

УДК 519.25

В.П. Долгин, доцент, канд. техн. наук*Севастопольский национальный технический университет**ул. Университетская, 33, г. Севастополь, Украина, 99053**E-mail: root@sevgtu.sebastopol.ua***ОЦЕНКА ТРЕБУЕМОГО ОБЪЕМА ВЫБОРКИ**

Изложен метод оценки требуемого объема случайной выборки по критерию погрешности ее математического ожидания.

Ключевые слова: оценка, математическое ожидание, дисперсия, погрешность, критерий.

При анализе стохастических процессов одной из актуальных является задача оценки требуемого объема выборки, обеспечивающего надежное определение статистических параметров. Наиболее важным из параметров можно считать математическое ожидание, которое входит в выражение 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 $m_x(t)$. С увеличением n погрешность оценки математического ожидания уменьшается.

Получить точное значение требуемого объема выборки не представляется возможным ввиду случайного характера изменений самой 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 иметь реализацию, анализ которой позволяет получить желаемый результат [1, 2], что предполагает применение итерационной процедуры, в процессе которой необходимо задать объем выборки, вычислить математическое ожидание и дисперсию статистической совокупности, затем уточнить параметры с новым массивом данных до тех пор, пока не будет обеспечена требуемая точность.

Целью статьи является 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 как

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

где $m_k = \frac{1}{k} \sum_{i=1}^k x_i$ – оценка математического ожидания; k – объем выборки, x_k – k -й элемент выборки;

m_{k-1} – оценка математического ожидания на предыдущем шаге (для объема выборки, равном $k-1$).

Абсолютное значение относительной погрешности δ математического ожидания на очередном шаге k определим как

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

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

$$\delta \leq \left| \frac{m_{k-1}}{m_O k} \right|.$$

Так как предельные значения $m_k \xrightarrow{k \rightarrow \infty} m_O$ и $m_{k-1} \xrightarrow{k \rightarrow \infty} m_O$, то для больших значений k можно принять

$$\delta \leq \frac{1}{k}, \quad (2)$$

где k представляет собой требуемый объем выборки.

Таким образом, по заданной погрешности δ в соответствии с (2) можно оценить требуемый объем выборки, положив $k = n$

$$n = 1/\delta, \quad (3)$$

так как при росте числа реализаций значение математического ожидания существенно изменяться не будет, оставаясь в пределах погрешности δ .

На рисунке 1 изображен для примера процесс изменения математического ожидания M_x при изменении объема выборки для равномерного закона распределения случайной величины в диапазоне $[a, b]$; $a = 30$; $b = 150$. Теоретическая (истинная) величина математического ожидания $[1, 2]$ составляет величину $M_O = \frac{a+b}{2} = 90$.

На рисунке 2 изображен график границы погрешности GMo, построенный по формуле (2) при переменной k , изменяющейся в диапазоне $[1, n]$. Для заданного уровня допустимой погрешности $\delta_{\max} = 0,01$ величина объема выборки принята в соответствии с формулой (3) равной $n = 100$. График GMx изображает относительные приращения математического ожидания $(M_k - M_{k-1})/M_O$, где M_O – теоретическое значение математического ожидания, при увеличении текущего объема выборки k ($k = [1, n]$).

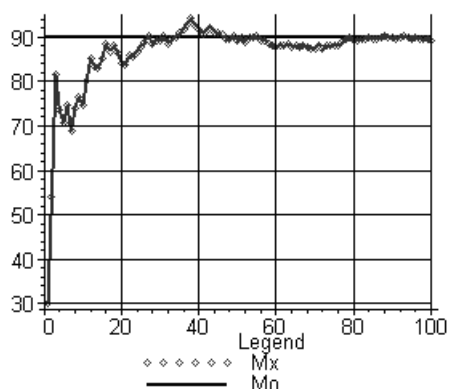


Рисунок 1 – Изменение математического ожидания M_x при изменении объема выборки k ($k = [1, n]$)

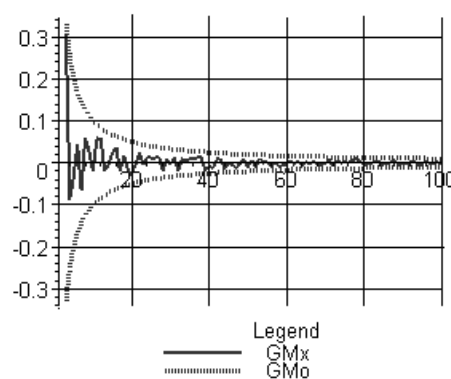


Рисунок 2 – Изменение относительного приращения математического ожидания $(M_k - M_{k-1})/M_O$ при изменении объема выборки k

Как видно из графиков, кривая изменения относительного приращения математического ожидания GMx не выходит за пределы границы (кривая GMo), рассчитанной по формуле (2), что обеспечивает высокую достоверность результата определения требуемого объема выборки (4) при заданной погрешности δ оценки математического ожидания статистической совокупности.

Для проверки корректности полученной оценки объема выборки (3) определим необходимый ее объем, учитывающий параметры статистической совокупности $[1, 2]$

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

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

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

$P_x = P\{|M_x - M_O| < \Delta_x\}$ – вероятность выполнения условия $|M_x - M_O| < \Delta_x$, M_O – истинное значение математического ожидания статистической совокупности.

Значение t_x , входящее в выражение (4), получается как результат решения уравнения Лапласа в форме

$$P_x = \text{erf}(t_x / \sqrt{2}). \quad (5)$$

Представим уравнение (4) в виде

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

где $K_x = \frac{\delta^2 M_x^2 (N_x - 1)}{D_x t_x^2}$ – коэффициент, и упростим его, положив $\delta = 1/N_x$ ($N_x \gg 1$) в соответствии с формулой (3). Тогда выражение для K_x сведется к виду

$$K_x = \frac{M_x^2}{D_x t_x^2 N_x}. \quad (6)$$

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

$$N \leq N_x.$$

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

Для экспоненциального распределения отношение $M_x^2/D_x = 1$ и величина коэффициента $K_x > 0,0041/N_x$, что позволяет принять объем выборки по формуле (3), обеспечивающий значение математического ожидания с погрешностью не более δ , равный $N = n$.

Можно показать, что для равномерного закона распределения коэффициент $K_x > 0,011/N_x$ 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], что является следствием изложенного хода доказательства.

Таким образом, применение формулы (3) при анализе случайной последовательности гарантирует получение достоверной оценки математического ожидания стохастической модели воздействия с заданной точностью при выборе объема выборки n , равном величине $n = \delta^{-1}$.

Важно отметить, что можно решать и обратную задачу определения погрешности δ оценки математического ожидания при известном объеме выборки n с высокой степенью достоверности, определяемой доверительной вероятностью

$$P(m_0 - \delta < m_x < m_0 + \delta) = 1,$$

где m_0 – истинное; m_x – искомое значение математического ожидания.

В некоторых случаях представляет интерес динамика процесса изменения дисперсии с ростом объема выборки. Известно [1,2], что погрешность вычисления дисперсии уменьшается с ростом объема выборки (рисунок 3). На рисунке 3 изображен график изменения дисперсии D_x при изменении объема выборки k ($k = [1, n]$) и нанесено теоретическое значение дисперсии $D_O = 1200$ для данных предыдущего примера.

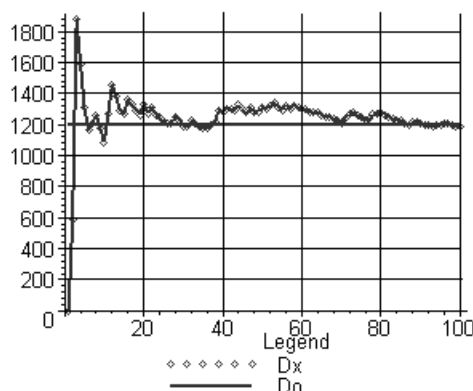


Рисунок 3 – Изменение дисперсии D_x при изменении объема выборки k ($k = [1, n]$)

Найдем изменение дисперсии за один шаг, обозначив

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

где $D_k = \frac{1}{k} \sum_{i=1}^k x_i^2 - m_k^2$ – дисперсия выборки, состоящей из k элементов x_i ($i = 1 \dots k$);

$D_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} x_i^2 - m_{k+1}^2$ – дисперсия выборки, содержащей $k+1$ элементов x_i ($i = 1 \dots k+1$);

$m_k = \frac{1}{k} \sum_{i=1}^k x_i$ – математическое ожидание выборки из k элементов;

$m_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} x_i$ – математическое ожидание выборки из $k+1$ элементов.

Выполнив подстановку в (7), 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

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

Преобразуем выражение (8) разложив разность квадратов математических ожиданий

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

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

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

Учтя, что $m_k = \frac{1}{k} \sum_{i=1}^k x_i$, найдем

$$m_k - m_{k+1} = \frac{1}{k+1} (m_k - x_{k+1}).$$

Подставив выражение в (9), а результат – в (8), получим

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

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

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

Имея ввиду, что $D_k = \frac{1}{k} \sum_{i=1}^k x_i^2 - m_k^2$, приведем последнее выражение к окончательному виду для

вычисления абсолютного значения относительной погрешности $\epsilon_k = \left| \frac{\Delta D_k}{D_k} \right|$ на k -м шаге

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

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

$$\epsilon_n \leq \frac{1}{n} \left| 1 - m_n^2 / D_n \right|, \quad (10)$$

где m_n и D_n – математическое ожидание и дисперсия выборки соответственно.

Следует заметить, что при удовлетворении условия $1 \gg m_n^2 / D_n$, выражение погрешности (10) упрощается до вида

$$\epsilon_n \approx 1/n. \quad (11)$$

На рисунке 4 изображены кривые, иллюстрирующие зависимости (7), (10), (11). Жирной сплошной линией нанесена кривая, соответствующая границе погрешности (10). Кривая, соответствующая упрощенной зависимости (11), нанесена штриховой линией. Относительное изменение приращений дисперсии (7) обозначено D_x .

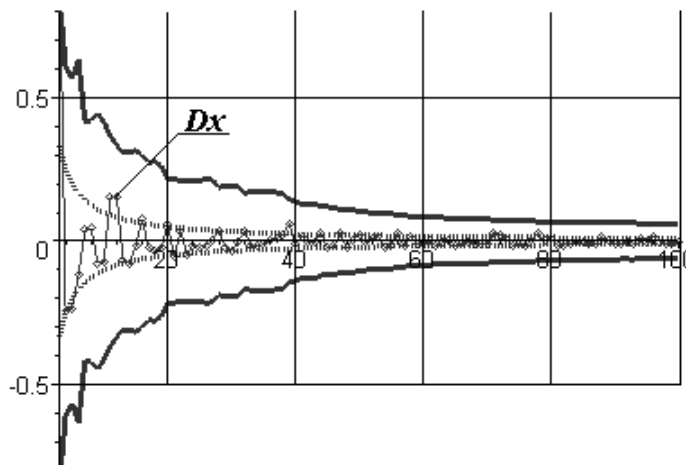


Рисунок 4 – Изменение относительного приращения дисперсии $D_x = (D_k - D_{k-1}) / D_0$ при изменении объема выборки k ($k = [1, n]$)

В результате имитационного моделирования при значениях $\delta = 0,01$; $n = 100$ для выборки случайных величин с равномерным законом распределения и параметрами $a = 30$; $b = 150$ при теоретических значениях математического ожидания $M_O = 90$ и дисперсии $D_O = 1200$ вычисленные параметры распределения составили $M_x = 89,25$ 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.

Применение изложенного метода не требует априорной информации о характере случайного распределения и его параметрах, что существенно упрощает благодаря его универсальности 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.

Библиографический список использованной литературы

1. Вентцель Е.С. Теория вероятностей / Е.С. Вентцель. — М.: Высш. шк., 1999. — 576 с.
2. Решение задач надежности и эксплуатации на универсальных ЭЦВМ / Б.П. Креденцер, М.М. Ластовченко, С.А. Сенецкий и др., 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. пр. — Луганськ, 2004. — С. 182–191.

Поступила в редакцию 09.11.2011 г.

Долгін В.П. Оцінка необхідного об'єму вибірки

Викладено метод оцінювання необхідного об'єму випадкової вибірки за критерієм похибки математичного очікування.

Ключові слова: оцінка, математичне очікування, дисперсія, похибка, критерій.

Dolgin V.P. Estimation of the required sample size

The method of estimation of the required accidental sample size on the criterion of the expected value error is explained.

Keywords: estimation, expected value, variance, error, criterion.