УДК 519.25

В. П. Долгин, канд. техн. наук
Севастопольский национальный технический университет
(Украина, 99053, Севастополь, ул. Университетская, 33,
тел. (0692) 543570; e-mail: root@sevgtu.sebastopol.ua)

Метод оценки требуемого объема выборки

Изложен метод оценки требуемого объема случайной выборки по критерию погрешности ее математического ожидания. Получены аналитические зависимости для граничных значений погрешности и дисперсии. Приведены результаты имитационного моделирования.

Викладено метод оцінки необхідного об'єму випадкової вибірки по критерію погрішності математичного очікування. Отримано аналітичні залежності для граничних значень погрішності та дисперсії. Наведено результати імітаційного моделювання.

К л ю ч е в ы е с л о в а: оценка, математическое ожидание, дисперсия, погрешность, критерий.

Одной из актуальных задач при анализе стохастических процессов является задача оценки требуемого объема выборки, обеспечивающего надежное определение статистических параметров. Наиболее важным параметром можно считать математическое ожидание, входящее в выражения для многих статистических законов распределения.

По определению [1, 2] математическое ожидание $m_x(t)$ дискретно распределенной случайной величины x можно оценить, имея выборку, состоящую из n элементов:

$$m_x(t) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

где x_i — значение случайной величины в момент времени t_i ($x_i = x(t_i)$). В упрощенной форме выражение математического ожидания записывают в виде $m_x(t) = M[X(t)]$ как начальный момент первого порядка [1].

Приведенное выражение позволяет вычислить математическое ожидание $m_x(t)$ с некоторой погрешностью. Объем выборки n (число реализаций) случайной величины x_i определяет погрешность, с которой может быть получена оценка математического ожидания $m_x(t)$. При увеличении значения n погрешность оценки уменьшается.

Получить точное значение требуемого объема выборки не представляется возможным ввиду случайного характера изменений величины x, но можно найти значение $m_x(t)$ с погрешностью, не превышающей заданного значения, с некоторой доверительной вероятностью P_x. Для решения этой задачи необходимо применение итерационной процедуры, в процессе которой следует задать объем выборки, вычислить математическое ожидание и дисперсию статистической совокупности, а затем уточнять параметры с новым массивом данных до тех пор, пока не будет обеспечена требуемая точность.

Предлагается упрощение процедуры оценки требуемого объема выборки, исключающее итерационный метод решения задачи. Найдем оценку требуемого объема выборки при заданном значении погрешности. Текущая оценка математического ожидания может быть вычислена рекуррентно на основе его предыдущего значения $m_k(t) = \frac{1}{k}\sum_{i=1}^{k} x_i$:

$$m_k = \frac{(k-1)\, m_{k-1} + x_k}{k}, \tag{1}$$

где k — объем выборки; x_k — k-й элемент выборки; m_{k-1} — оценка математического ожидания на предыдущем шаге (для объема выборки, равного $k-1$).

Определим абсолютное значение относительной погрешности δ математического ожидания на очередном шаге k:

$$\delta = \left| \frac{m_k - m_{k-1}}{m_0} \right|,$$

где m_0 — истинное значение математического ожидания. Подставив в правую часть значение m_k из формулы (1), после преобразований получим

$$\delta = \left| \frac{m_{k-1} - x_k}{m_0 k} \right|.$$

Исключив слагаемое x_k, найдем максимально возможное значение погрешности:

$$\delta \le \left| \frac{m_{k-1}}{m_0 k} \right|.$$

Поскольку предельные значения $\lim\limits_{k\to\infty} m_k$ и $\lim\limits_{k\to\infty} m_{k-1}$ стремятся к m_0, для больших значений k можно принять

$$\delta \le 1/k, \tag{2}$$

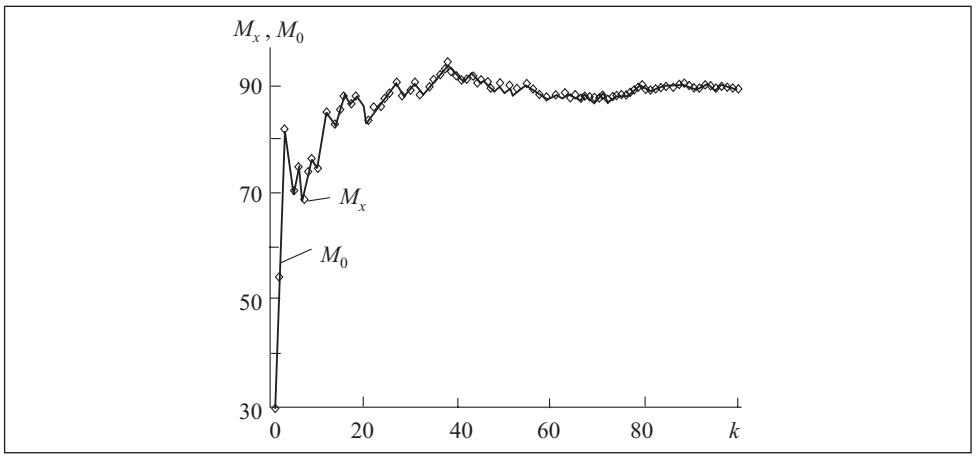

Рис. 1. Графики зависимости математического ожидания M_x от изменения объема выборки $k = [1, n]$

где k — требуемый объем выборки. Таким образом, по заданной погрешности δ (2), положив $k = n$, можно оценить требуемый объем выборки:

$$n \le 1/\delta, \tag{3}$$

так как при возрастании числа реализаций значение математического ожидания существенно не изменяется, оставаясь в пределах погрешности δ.

На рис. 1 показан процесс изменения математического ожидания M_x в зависимости от изменения объема выборки при равномерном законе распределения случайной величины в диапазоне $[a, b]$, где $a = 30$, $b = 150$. Теоретическая (истинная) величина математического ожидания [1, 2] составляет $M_0 = \dfrac{a+b}{2} = 90$.

На рис. 2 представлен график границы погрешности GM_0, построенный по формуле (2) при изменении переменной k в диапазоне $[1, n]$. Для заданного уровня допустимой погрешности $\delta_{\max} = 0{,}01$ величина объема выборки в соответствии с формулой (3) принята $n = 100$. График GM_x соответствует относительным приращениям математического ожидания $(M_k - M_{k-1})/M_0$, где M_0 — теоретическое значение математического ожидания при увеличении текущего объема выборки $k = [1, n]$.

Как видно из рис. 2, кривая изменения относительного приращения математического ожидания GM_x не выходит за пределы границы (кривая GM_0), рассчитанной по формуле (2), что обеспечивает высокую достоверность результата определения требуемого объема выборки при заданной погрешности δ оценки математического ожидания статистической совокупности.

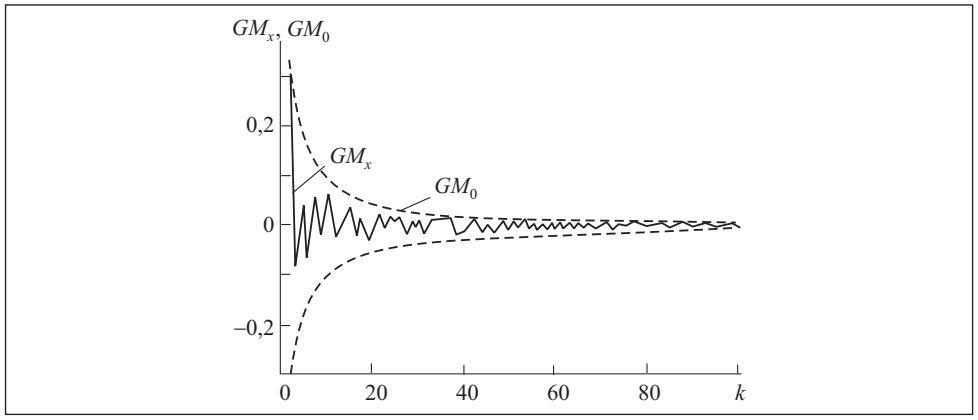

Рис. 2. График относительного приращения математического ожидания $(M_k - M_{k-1})/M_0$ при изменении объема выборки k

Для проверки корректности полученной оценки объема выборки (3) определим ее необходимый объем с учетом параметров статистической совокупности [1, 2]:

$$N = \frac{t_x^2 D_x N_x}{\Delta_x^2 (N_x - 1) + t_x^2 D_x}. \tag{4}$$

Здесь N_x и D_x — объем и дисперсия; $\Delta_x = \delta M_x$, где M_x — математическое ожидание статистической совокупности; t_x — корень уравнения Лапласа

$$P_x = \sqrt{\frac{2}{\pi}} \int_0^{t_x} \exp(-x^2/2)\, dt.$$

При этом $P_x = P\{|M_x - M_0| < \Delta_x\}$ — вероятность выполнения условия $|M_x - M_0| < \Delta_x$. Значение t_x, входящее в выражение (4), находим в результате решения уравнения Лапласа в форме

$$P_x = \mathrm{erf}\,(t_x/\sqrt{2}). \tag{5}$$

Представим уравнение (4) в виде

$$N = \frac{N_x}{1 + K_x},$$

где K_x — коэффициент,

$$K_x = \frac{\delta^2 M_x^2 (N_x - 1)}{D_x t_x^2},$$

и упростим его, положив $\delta = 1/N_x$ ($N_x >> 1$) в соответствии с (3). Тогда выражение для K_x примет вид

$$K_x = \frac{M_x^2}{D_x t_x^2 N_x}.$$

Приняв максимально допустимое значение $P_x = 1$ и решив уравнение Лапласа (5), получим $t_x^{-2} = 0{,}0041$. В большинстве практических случаев $N_x >> 1$ и $\delta << 1$, что позволяет уравнение (4) свести к неравенству

$$N \le N_x. \qquad (6)$$

Для распределения Вейбулла $M_x = 0$ и $K_x = 0$. Поэтому объем выборки, гарантирующий погрешность не более δ, соответствующую (3), составляет $N = n$, что получаем в результате замены $N_x = n$ в формуле (6).

При экспоненциальном законе распределения $M_x^2 / D_x = 1$ и $K_x > 0{,}0041/N_x$, что позволяет принять объем выборки согласно (3) $N = n$, обеспечивающий значение M_0 с погрешностью не более δ.

При равномерном законе распределения $K_x > 0{,}011/N_x$ в стандартном интервале изменения случайной величины [0,1], что позволяет принять объем выборки $N = n$, обеспечивающий требуемую точность определения ее математического ожидания.

Полученная оценка погрешности является состоятельной, несмещенной и эффективной [1, 2]. Таким образом, применение формулы (3) при анализе случайной последовательности гарантирует получение достоверной оценки математического ожидания стохастической модели воздействия с заданной точностью при объеме выборки $n = \delta^{-1}$.

Следует заметить, что можно решать с высокой степенью достоверности и обратную задачу определения погрешности оценки математического ожидания при известном объеме выборки, определяемой доверительной вероятностью $P(m_0 - \delta < m_x < m_0 + \delta) = 1$, где m_0 и m_x — истинное и искомое значения математического ожидания.

В некоторых случаях представляет интерес динамика процесса изменения дисперсии с увеличением объема выборки. Известно [1, 2], что погрешность вычисления дисперсии уменьшается при возрастании объема выборки. На рис. 3 изображен график изменения дисперсии D_x при изменении объема выборки $k = [1, n]$ и показано теоретическое значение дисперсии $D_0 = 1200$.

Найдем изменение дисперсии за один шаг, обозначив

$$\Delta D_k = D_{k+1} - D_k, \qquad (7)$$

где D_k — дисперсия выборки, состоящей из k элементов x_i,

$$D_k = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - m_k^2, \quad i = 1, \ldots, k;$$

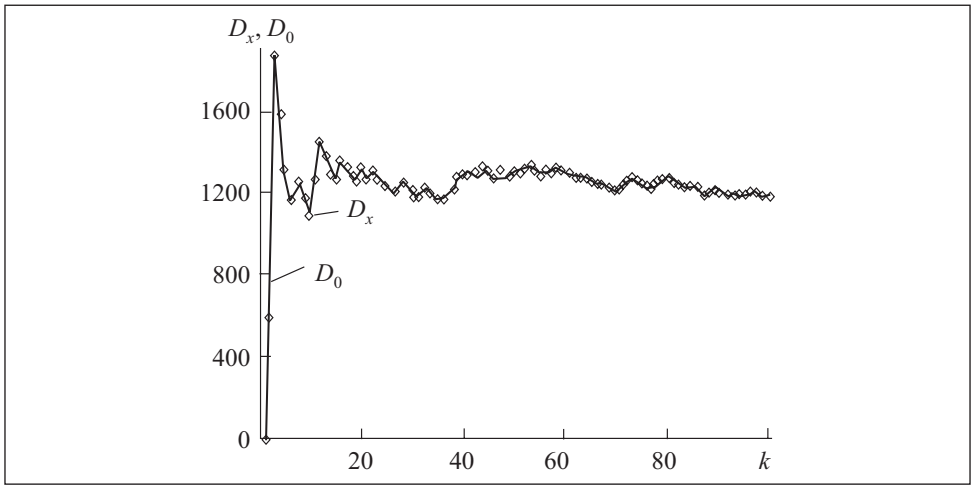

Рис. 3. График изменения дисперсии D_x при изменении объема выборки $k=[1,n]$

D_{k+1} — дисперсия выборки, содержащей $k+1$ элементов x_i,

$$D_{k+1}=\frac{1}{k+1}\sum_{i=1}^{k+1}x_i^2-m_{k+1}^2,\ \ i=1,...,k+1;$$

m_k и m_{k+1} — математическое ожидание выборки из k и $k+1$ элементов,

$$m_k=\frac{1}{k}\sum_{i=1}^{k}x_i,\ \ m_{k+1}=\frac{1}{k+1}\sum_{i=1}^{k+1}x_i.$$

Выполнив подстановку, получим

$$\Delta D_k=\frac{1}{k+1}\sum_{i=1}^{k+1}x_i^2-m_{k+1}^2-\frac{1}{k}\sum_{i=1}^{k}x_i^2+m_k^2.$$

Представив $D_{k+1}=\frac{1}{k+1}\sum_{i=1}^{k}x_i^2+\frac{1}{k+1}x_{k+1}^2-m_{k+1}^2$, после преобразований найдем

$$\Delta D_k=\frac{1}{k+1}\left(x_{k+1}^2-\frac{1}{k}\sum_{i=1}^{k}x_i^2\right)+m_k^2-m_{k+1}^2. \qquad (8)$$

Преобразуем выражение (8), разложив разность квадратов математических ожиданий:

$$m_k^2-m_{k+1}^2=(m_k-m_{k+1})(m_k+m_{k+1}). \qquad (9)$$

Преобразуем разность математических ожиданий:

$$m_k - m_{k+1} = \frac{1}{k}\sum_{i=1}^{k} x_i - \frac{1}{k+1}\sum_{i=1}^{k} x_i - \frac{1}{k+1} x_{k+1}.$$

Сгруппировав слагаемые, получим

$$m_k - m_{k+1} = \frac{1}{k(k+1)}\sum_{i=1}^{k} x_i - \frac{1}{k+1} x_k.$$

Учитывая, что $m_k = \frac{1}{k}\sum_{i=1}^{k} x_i$, находим $m_k - m_{k+1} = \frac{1}{k+1}(m_k - x_k)$. Подставив это выражение в (9), а результат — в (8), получим

$$\Delta D_k = \frac{1}{k+1}\left[(m_k + m_{k+1})(m_k - x_k) - \frac{1}{k}\sum_{i=1}^{k} x_i^2 + x_{k+1}^2\right]. \qquad (10)$$

Упростив выражение (10), исключив конечный элемент выборки и приняв $m_{k+1} = m_k$, найдем

$$\Delta D_k \le \frac{1}{k}\left(2m_k^2 - \frac{1}{k}\sum_{i=1}^{k} x_i^2\right). \qquad (11)$$

Учитывая, что $D_k = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - m_k^2$, приводим выражение (11) к окончательной форме для вычисления абсолютного значения относительной погрешности $\varepsilon_k = \left|\frac{\Delta D_k}{D_k}\right|$ на k-м шаге:

$$\varepsilon_k \le \frac{1}{k}\left|\frac{D_k - m_k^2}{D_k}\right|.$$

При оценке дисперсии выборки с числом элементов n абсолютное значение границы относительной погрешности определения дисперсии D_n не превосходит величины

$$\varepsilon_n \le \frac{1}{n}\left|1 - m_n^2 / D_n\right|, \qquad (12)$$

где m_n и D_n — математическое ожидание и дисперсия выборки. Следует заметить, что при удовлетворении условия $1 >> m_n^2 / D_n$ выражение (12) упрощается к виду

$$\varepsilon_n \approx 1/n. \qquad (13)$$

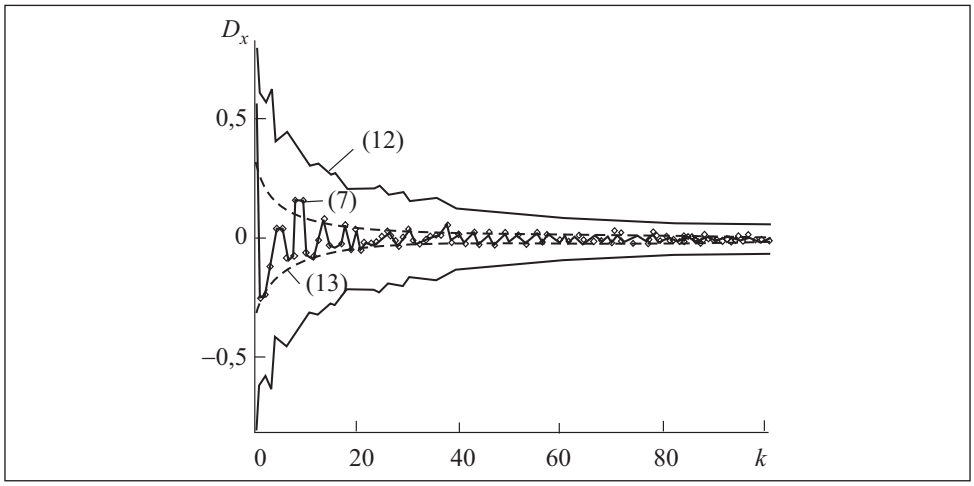

Рис. 4. Графики зависимости относительного приращения дисперсии $D_x = (D_k - D_{k-1}) / D_0$ от изменения объема выборки $k = [1, n]$: ◊ — относительное изменение приращений дисперсии (7); ——— — граница погрешности (12); - - - - - — упрощенная зависимость (13)

На рис. 4 представлены кривые, соответствующие зависимостям (7), (12), (13).

В результате имитационного моделирования при $\delta = 0{,}01$, $n = 100$ для выборки случайных величин с равномерным законом распределения и параметрами $a = 30$, $b = 150$ при $M_0 = 90$ и $D_0 = 1200$ получены параметры распределения $M_x = 89{,}25$ и $D_x = 1187$. Вычисленное по формуле (4) значение N совпало с полученным по формуле (3) и составило $N = n = 100$, что свидетельствует об эффективности рассмотренного метода, позволяющего непосредственно оценить требуемый объем выборки без вычисления ее статистических параметров.

Выводы

Применение изложенного метода не требует априорной информации о характере случайного распределения и его параметрах, что существенно упрощает процедуру имитационного моделирования стохастических процессов, обеспечивая надежную оценку требуемого объема выборки на основании допустимой погрешности модели.

Полученные результаты могут найти применение при моделировании и анализе технологических [3], автотранспортных, радиотехнических и других стохастических систем автоматического управления.

The method is stated for estimation of the required accidental sample size by the criterion of the expected value error. Analytical dependences for the boundary values of the error and variance are obtained. The results of imitation modeling are presented.

1. *Вентцель Е.С*. Теория вероятностей. — М.: Высш. шк., 1999. — 576 с.
2. *Креденцер Б.П., Ластовченко М.М., Сенецкий С.А. и др.* Решение задач надежности и эксплуатации на универсальных ЭЦВМ. Под ред. Н.А. Шишонка. — М.: Сов. радио, 1967. — 400 с.
3. *Братан С.М.* Построение подсистемы динамической диагностики для оценки непосредственно наблюдаемых параметров при чистом и тонком шлифовании // Ресурсозберігаючі технології виробництва та обробки тиском матеріалів у машинобудуванні: Зб. наук. праць в 2-х частинах. Ч. 2. — Луганськ: вид-во СНУ ім. Даля, 2004. — С. 182—191.

Поступила 22.11.11

ДОЛГИН Владимир Прохорович, канд. техн. наук, доцент кафедры автомобильного транспорта Севастопольского Национального технического университета. В 1958 г. окончил Военно-морское инженерное училище им. Ф. Э. Дзержинского (г. Ленинград), в 1965 г. — Севастопольский приборостроительный ин-т. Область научных исследований — адаптивные модели в системах управления технологическими объектами.