УДК 519.217.2+338+331

Ю. В. Щербина, к.т.н., С. Д. Асабашвілі, А. М. Прокопенко*Одеська державна академія технічного регулювання та якості, м. Одеса***МОДЕЛЮВАННЯ НЕСИМЕТРИЧНО РОЗПОДІЛЕНИХ ПРОЦЕСІВ**

У статті розглянуто проблеми моделювання несиметрично розподілених стохастичних процесів та виконання оцінки якості процесу моделювання на основі використання критерію χ^2 -квадрат Пірсона. Визначені проблеми, що виникають у кожному індивідуальному випадку моделювання. Обґрунтовано вибір непараметричного методу дослідження на основі гістограм та надано пропозиції, щодо вибору їх основних параметрів. Показано, що на успіх моделювання суттєво впливає якість статистичного матеріалу і окремою проблемою є його фільтрація попередня до визначення показника якості. Надані рекомендації щодо послідовності процедур формування моделі стохастичного процесу та обробки результатів обчислення показників гістограм.

Ключові слова: непараметричний метод гістограм, інтервали гістограм, Критерій Пірсона χ^2 -квадрат, генератор Мерсена, метод зворотних функцій, метод ступінчастої апроксимації, метод усічення, тест Колмогорова-Смірнова, тест Андерсона-Дарлінга, квартиль.

Ю. В. Щербина, к.т.н., С. Д. Асабашвілі, А. М. Прокопенко**МОДЕЛИРОВАНИЕ НЕСИММЕТРИЧНЫХ РАСПРЕДЕЛЕННЫХ ПРОЦЕССОВ**

В статье рассмотрены проблемы моделирования несимметрично распределенных стохастических процессов и выполнения оценки качества процесса моделирования на основе использования критерия χ^2 -квадрат Пирсона. Определены проблемы, возникающие в каждом индивидуальном случае моделирования. Обоснован выбор непараметрического метода исследования на основе гистограмм и представлены предложения по выбору их основных параметров. Показано, что на успех моделирования существенно влияет качество статистического материала и отдельной проблемой является его фильтрация предшествующая определению показателя качества. Определена последовательность процедур формирования модели стохастического процесса и обработки результатов вычисления параметров гистограммы.

Ключевые слова: непараметрический метод гистограмм, интервалы гистограмм, Критерий Пирсона χ^2 -квадрат, генератор Мерсенна, метод обратных функций, метод ступенчатой аппроксимации, метод усечения, тест Колмогорова-Смирнова, тест Андерсона-Дарлинга, квартиль.

Yu. V. Shcherbina, PhD, S. D. Asabashvili, A. M. Prokopenko**SIMULATION OF ASYMMETRICALLY DISTRIBUTED PROCESSES**

The article deals with the problems of modeling asymmetrically distributed stochastic processes and assessing the quality of the modeling process based on the use of the Pearson chi-square criterion. The problems that arise in each individual case of modeling are identified. As a method of generating an output stream of random numbers distributed according to a given law of probability distribution, the method of inverse functions is proposed, and as a generator of the input stream, it is proposed to use a stream generated by the Mersenne generator. The analysis of research methods and evaluation of computer models of stochastic processes is given and the choice of a nonparametric research method based on histograms is substantiated. Suggestions on the choice of their main parameters are provided: the size of the interval for separating the statistical set of values of the output stream of the model and the number of such intervals, calculated on the basis of the size of the statistical sample. It is shown that the quality of statistical material significantly affects the success of modeling. In the case of an asymmetric nature of the distribution of the model's output stream, the accuracy of calculating the chi-square exponent is significantly affected by those values that fall into the extreme intervals of the histogram. It is shown that the reason for this is the incommensurability of their values with the values in the central part of the histogram and this is a separate modeling

problem, as a solution to which there can be filtering of the statistical set, preceding the determination of the quality indicator. The sequence of procedures for forming a model of a stochastic process and processing the results of calculating the parameters of the histogram is determined.

Keywords: Nonparametric histogram method, histogram intervals, Pearson's chi-square test, Mersenne generator, inverse function method, step approximation method, truncation method, Kolmogorov-Smirnov test, Anderson-Darling test, quartile.

DOI 10.32684/2412-5288-2020-1-16-42-47

Вступ

Проектування складних систем на попередніх етапах передбачає їх декомпозицію на складові частини та виявлення математичних і логічних взаємозв'язків, які їх об'єднують. Наявність таких моделей дає можливість отримувати загальну уяву про поведінку таких систем у середовищі експлуатації і виконувати попередній вибір параметрів і характеристик системи.

Теорія ймовірностей надає дослідникам набір математичних моделей для імітації гіпотетичних реальних явищ або систем стохастичної природи і, якщо взаємозв'язки, які утворюють модель досить прості, то для отримання точної інформації про поведінку системи можна використовувати математичні методи.

Проблема полягає у тому, що більшість існуючих систем є дуже складними, і для них неможливо створити реальну аналітичну модель. Вихід полягає у тому, щоб використовуючи методи математичної статистики, виконати обґрунтований вибір із сукупності можливих моделей тієї, яка найкращим чином характеризує реальну поведінку конкретної реальної системи. У такому випадку створюють імітаційну модель, яка представляє собою статистичний комп'ютерний аналог процесу, поведінка якого максимально наближається до поведінки об'єкту, що підлягає дослідженню. Постановка експериментів з такою моделлю дає можливість зрозуміти поведінку системи, або оцінити різноманітні стратегії, що забезпечують її функціонування. Імітаційне моделювання має наступні цілі:

1. Описання поведінки систем;
2. Побудова теорії та гіпотез, які дають можливість зрозуміти поведінку системи або процесу;
3. Прогнозування поведінки системи у різних режимах її роботи;

Аналіз останніх досліджень і публікацій

Проблема побудови послідовностей випадкових чисел існує давно. На створення генераторів, що дають можливість отримувати псевдовипадкові числа із заданим законом розподілу витрачено кілька десятиріч. Зусиллями математиків і програмістів було винайдено багато способів аналізу сформованих числових послідовностей

на предмет їх «випадковості» і, нажаль, результати проведених досліджень показують, що ця проблема остаточно ще не вирішена [1].

Основним способом моделювання випадкових величин є метод зворотних функцій, суть якого полягає у тому, що на підставі умов існування зворотних функцій та теореми, згідно з якою безперервна випадкова величина x , яка має функцію розподілу ймовірностей $F(x)$ визначає безперервну рівномірно розподілену на інтервалі $(0, 1)$, випадкову величину $\xi = F^{-1}(\gamma)$ [2]. Метод зворотних функцій добре працює у тих випадках, коли процес можна описати аналітично і зворотна функція для нього існує. У протилежному випадку, як, наприклад, при моделюванні нормально розподіленого процесу, задача суттєво ускладнюється. Існують й інші методи, такі як, наприклад, метод ступінчастої апроксимації (Stepwise approach method) або метод усічення (Rejection Method) [3], але всі вони менш ефективні з точки зору необхідних обчислювальних ресурсів і точності моделювання.

Із сказаного витікає, що джерелом вхідного потоку для процесу моделювання є генератор рівномірно розподілених псевдовипадкових числових послідовностей (ПВП), і якість такого моделювання визначається тим, наскільки рівномірно розподіленою є послідовність чисел, що формується закладеним в основу генератора алгоритмом. Бінарний потік на виході генератора повинен мати максимально можливий період повторення і сторонній спостерігач не повинен мати можливості вгадати наступне число з ймовірністю, що відрізняється від 0,5. Зважаючи на серйозність проблеми, американський інститут стандартизації NIST видав окремим виданням описання пакет тестів для перевірки якості генераторів ПВП [4]. Сьогодні для вимог програмування найбільш придатним вважається генератор Мерсена (Mersenne Twister) [5], розроблений на мові С. Він використовує 128 бітні операції і має період повторення, що складає $2^{19937} - 1$ біт. Цей алгоритм реалізований у вигляді окремої функції, і входить до складу бібліотек таких програмних середовищ як Boost, Glib, C++, Python, Ruby, R, PHP, MATLAB та Autoit. Єдиним для нього обмеженням є використання у криптографічних

перетвореннях.

Робота над створенням моделі фізичного процесу має закінчуватись випробуваннями, на підставі яких можна було б зробити висновок про відповідність розподілу вихідного потоку цілям моделювання. Для цього використовують тести на відповідність, ідея яких полягає у тому, щоб отримати числовий показник належності сформованої вибірки заявленому розподілу ймовірностей. Найбільш ефективними визнані тести Колмогорова-Смірнова, Андерсона-Дарлінга і тест Пірсона (χ^2 -тест). Тести Колмогорова-Смірнова і Андерсона-Дарлінга обмежені безперервними розподілами, в той час як тест χ^2 може застосовуватися як до дискретних, так і до безперервних розподілів [6]. У разі комп'ютерного моделювання найбільш поширеним для такого роду задач є тест Пірсона через його неможливість до обчислювальних ресурсів [7].

Сучасний рівень розвитку комп'ютерної техніки та втілення у практику наукових досліджень різноманітних цифрових приладів дозволяє значно підвищити точність та ефективність моделювання систем і фізичних процесів. Існує велика кількість прикладних пакетів програм, призначених для виконання задач статистики, до складу яких можна віднести Statistica, Mathematica, Math, SPCC, Stata, SYSTAT, SPSS, Maple та інших. Однак, зазвичай, вони досить важкі в освоєнні, коштують занадто дорого, їх використання вимагає значних обчислювальних ресурсів і, головне, кожна конкретна задача моделювання може бути вирішена простіше і більш наглядно за допомогою відомих середовищ програмування.

Мета і завдання досліджень

Враховуючи, що важливим завданням моделювання є відповідність вихідного процесу заданим умовам, метою роботи є пошук можливих шляхів аналізу побудованих моделей на предмет їх відповідності заданому закону розподілу ймовірностей $F(\xi < x)$, які дозволяли б з визначеною долею достовірності дати відповідь на питання про те, чи є створена модель прийнятною для досліджень поведінки системи у тих або інших умовах експлуатації системи.

Викладення основного матеріалу

Головна проблема моделювання полягає в тому, що у кожному конкретному випадку доводиться шукати індивідуальний підхід до вирішення задачі, виходячи із складності процесу, який доводиться моделювати. Спочатку слід визначити, як математично його описати, і чи можливо це зробити взагалі. Якщо ні, то слід виконати його декомпозицію на складові частини, для кожної з яких будувати окрему модель, а

потім перейти до розроблення відповідного до визначеної моделі генератора псевдовипадкових чисел x_i , розподілених за необхідним законом $F(\xi < x)$. Це у більшості випадків приводить до того, що загальна модель має складний несиметричний розподіл, що у свою чергу, ускладнює оцінку отриманих результатів. Для виконання оцінки якості сформованого процесу обирають найбільш придатний у даній ситуації критерій.

Будь-який фізичний процес уявляє собою послідовність переходів системи із одного стану у будь-який інший стан. Через це його можна описати як послідовність величин вихідної змінної x_i , $i = 1, 2, \dots, k$, розподіл якої може бути відомим або невідомим. Взагалі, під час проектування розробники намагаються створити систему з бажаним розподілом. Якщо ж припущення про його тип зробити не вдається, то моделюються складові частини системи і накопичується статистичний матеріал, на основі досліджень якого робиться висновок про те, до якого з відомих розподілів вихідний процес системи стоїть найближче. Якщо ж розподіл відомий і відомі його параметри, то переходять до створення програмної моделі процесу [8].

Для простоти розуміння розглянемо методологію створення моделі несиметрично розподіленого процесу, що описується експоненціальним законом вихідної величини x . Тобто, нехай

$$F(\xi < x) = \begin{cases} 0 & \text{при } x \leq 0 \\ 1 - e^{-\lambda x} & \text{при } x > 0 \end{cases},$$

а функція щільності розподілу виглядає як

$$p(x) = \begin{cases} 0 & \text{при } x \leq 0 \\ e^{-\lambda x} & \text{при } x > 0 \end{cases}.$$

Зворотною функцією по відношенню до функції $F(x)$, є функція

$$y = F^{-1}(x) = -\frac{\ln(1-x)}{\lambda}. \quad (1)$$

Аргументом цієї функції має бути потік рівномірно розподілених на інтервалі $[0, 1]$ випадкових чисел x_i , $i = 1, 2, \dots, n$. Чим більше він буде відповідати вимогам рівномірності, тим більше величина на виході моделі, буде відповідати експоненційному розподілу ймовірностей. Для формування такого потоку на мові C++ можна скористатися функцією

```
random_device rd; // Ініціалізація генератора Мерсена
mt19937 merseenne(rd()) // випадковим стартовим числом.
```

Вона дає найбільш рівномірний розподіл у порівнянні з іншими генераторами, що входять

до бібліотеки `<math.h>`. Тепер, скориставшись залежністю (1), можна сформувати необхідний ряд чисел, що описуються цим законом, а вже потім провести його дослідження на предмет того, наскільки точно вихідний потік чисел $y_i, i = 1, 2, \dots, n$, відповідає експоненціальному розподілу.

Для вирішення цієї задачі скористаємося таким непараметричним методом дослідження і оцінки розподілу, як побудова гістограми і визначення на її основі показника якості моделювання [9]. Щоб побудувати гістограму, необхідно по-перше мати вихідний статистичний набір чисел $y_i, i = 1, 2, \dots, n$, на виході моделі, що повинен відповідати обраному розробником закону розподілу і, по-друге, визначити розмір інтервалів та їх кількість. Проблема полягає у тому, що оптимальний розмір цих показників суттєво залежить від типу розподілу, закладеного в ідею моделі.

У перше дослідження стохастичних процесів методом гістограм описав у 1926 році Стержес у роботі [10], де визначив формулу, згідно з якою, оптимальна кількість інтервалів, на які розділяється варіаційний ряд чисел y_i , визначається як

$$k = 1 + \lfloor \log_2 n \rfloor. \quad (2)$$

Ця формула була виведена для «ідеального» біноміального розподілу і об'єму статистичної вибірки, розмір якої дорівнює ступеню двійки. Вона добре працює лише для симетричних розподілів.

Оскільки формула (2) не враховує стандартного відхилення вибірки, у 1979 році в роботі [11] Скотом було обґрунтовано величину півінтервалу як:

$$h = 3,5 \cdot S \cdot n^{-1/3}, \quad (3)$$

а, Фрідман і Діаконіс у 1981 році у роботі [10] запропонували метод, що дає величину інтервалу, виражену через міжквартильний ранг (відстань між кінцем першого і початком останнього квартилю IQ):

$$h = 2 \cdot (IQ) \cdot n^{-1/3}. \quad (4)$$

Ширина інтервалу, розрахована за формулою (6), буде дещо менше ніж ширина, визначена за формулами (2) і (3), а кількість інтервалів визначається як

$$k = \frac{y_{\max} - y_{\min}}{h}, \quad (5)$$

де $y_{\max}$ і $y_{\min}$ максимальне і мінімальне значення

членів варіаційного ряду, відповідно. Зважаючи на це, є підстави вважати, що точність оцінки розрахунків буде вищою. Це важливо саме у разі, коли очікуваний розподіл має несиметричний характер.

Справа у тому, що необхідно отримати статистичний матеріал, для якого у кожний інтервал гістограми попадає не менше чотирьох чисел вихідного потоку y_i [6], а у разі несиметричних розподілів у «хвості» гістограм потрапляє вкрай мала їх кількість. Але якщо вони туди потрапляють, то потрібного усереднення їх кількості не відбувається, а це, у свою чергу, суттєво потворить точності оцінки якості моделювання.

У разі використання для визначення якості моделювання критерію Пірсона, показник якості визначається як

$$\chi^2 = \frac{(n_i - n_i')^2}{n_i},$$

де n_i і n_i' – це кількість випадкових чисел вихідного потоку, що попала в i -й інтервал та їх очікувана кількість, відповідно.

Якщо очікувана кількість чисел n_i' у «хвості» варіаційного ряду дуже мала, а реальна кількість n_i – ціле число, то складова частина критерію χ^2 , що відповідає i -тому інтервалу у «хвості» гістограми суттєво збільшує його загальну величину і наближає до критичного значення $\chi^2_{кр.}$. У таблиці 1 наведено дані про розподіл попадання 250-ти випадкових величин експоненціального потоку з параметром $\lambda = 0,15$. Розмір інтервалу розраховано за правилом (4), а кількість інтервалів за правилом (5).

Таблиця 1 – Розподіл попадання випадкових чисел в інтервали гістограми

№	$y_{\min}$	$y_{\max}$	P_i	n_i	n_i'	$(n_i - n_i')^2/n_i'$
1	0	3	0,354172848637097	80	89	0,910112359550562
2	3	6	0,228734441925381	61	57	0,280701754385965
3	6	9	0,147722913047252	36	37	0,027027027027027
4	9	12	0,0954034681243366	25	24	0,0416666666666667
5	12	15	0,0616141500488818	20	15	1,666666666666667
6	15	18	0,0397920910097158	14	10	1,6
7	18	21	0,0256988127835781	8	6	0,666666666666667
8	21	24	0,0165969910534268	2	4	1
9	24	27	0,0107187874532302	2	3	0,333333333333333
10	27	30	0,00692248396698409	1	2	0,5
11	30	33	0,0044707281007527	1	1	0

З даної таблиці видно, що починаючи з 8-го інтервалу (8, 9 і 10-й інтервали) при кількості попадань випадкових чисел менше 4-х, внесок у загальний показник χ^2 , більший ніж у перших інтервалах, де кількість попадань складає декілька десятків, і є підстави сумніватись у достовір-

ності цих складових.

Можливим рішенням проблеми могло б бути видалення із вибірки статистичного матеріалу тих його членів, де їх очікується менше 4-х, але у цьому разі доведеться перераховувати величину самого інтервалу h та їх кількості k . Втім це необов'язково, оскільки якщо ця зміна незначна і на результат обчислень це суттєво не вплине.

На рисунку 1 наведена гістограма, яка побудована на основі перших семи інтервалів. У разі, якщо провести обчислення показників по всій вибірці статистичного матеріалу, показник $\chi^2 = 7,0261744 > \chi^2_{кр.} = 3,3512843$. Показник, розрахований на основі перших семи інтервалів залишається меншим ніж критичне його значення ($\chi^2 = 5,3251128 > \chi^2_{кр.} = 5,1928411$), але він стоїть значно ближче до критичного порогу і у більшості реалізацій процесу показник χ^2 задовольняє вимогам $\chi^2 < \chi^2_{кр.}$.

Рисунок 1 – Гістограма, побудована за результатами експерименту

У розглянутому прикладі критичний поріг розрахований виходячи із точності вимірювання у 5%, але, зазвичай, моделювання виконується на перших етапах проектування систем і не вимагає такої точності.

Висновки

Проведені дослідження показали, що реалізація задач моделювання з використанням сучасних спеціалізованих пакетів прикладного програмного забезпечення вимагає значних обчислювальних ресурсів. Виходячи із цього, комп'ютерні моделі стохастичних процесів краще виконувати за допомогою загальних засобів програмування на мовах, які дозволяють створити економічний програмний код. Найкращим варіантом слід вважати мову C++, через те що для неї розроблені найкращі, з точки зору рівномірності розподілення, генератори псевдовипадкових чисел.

Для оцінки якості моделювання в роботі обґрунтовано вибір непараметричного методу гістограм, що дозволяє на основі великої вибірки

оцінювати параметри стохастичного процесу, а також отримати візуальну інформацію про форму розподілу процесу.

Сама оцінка виконувалась на основі обчислення показника якості χ^2 -Пірсона, вибір якого обумовлений саме тим, що його легко обчислити на основі параметрів гістограм.

Показано, що розмір інтервалу розподілення варіаційного ряду, у разі несиметричних розподілень краще визначати за формулою Фрідмана-Діаконіса, оскільки вона дає меншу величину інтервалу, і це підвищує достовірність оцінки. Взагалі, якщо такий інтервал великий, то не буде наглядного образу розподілу, а, якщо він дуже малий, то фільтрування випадкових величин, що не вписуються у закон розподілу, відбувається слабо.

Список використаних джерел (References)

1. Devroye, Luc. Non-uniform random variate generation : / L. Devroye. New York : Springer-Verlag, 1986. 857 p. ISBN 0-387-96305-7. Available via <<http://www.eirene.de/Devroye.pdf>> [accessed 07-August-2020].
2. Baron, M. Probability and Statistics for Computer Scientists : / M. Baron, USA CRC Press by Taylor & Francis Group, LLC : University of Texas at Dallas Richardson, 2014. ISBN 9781138044487. Available via <<https://prakashgautam.com.np/dipit02/books/rps-michael-byron.pdf>> [accessed 07-August-2020].
3. Sanda Micula, Ioana D. Pop. (2010) Simulations of continuous random variables and Monte Carlo methods. Journal of information systems & operations management, Vol. 72(3), 435-447. Available via <<http://jisom.rau.ro/Vol.10%20No.2%20-%202016/JISOM-WI16-A16.pdf>> [accessed 07-August-2020].
4. A Statistical Test Suite for the Validation of Random Number Generators and Pseudo Random Number Generators for Cryptographic Applications. NIST Special Publication 800-22. April 2010. Available via <<https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-22r1a.pdf>> [accessed 07-August-2020].
5. Saito, M. An Application of Finite Field: Design and Implementation of 128-bit Instruction-Based Fast Pseudorandom Number Generator. Dept. of Math. Graduate School of Science, February 9th, 2007. Available via <<http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/SFMT/M062821.pdf>> [accessed 07-August-2020].

07-August-2020].

6. Knuth, D. E. The Art of Computer Programming. Volume 2. Seminumerical Algorithms. 3rd edition / D. E. Knuth. Boston, Mass, USA : Addison-Wesley, Longman Publishing, 1997. 762 p. ISBN 0-201-89683-4. Available via <https://doc.lagout.org/science/0_Computer%20Science/2_Algorithms/The%20Art%20of%20Computer%20Programming%20%28vol.%202_%20Seminumerical%20Algorithms%29%20%283rd%20ed.%29%20%5BKnuth%201997-11-14%5D.pdf> [07-August - 2020].

7. Taylor, J. R. An introduction to error analysis / J. R. Taylor. University of Colorado : University Science Books. 1982. 327 p. ISBN 0-935702-42-3. Available via <https://getthisat.files.wordpress.com/2016/04/introduction_to_error_analysis_taylor_john.pdf> [accessed 07-August-2020].

8. Richard V. Field, Jr., Stochastic Models: Theory and Simulation. Sandia National Laboratories Albuquerque, NM 87185-0346, March 2008.

Available via <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1071.3216&rep=rep1&type=pdf>> [accessed 07-August-2020].

9. Scott, D.W. (1979) On optimal and data-based histograms. Biometrika, 66, 605–610. Available via <<http://www.stat.cmu.edu/~rnugent/PCMI2016/papers/ScottBandwidth.pdf>> [07-August-2020].

10. Freedman, D. and Diaconis, P. (1981) On this histogram as a density estimator: L2 theory. Zeit. Wahr. ver. Geb., 57, 453–476. Available via <https://bayes.wustl.edu/Manual/FreedmanDiaconis1_1981.pdf> [07-August-2020].

11. Sturges, H. (1926) The choice of a class-interval. J. Amer. Statist. Assoc., 21, P. 65–66. Available via <<https://www.tandfonline.com/doi/abs/10.1080/01621459.1926.10502161>> [07-August-2020].

Надійшла до редакції 10.08.2020