частного вида случайных функций нам понадобятся некоторые дополнительные понятия. Пусть случайная функция имеет вид

$$\boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i(x),$$

где $g_i \in C^m[a, b]$ — пространство функций, определенных на $[a, b]$ и имеющих на этом интервале m непрерывных производных. Тогда формальную производную от $\boldsymbol{f}(x)$ определим таким образом:

$$\partial^k \boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i^{(k)}(x), \quad k = 0, \ldots, m.$$

Далее запись $a \in \boldsymbol{a}$ будет означать, что a есть некоторая реализация случайной величины $\boldsymbol{a}$. Соответственно, функцию

$$f(x) = \sum_{i=1}^{n} a_i g_i(x), \quad a_i \in \boldsymbol{a}_i,$$

будем называть *сужением функции* $\boldsymbol{f}$ по параметрам $\boldsymbol{a}_i$.

2. Задачи интерполяции

2.1. Случайные полиномы Лагранжа

Задачу интерполяции сформулируем следующим образом: пусть для некоторой функции f в точках $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ известны аппроксимации $f(x_i)$ случайными величинами $\boldsymbol{f}_i$ и задана их совместная функция плотности вероятности $p(f_0, f_1, \ldots, f_n)$. Необходимо построить случайный интерполяционный полином $\boldsymbol{l}_n(x)$: $\boldsymbol{l}_n(x_i) = \boldsymbol{f}_i$.

Рассмотрим интерполяционные полиномы Лагранжа для случая линейной интерполяции. Пусть для некоторой функции f в точках x_1, x_2 известны значения f_1, f_2.

В случае линейной интерполяции имеем точное равенство

$$f(x) = l_1(x) + \frac{(x - x_1)(x - x_2)}{2} f''(\xi),$$

где l_1 — полином Лагранжа первой степени,

$$l_1(x) = f_1 \frac{x_2 - x}{x_2 - x_1} + f_2 \frac{x - x_1}{x_2 - x_1}, \quad \xi \in [x_1, x_2].$$

В случае, когда $f_1 \in \boldsymbol{f}_1$, $f_2 \in \boldsymbol{f}_2$ известны неточно, необходимо построить линейную случайную функцию $\boldsymbol{l}(x)$, такую что выполнены условия интерполяции $\boldsymbol{l}(x_1) = \boldsymbol{f}_1$ и $\boldsymbol{l}(x_2) = \boldsymbol{f}_2$. Таким образом, используя естественные вероятностные расширения, построим случайный полином Лагранжа первой степени

$$\boldsymbol{l}(x) = \boldsymbol{f}_1 \frac{x_2 - x}{x_2 - x_1} + \boldsymbol{f}_2 \frac{x - x_1}{x_2 - x_1}.$$

Заметим, что условия интерполяции выполнены и $\boldsymbol{l}(x)$ принимает соответствующие значения в узлах интерполяции. Отметим, что здесь сужение случайной линейной функции по параметрам $\boldsymbol{f}_i$ будет вещественной линейной функцией.

Далее, если требуется построить случайную функцию $\boldsymbol{l}$, для которой при всех $x \in [x_1, x_2]$ выполнено включение $f \in \boldsymbol{l}$, то необходимо знать априорную информацию о плотности вероятности $f'' \in \boldsymbol{f}''$ на интервале $[x_1, x_2]$. Заметим, что

$$f(x) = l(x) + \frac{(x - x_1)(x - x_2)}{2} f''(\xi).$$

Получаем оценку

$$f(x) \in \boldsymbol{l}(x) + \frac{(x - x_1)(x - x_2)}{2} \boldsymbol{f}''.$$

Рассмотрим интерполяционный полином Лагранжа в общем случае. Справедливо представление

$$\boldsymbol{l}_n(x) = \sum_{i=0}^{n} \boldsymbol{f}_i \omega_i(x),$$

где базисные функции

$$\omega_i(x) = \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)}.$$

Таким образом, вычисление интерполяционного полинома Лагранжа в произвольной точке сводится к вычислению суммы $\boldsymbol{f}_i$ с весами. Наиболее просто это осуществляется в случае независимости случайных величин $\boldsymbol{f}_i$, поскольку подпадает под условия теоремы 1.

2.2. Случайная кусочно-линейная интерполяция

При значениях $n > 5$ использование интерполяционных полиномов Лагранжа неэффективно. В этом случае для простоты можно использовать кусочно-линейную интерполяцию, при которой на каждом подынтервале $[x_i, x_{i+1}]$ реализуется полином Лагранжа первой степени:

$$\boldsymbol{l}_1(x) = \boldsymbol{f}_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + \boldsymbol{f}_{i+1} \frac{x - x_i}{x_{i+1} - x_i}.$$

Оценим математическое ожидание интерполяционных полиномов Лагранжа. В силу линейности математическое ожидание интерполяционного полинома будет линейной комбинацией математических ожиданий значений функции и совпадать с интерполяционным полиномом Лагранжа, проведенным через математические ожидания значений функции:

$$\mathsf{M}[\boldsymbol{l}_1(x)] = \mathsf{M}[\boldsymbol{f}_i] \frac{x_{i+1} - x}{x_{i+1} - x_i} + \mathsf{M}[\boldsymbol{f}_{i+1}] \frac{x - x_i}{x_{i+1} - x_i}.$$

В тех случаях, когда для математического ожидания случайной функции $\boldsymbol{f}$ известны оценки второй производной $\max\limits_{x \in [a,b]} |\mathsf{M}[\boldsymbol{f}^{(2)}]|$, справедлив следующий результат.

Теорема 2. *Пусть $\boldsymbol{l}_1$ — кусочно-линейная интерполяция случайной функции $\boldsymbol{f}$. Тогда справедлива оценка*

$$\left| \mathsf{M}[\boldsymbol{l}_1] - \mathsf{M}[\boldsymbol{f}] \right| \le K h^2 \max_{x \in [a,b]} \left| \mathsf{M}[\boldsymbol{f}^{(2)}] \right|,$$

где K — константа, не зависящая от h.

Доказательство повторяет аналогичное доказательство для вещественных интерполяционных полиномов.

Рассмотрим теперь свойства дисперсии кусочно-линейной интерполяции $\boldsymbol{l}_1$ случайной функции $\boldsymbol{f}$.

Теорема 3. *Дисперсия кусочно-линейной интерполяции $\boldsymbol{l}_1$ удовлетворяет следующей оценке:*

$$\mathsf{D}[\boldsymbol{l}_1] \ \le \max_{0\le i\le n-1} \max\big\{ \mathsf{D}[\boldsymbol{F}_i], \mathsf{D}[\boldsymbol{F}_{i+1}], K_{i,i+1}\big\}.$$

Доказательство. Дисперсия линейной функции l_1 на интервале $[x_i, x_{i+1}]$ выражается формулой [7]

$$\begin{aligned}\mathsf{D}[\boldsymbol{f}_i a_i + \boldsymbol{f}_{i+1}a_{i+1}] \ &= \mathsf{D}[\boldsymbol{f}_i]\, a_i^2 + \mathsf{D}[\boldsymbol{f}_{i+1}]\, a_{i+1}^2 + 2a_i a_{i+1} K_{i,i+1} \le \\ &\le \max\big\{\mathsf{D}[\boldsymbol{f}_i], \mathsf{D}[\boldsymbol{f}_{i+1}], K_{i,i+1}\big\}\, (a_i^2 + a_{i+1}^2 + 2a_i a_{i+1}),\end{aligned}$$

где

$$a_i = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \quad a_{i+1} = \frac{x - x_i}{x_{i+1} - x_i};$$

$K_{i,i+1}$ — корреляционный момент $\boldsymbol{f}_i$, $\boldsymbol{f}_{i+1}$. Поскольку $a_i + a_{i+1} = 1$, $a_i \in [0,1]$, то $a_i^2 + a_{i+1}^2 + 2a_i a_{i+1} = (a_i + a_{i+1})^2 = 1$.

2.3. Случайные сплайны

Рассмотрим вопросы построения случайных сплайнов. Они вводятся как вероятностные расширения соответствующих вещественных сплайнов по некоторым их параметрам. При небольшой априорной информации об интерполируемой функции строятся ее вероятностные приближения.

Приведем необходимые для дальнейшего изложения элементы теории сплайнов [14]. Пусть для некоторого целого $N \ge 2$ на интервале $[a,b]$ задана сетка

$$\omega = \{\, x_i \mid a = x_0 < x_1 < \ldots < x_N = b\,\}$$

с шагами $h_i = x_{i+1} - x_i$, $h = \max\limits_{0\le i\le N-1} h_i$.

Функцию s называют сплайном степени n дефекта k (k — целое, $1 \le k \le n$) с узлами из ω, если:

$$s(x) = \sum_{i=0}^{n} a_{ij}(x - x_j)^i \quad \text{на} \quad [x_j, x_{j+1}], \quad j = 0, \ldots, N-1; \tag{6}$$

$$s \in C^{n-k}[a,b]. \tag{7}$$

Множество сплайнов, удовлетворяющих этим условиям, обозначим через S_n^k.

Обратимся к вопросу интерполирования заданных функций сплайнами нечетных степеней. Положим целое $m = (n-1)/2$ и будем считать дефект $k \le m+1$.

Пусть $f \in C^{n-k}[a,b]$. Поставим задачу определения сплайна $s \in S_n^k$, интерполирующего функцию f в следующем смысле:

$$s^{(j)}(x_i) = f^{(j)}(x_i), \quad i = 1, \ldots, N-1, \quad j = 0, \ldots, k-1. \tag{8}$$

Дополнительно задается еще по $m+1$ краевому условию на $[a,b]$. Часто они берутся в виде

$$s^{(j)}(a) = f^{(j)}(a), \quad s^{(j)}(b) = f^{(j)}(b), \quad j = 0,\dots,m. \tag{9}$$

Условия (8), (9) приводят, соответственно, к $2(N-1)k$ и $(n+1)$ уравнениям для коэффициентов a_{ij}. Еще $(n-2k+1)(N-1)$ уравнений вытекают из условия непрерывности производных в соответствии с (7). В итоге получается система линейных алгебраических уравнений с квадратной матрицей $A \in \mathbb{R}^{N(n+1)\times N(n+1)}$, вектором неизвестных $a \in \mathbb{R}^{N(n+1)}$, известной правой частью $b \in \mathbb{R}^{N(n+1)}$. Не исследуя общего случая, в дальнейшем выпишем эту систему для конкретных примеров и исследуем ее разрешимость.

Теперь перейдем к определению случайных сплайнов от вещественного аргумента. Необходимость в их использовании возникает, когда вместо точных значений функции f и ее производных известны только какие-то случайные величины. Пусть даны случайные константы

$$f^{(j)}(x_i) \in \boldsymbol{f}_i^j, \quad i = 1,\dots,N-1, \quad j = 0,\dots,k-1; \tag{10}$$

$$f^{(j)}(a) \in \boldsymbol{f}_0^j, \quad f^{(j)}(b) \in \boldsymbol{f}_N^j, \quad j = 0,\dots,m. \tag{11}$$

Случайным сплайном назовем функцию $\boldsymbol{s} : [a,b] \to \boldsymbol{R}$ со случайными значениями, определяемую как

$$\begin{aligned}\boldsymbol{s}(x) = \{\, s(x) \mid s \in S_n^k,\ s^{(j)}(x_i) \in \boldsymbol{f}_i^j, \quad i = 1,\dots,N-1,\ j = 0,\dots,k-1;\\ s^{(j)}(a) \in \boldsymbol{f}_0^j, \quad s^{(j)}(b) \in \boldsymbol{f}_N^j, \quad j = 0,\dots,m \,\}.\end{aligned} \tag{12}$$

Эта случайная функция является вероятностным расширением соответствующего вещественного сплайна.

Для сплайнов известны следующие оценки [14].

Теорема 4. *Пусть $f \in C^p[a,b]$, $1 \le p \le n+1$, и сплайн s, $s \in S_n^k$, интерполирует функцию f в смысле (8), (9). Тогда при $0 \le j \le p$ имеет место*

$$\|f^{(j)} - s^{(j)}\|_\infty \le K_j h^{p-j} \|f^{(p)}\|_\infty, \tag{13}$$

где K_j — константы, не зависящие от f и h; $\|\cdot\|_\infty$ — чебышевская норма в пространстве функций.

Далее мы подробнее рассмотрим сплайны степени 3, т. е. кубические сплайны, наиболее популярные в современных вычислительных технологиях. Рассмотрим случайную функцию $\boldsymbol{s} : [a,b] \to \boldsymbol{R}$ со значениями

$$\boldsymbol{s}(x) = \{\, s(x) \mid s \in S_3^1, \quad s(x_i) \in \boldsymbol{f}_i^0, \quad i = 0,\dots,N, s''(a) \in \boldsymbol{f}_0^2, \quad s''(b) \in \boldsymbol{f}_N^2 \,\}.$$

В итоге кубический сплайн на подынтервалах $[x_{j-1}, x_j]$, $j = 1,\dots,N$, имеет два представления [14]:

$$\begin{aligned} s(x) = M_{j-1}(x_j - x)^3/(6h_j) + M_j(x - x_{j-1})^3/(6h_j) + \\ +(f_{j-1} - M_{j-1}h_j^2/6)(x_j - x)/h_j + (f_j - M_j h_j^2/6)(x - x_{j-1})/h_j, \end{aligned} \tag{14}$$

$$s(x) = m_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 - m_j(x - x_{j-1})^2(x_j - x)/h_j^2 +$$
$$+ f_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 + f_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3, \tag{15}$$

где $M_j = s''(x_j)$, $m_j = s'(x_j)$, $f_j = f(x_j)$.

Заменяя M_j, m_j, f_j на $\boldsymbol{M}_j$, $\boldsymbol{m}_j$, $\boldsymbol{f}_j \equiv \boldsymbol{f}_j^0$, получаем соответствующее представление случайного кубического сплайна. Выпишем случайные системы линейных алгебраических уравнений:

для $\boldsymbol{M}_j$

$$\begin{gathered}
\mu_j \boldsymbol{M}_{j-1} + 2\boldsymbol{M}_j + \lambda_j \boldsymbol{M}_{j+1} = \boldsymbol{D}_j,\\
\boldsymbol{M}_0 = \boldsymbol{f}_0^2, \quad \boldsymbol{M}_N = \boldsymbol{f}_N^2,\\
\boldsymbol{D}_j = 6((\boldsymbol{f}_{j+1} - \boldsymbol{f}_j)/h_{j+1} - (\boldsymbol{f}_j - \boldsymbol{f}_{j-1})/h_j)/(h_j + h_{j+1}),\\
\lambda_j = h_{j+1}/(h_j + h_{j+1}), \quad \mu_j = 1 - \lambda_j,\\
j = 1, \ldots, N-1;
\end{gathered} \tag{16}$$

для $\boldsymbol{m}_j$

$$\begin{gathered}
\lambda_j \boldsymbol{m}_{j-1} + 2\boldsymbol{m}_j + \mu_j \boldsymbol{m}_{j+1} = \boldsymbol{d}_j,\\
2\boldsymbol{m}_0 + \boldsymbol{m}_1 = 3(\boldsymbol{f}_1 - \boldsymbol{f}_0)/h_1 - h_1 \boldsymbol{f}_0^2/2,\\
2\boldsymbol{m}_N + \boldsymbol{m}_{N-1} = 3(\boldsymbol{f}_N - \boldsymbol{f}_{N-1})/h_N + h_N \boldsymbol{f}_N^2/2,\\
\boldsymbol{d}_j = 3\lambda_j(\boldsymbol{f}_j - \boldsymbol{f}_{j-1})/h_j + 3\mu_j(\boldsymbol{f}_{j+1} - \boldsymbol{f}_j)/h_{j+1},\\
j = 1, \ldots, N-1.
\end{gathered} \tag{17}$$

Матрицы систем (16) и (17) имеют вещественные элементы, а правые части содержат случайные переменные. Кроме того, эти матрицы являются трехдиагональными и имеют строгое диагональное преобладание.

Таким образом, в силу вещественности матриц (16), (17) решения $\boldsymbol{M}$ и $\boldsymbol{m}$ можно представить как линейные комбинации от правых частей

$$\boldsymbol{M} = B_1 \boldsymbol{f} \quad \text{или} \quad \boldsymbol{m} = B_2 \boldsymbol{f},$$

где B_1, B_2 — обратные матрицы к матрицам систем (16), (17).

С помощью найденных решений, используя вероятностные расширения, можно построить случайные функции $\boldsymbol{s}(x) : [a, b] \to \boldsymbol{R}$ со значениями

$$\begin{aligned}
\boldsymbol{s}(x) = {} & \boldsymbol{M}_{j-1}\left((x_j - x)^3/(6h_j) - (x_j - x)h_j/6\right) +\\
& + \boldsymbol{M}_j((x - x_{j-1})^3/6h_j) - (x - x_{j-1})h_j/6) +\\
& + \boldsymbol{f}_{j-1}(x_j - x)/h_j + \boldsymbol{f}_j(x - x_{j-1})/h_j
\end{aligned} \tag{18}$$

или

$$\begin{aligned}
\boldsymbol{s}(x) = {} & \boldsymbol{m}_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 -\\
& - \boldsymbol{m}_j(x - x_{j-1})^2(x_j - x)/h_j^2 +\\
& + \boldsymbol{f}_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 +\\
& + \boldsymbol{f}_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3
\end{aligned} \tag{19}$$

при $x \in [x_{j-1}, x_j]$, $j = 1, \ldots, N$.

Заметим, что если в (18) брать сужение $\boldsymbol{s}$ по случайным параметрам $\boldsymbol{M}_j$, то в общем случае мы не получим кубического сплайна, так как $\boldsymbol{M}_j$ могут не удовлетворять системе (16) и, как следствие, будем иметь $s' \notin C[a,b]$. Если брать сужение по $\boldsymbol{m}_j$ в (19), то $s' \in C[a,b]$, но $s'' \notin C[a,b]$.

Заметим, что сплайн на каждом подынтервале $[x_{j-1}, x_j]$ можно представить как линейную комбинацию

$$\boldsymbol{s}(x) = \sum_i \boldsymbol{f}_i \psi_i(x),$$

где $\psi_i(x)$ — некоторые вещественные функции. В этом представлении сужение $\boldsymbol{s}$ по случайным параметрам $\boldsymbol{f}_i$ не нарушает гладкости сплайна.

Рассмотрим случайные эрмитовы кубические сплайны $\boldsymbol{s}_1 : [a,b] \to \boldsymbol{R}$. На каждом интервале $[x_{j-1}, x_j], j = 1, \dots, N$, согласно формуле (15), эти сплайны представимы в виде

$$\begin{aligned} \boldsymbol{s}_1(x) = \boldsymbol{f}_{j-1} v((x - x_{j-1})/h_j) + \boldsymbol{f}^1_{j-1} w((x - x_{j-1})/h_j) + \\ + \boldsymbol{f}_j v((x - x_j)/h_j) + \boldsymbol{f}^1_j w((x - x_j)/h_j), \end{aligned}$$

где $v(x) = (|x| - 1)^2(2|x| + 1)$, $w(x) = x(|x| - 1)^2$. Кроме того, сужение $\boldsymbol{s}_1$ по параметрам $\boldsymbol{f}_j$, $\boldsymbol{f}^1_j$ будет эрмитовым кубическим сплайном. Следовательно, при аппроксимации функций эрмитовыми сплайнами наличие ошибок в данных не снижает гладкости сплайна.

3. Гарантированные оценки функции распределения

Рассмотрим построение гарантированных оценок эмпирической функции распределения.

Пусть $(\xi_1, \dots, \xi_n)$ — выборка случайной величины X с функцией распределения $F(t), t \in [a,b]$. Соответствующая эмпирическая функция распределения находится, как известно, следующим образом:

$$F_n(t) = \frac{m_t}{n}, \tag{20}$$

где m_t — число элементов, удовлетворяющих $\xi_i < t$.

Пусть $z_i = F(\xi_i)$, $i = 1, \dots, n$. Заметим, что z_i, $i = 1, \dots, n$, — равномерно распределенные случайные величины на $[0,1]$. Если $z_1 \le z_2 \le \dots \le z_n$, то z_k — k-я порядковая статистика с математическим ожиданием $\mathsf{M}[z_k] = k/(n+1)$ [15]. Далее будем использовать точки $(\xi_i, i/(n+1))$ для построения подходящей аппроксимации функции распределения F.

Пусть $\omega = \{a < \xi_1 < \xi_2 < \dots < \xi_n < b\}$ — сетка на $[a,b]$. На этом интервале, используя сетку ω, построим кусочно-линейную функцию s, такую что

$$\begin{gathered} s(\xi_i) = i/(n+1), \quad i = 1, \dots, n, \\ s(a) = 0, \quad s(b) = 1. \end{gathered}$$

Заметим, что если вместо математических ожиданий $i/(n+1)$ можно было бы использовать точные значения z_i, то погрешность кусочно-линейной функции $s(x)$ на сетке $\{\, \xi_i, i = 0, 1, \dots, n \,\}$ с $h = \max\limits_{1 \le i \le n} (\xi_i - \xi_{i-1})$ удовлетворяла бы оценке

$$\|F - s\| \le K h^2 \|F^{(2)}\|.$$

Таким образом, даже при относительно небольших n построенные оценки достаточно хорошо аппроксимируют функцию распределения F. Относительно z_i известно, что они образуют порядковые статистики.

Плотность вероятности k-й порядковой статистики

$$p_k(z) = \frac{n!}{(n-k)!(k-1)!}\, z^{k-1}(1-z)^{n-k}, \quad z \in [0,1].$$

Совместная плотность вероятности вектора (z_j, z_k) выражается следующим образом:

$$p_{j,k}(z_j, z_k) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!}\, z_j^{j-1}(z_k - z_j)^{k-j-1}(1-z_k)^{n-k},$$

$$j < k, \quad 0 \le z_j \le z_k \le 1.$$

Каждому случайному вектору $(z_1, z_2, \ldots, z_n)$ соответствует кусочно-линейная функция s. Перебирая все возможные случайные векторы $(z_1, z_2, \ldots, z_n)$, получаем все множество кусочно-линейных функций $\{s\}$. Заметим, что $\{s\}$ содержит интерполянт функции распределения F. Таким образом, используя для значения в узле ξ_k плотность вероятности k-й порядковой статистики, множество $\{s\}$ можно представить в виде случайной кусочно-линейной функции $\boldsymbol{L}$. Соответственно, $\boldsymbol{L}$ — гарантированная оценка эмпирической функции распределения.

Рассмотрим аппроксимацию случайной кусочно-линейной функции $\boldsymbol{L}$ в виде гистограммной функции распределения [16]. Построим две сетки $\omega_y = \{y_i = i/N_y, i = 0, 1, \ldots, N_y\}$ и $\omega_t = \{0 = t_0 < t_1 < \ldots < t_{N_t}\}$. Не ограничивая общности, будем считать $\omega \subset \omega_t$. На сетке ω построим гистограммную кусочно-линейную функцию. Для значений в узле ξ_k будем использовать плотность вероятности k-й порядковой статистики. Поскольку известны совместные плотности вероятности для порядковых статистик, используя численные операции над плотностями случайных величин, построим на сетке ω_y гистограммные значения h_t в узлах сетки ω_t.

Рассмотрим сетку $\Omega_h = \omega_t \times \omega_y$. Далее, используя гистограммы h_t, несложно в центре каждой ячейки сетки Ω_h восстановить среднее значение. Таким образом, на сетке Ω_h получаем кусочно-постоянную аппроксимацию, где H_{ij} — матрица значений в ячейках сетки.

Перепишем матрицу H_{ij} в виде списка гистограммной функции распределения (h_j, p_j), где h_j — гистограммы, образованные значениями H_{ij} при фиксированном j.

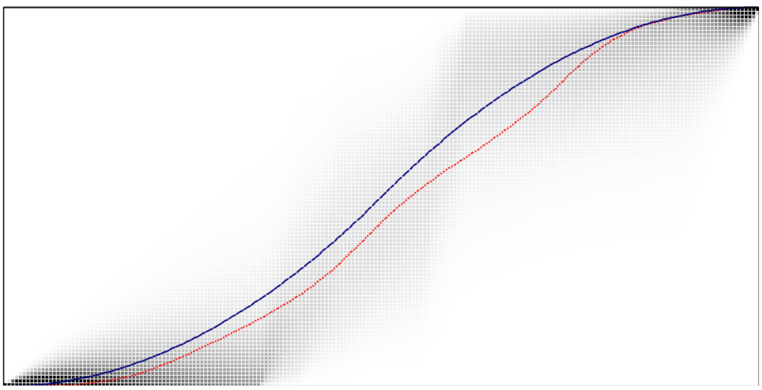

Рис. 1. Гарантированная оценка функции распределения

На рис. 1 представлена кусочно-постоянная аппроксимация гарантированной оценки эмпирической функции распределения, размерность выборки 7. Оттенками серого представлены плотности вероятности, сплошной линией показана точная функция распределения, точечная линия — математические ожидания гистограмм h_j.

4. Гарантированные оценки функции плотности вероятности

Основным недостатком кусочно-линейных функций является невозможность вычислить от них производные и получить оценки функции плотности вероятности. Рассмотрим использование эрмитовых сглаживающих сплайнов для построения аппроксимаций функций распределения.

Пусть $(\xi_1, \ldots, \xi_n)$ — выборка случайной величины X с функцией распределения $F(t)$, $t \in [a, b]$, сетка $\{x_i, i = 0, 1, \ldots, N\}$, $x_0 = a$, $x_N = b$, φ_l — базис в пространстве сплайнов Эрмита. Будем искать сплайн в виде

$$s(x) = \sum_{l=1}^{m} s_l \varphi(x)$$

с граничными условиями

$$s'(a) = 0, \quad s(a) = 0, \quad s'(b) = 0, \quad s(b) = 1.$$

Для нахождения неизвестных параметров s_l будем использовать метод наименьших квадратов

$$\sum_{i=1}^{n} (s(\xi_i) - z_i)^2 + \alpha \|s''\|^2 \to \min,$$

где $\alpha \geq 0$ — параметр. Заметим, что построение кубического сплайна сводится к решению систем линейных алгебраических уравнений с трехдиагональной матрицей.

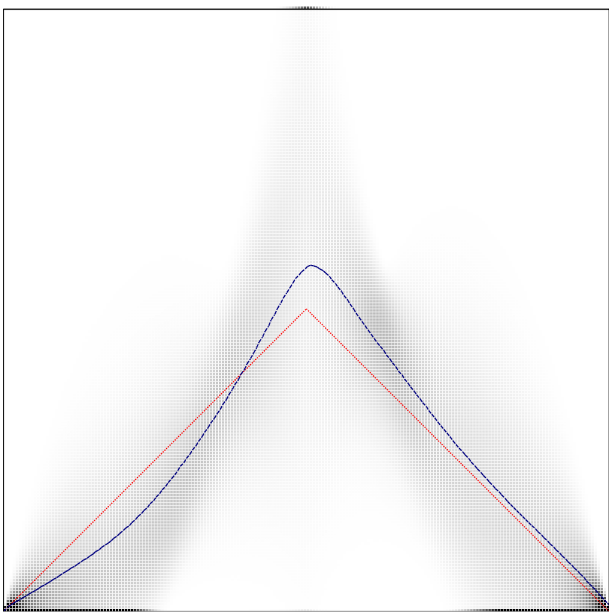

Рис. 2. Гарантированная оценка функции плотности вероятности

Таким образом, подобно описанному выше для точного вектора $(z_1, z_2, \dots, z_n)$ получаем оценку точности (13). Продифференцировав сплайн s, получаем аппроксимацию функции плотности вероятности и, используя плотности вероятности порядковых статистик, строим гистограмму второго порядка [6]. На рис. 2 представлена гистограмма второго порядка, приближающая треугольную плотность вероятности (точечная линия). Исходная выборка имеет размерность 7. Сплошной линией показано математическое ожидание гистограммы второго порядка.

В работе показано использование случайных интерполяционных многочленов и сплайнов для построения гарантированных оценок эмпирических функций распределения и функций плотности вероятности.

Список литературы / References

[1] **Добронец Б.С.** Приближения множеств решений параметрическими множествами // Журн. Сиб. федер. ун-та. Математика и физика. 2009. № 2(3). С. 305–311.
Dobronets, B.S. Approximation of the sets of solutions by parametric sets // SibFU J. Mathematics & Physics. 2009. No. 2(3). P. 305–311. (In Russ.)

[2] **Шарый С.П.** Конечномерный интервальный анализ: Электронная книга. Новосибирск: ИВТ СО РАН, 2016. 617 с. Адрес доступа:
`http://www.nsc.ru/interval/Library/InteBooks`
Shary, S.P. Finite-dimensional Interval Analysis. Novosibirsk: IVT SO RAN, 2016. 617 p. Available at: http://www.nsc.ru/interval/Library/InteBooks (In Russ.)

[3] **Соболь И.М.** Численные методы Монте-Карло. М.: Наука, 1973. 312 с.
Sobol, I.M. Numerical Monte-Carlo methods. M.: Nauka, 1973. 312 p. (In Russ.)

[4] **Шарый С.П.** Интервальный анализ или методы Монте-Карло? // Вычисл. технологии. 2007. Т. 12, № 1. С. 103–115.
Shary, S.P. Interval analysis or Monte-Carlo methods? // Comput. Technologies. 2007. Vol. 12, No. 1. P. 103–115. (In Russ.)

[5] **Dobronets, B.S., Krantsevich, A.M., Krantsevich, N.M.** Software implementation of numerical operations on random variables // SibFU J. Mathematics & Physics. 2013. No. 6(2). C. 168–173.

[6] **Добронец Б.С., Попова О.А.** Численный вероятностный анализ неопределенных данных. Красноярск: Сиб. федер. ун-т, 2014. 167 с.
Dobronets, B.S., Popova, O.A. The numerical probabilistic analysis of uncertainty data. Krasnoyarsk: Sib. Feder. Univ., 2014. 167 p. (In Russ.)

[7] **Вентцель Е.С.** Теория вероятностей. М.: Наука, 1969. 576 с.
Wenttsel, E.S. Probability. Moscow: Nauka, 1969. 576 p. (In Russ.)

[8] **Гнеденко Б.В.** Курс теории вероятностей. М.: Наука, 1988. 448 с.
Gnedenko, B.V. Course of probability theory. Moscow: Nauka, 1988. 448 p. (In Russ.)

[9] **Billard, L., Diday, E.** Symbolic data analysis: conceptual statistics and data mining. John Wiley & Sons, 2006.

[10] **Попова О.А.** Гистограммы второго порядка для численного моделирования в задачах с информационной неопределенностью // Изв. Южного федер. ун-та. Техн. науки. 2014. № 6(179). С. 6–13.

Popova, O.A. Second order histogram for numerical simulation problems with information uncertainty // Izv. SFedU. Engineering Sci. 2014. No. 6(179). P. 6–13. (In Russ.)

[11] **Dobronets, B.S., Popova, O.A.** Numerical probabilistic analysis under aleatory and epistemic uncertainty // Reliable Computing. 2014. Vol. 19. P. 274–289.

[12] **Добронец Б.С., Попова О.А.** Численные операции над случайными величинами и их приложения // Журн. Сиб. федер. ун-та. Математика и физика. 2011. Т. 4, № 2. С. 229–239.
Dobronets, B.S., Popova, O.A. Numerical operations on random variables and their application // SibFU J. Math. and Physics. 2011. Vol. 4, No. 2. P. 229–239. (In Russ.)

[13] **Добронец Б.С., Попова О.А.** Элементы численного вероятностного анализа // Вестн. Сиб. гос. аэрокосм. ун-та. им. акад. М.Ф. Решетнева. 2012. № 2(42). С. 19–23.
Dobronets, B.S., Popova, O.A. Elements of a numerical probabilistic analysis // Vestn. Sib. Gos. Aerokosm. Un-ta im. Akad. M.F. Reshetneva. 2012. No. 2(42). P. 19–23. (In Russ.)

[14] **Альберг Дж., Нильсон Э., Уолш Дж.** Теория сплайнов и ее приложения. М.: Мир, 1972. 318 с.
Ahlberg, J.H., Nilson, E.N., Walsh, J.L. The theory of splines and their applications. New York: Acad. Press, 1967. 283 p.

[15] **Дэйвид Г.** Порядковые статистики. М.: Наука, 1979. 336 с.
David, H.A. Order statistics. 2rd Edition. New York: John Wiley, 1981. 360 p.

[16] **Добронец Б.С., Попова О.А.** Представление и обработка неопределенности на основе гистограммных функций распределения и p-boxes // Информатизация и связь. 2014. № 2. С. 23–26.
Dobronets, B.S., Popova, O.A. Representation and processing of uncertainty based on histogram distribution functions and p-boxes // Informatization and Communication. 2014. No. 2. P. 23–26. (In Russ.)

Поступила в редакцию 6 февраля 2017 г.

Application of numerical probabilistic analysis in interpolation problems

Popova, Olga A.

Institute of Space and Informatic Technologies SFU, Krasnoyarsk, 660041, Russia

Corresponding author: Popova, Olga A. e-mail: OlgaArc@yandex.ru

The paper addresses the application of numerical probabilistic analysis for construction of function approximations under conditions of stochastic uncertainty. We introduced the definitions of random interpolation and random spline in this study. To define these concepts we use the probabilistic extensions as the base component for numerical probabilistic analysis. To construct these extensions the arithmetical operations for probability density functions of the random variables and numerical procedures for the probabilistic extensions have been applied. Lagrange interpolation polynomials and polynomial splines with input random data are discussed. To employ the numerical probabilistic analysis we propose the special class of the random Lagrange interpolating polynomials.

We argue that our approach to uncertainty representation allows us to build reliable estimates for random functions. In the case of a random piecewise linear interpolation we assess the accuracy of the mathematical expectation and variance.

To use the numerical probabilistic analysis we proposed special class of the random cubic splines. For the construction of the random splines we solved the random system of linear algebraic equations.

Our study shows that the application of the random splines for approximation of distribution functions and probability densities and allows the reliable estimate. This approach is based on the properties of order statistics. Reliable estimates are represented as the histogram P-boxes and second order histograms.

To demonstrate our theory we use numerical examples to build the histogram P-boxes and the second order histogram. It is important to note that numerical results show the effectiveness of our techniques to small samples.

Keywords: numerical probabilistic analysis, functions of random variables, interpolation, verified bounds.

Received 6 February 2017