Электронный научно-технический журнал Октябрь 2011
http://www.bru.mogilev.by

# ОСОБЕННОСТИ ПРОГРАММНОГО МОДУЛЯ ПОСТРОЕНИЯ КРИВЫХ ПЛОТНОСТЕЙ РАСПРЕДЕЛЕНИЯ ПИРСОНА ДЛЯ ИССЛЕДУЕМОЙ ПОСЛЕДОВАТЕЛЬНОСТИ ДАННЫХ [1]

*Е. А. Якимов, Е. М. Борчик, А. А. Ковалевич,* **О. М. Демиденко**

Аннотация. В статье представлена методика построения функции плотности распределения семейства Пирсона, реализованная в программном модуле BelSim2#.random. Показаны основные теги файла model.xml, приведен пример работы программного модуля.

Ключевые слова: имитационная модель, обобщенное распределение Пирсона, статистический критерий, закон распределения.

## 1. Введение

Пусть в ходе имитационных экспериментов получена последовательность данных, представленная выборкой

$$X = \{ x_i \mid x_i \in R, i = 1, \ldots, n \}.$$

Необходимо построить функцию плотности распределения (кривую), наилучшим образом описывающую выборку $X$ на заданном интервале $[a, b]$.

Предлагается применение плотностей обобщенного распределения Пирсона с последующей проверкой комплексом из трех статистических критериев (Пирсона, Колмогорова-Смирнова, Мизеса) соответствия закону распределения с построенной плотностью. Семейство кривых Пирсона включает семь основных типов и три частных случая распределений: равномерное, нормальное, экспоненциальное.

## 2. Методика построения функции плотности распределения семейства Пирсона

Методика построения функции плотности распределения семейства Пирсона, соответствующей эмпирическому распределению выборки $X$, реализована в программном модуле «BelSim2#.random» [1, 2] и состоит из следующих этапов:

Этап 1. Определение точечных оценок выборки $X$.

Рассчитываются следующие необходимые для работы системы точечные оценки выборки $X$: минимальное и максимальное значения выборки $Min$, $Max$; размах $Range$; начальный момент первого порядка $\nu_1$; центральные моменты до четвертого порядка включительно $\mu_0, \ldots, \mu_4$; дисперсия $D$; среднеквадратическое отклонение $\sigma = \sqrt{D}$; коэффициенты асимметрии и эксцесса, введенные Пирсоном, $\beta_1 = \mu_3^2 / \mu_2^3$, $\beta_2 = \mu_4 / \mu_2^2$; а также в соответствии со статистическим пакетом STATISTICA: $\gamma_3 = \sqrt{\beta_1}$, $\gamma_4 = \beta_2 + 3$.

[1] Работа выполнена по проекту Ф09М-171 при финансовой поддержке Белорусского республиканского фонда фундаментальных исследований

Этап 2. Построение функции $f(x)$ плотности распределения Пирсона в соответствии с классификацией.

Шаг 2.1. Классификация типа кривой плотности распределения Пирсона.

Для определения типа кривой плотности распределения Пирсоном рассматривается показатель

$$\aleph = \left(\beta_1(\beta_2+3)^2\right)\Big/\left(4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)\right). \quad (1)$$

Значение показателя $\aleph < 0$ соответствует типу I; $\aleph = 0$ – типам II, VII; $0 < \aleph < 1$ – определяет тип IV; $\aleph = 1$ – тип V; $\aleph > 1$ – тип VI; $\aleph = \pm\infty$ тип III. Для типа II выполняются условия: $\beta_1 = 0$, $\beta_2 \neq 3$; для типа VII: $\beta_1 = 0$, $\beta_2 = 3$. При $(\beta_1, \beta_2) = (0; 1{,}8)$ имеет место равномерное распределение; при $(\beta_1, \beta_2) = (0, 3)$ – нормальное; при $(\beta_1, \beta_2) = (4, 9)$ – экспоненциальное. Частный случай распределения типа I – бета-распределение 1-го рода; для типа II – равномерное распределение; для типа III – гамма-распределение и распределение хи-квадрат; для распределений Типа VI – бета-распределение 2-го рода и распределение Фишера-Снедекора; для типа VII – распределение Стьюдента.

Для оценки значений параметров функции плотности распределения $f(x)$ применяется классический метод моментов. Все значения параметров функции $f(x)$ выражаются через моменты $\mu_0, \ldots, \mu_4$. Для построенной функции $f(x)$, в соответствии с методикой, предложенной Пирсоном, определяется нормирующий множитель $N$, расчет которого производится исходя из условия

$$N \cdot \int_a^b f(x) = 1, \quad (2)$$

где $[a, b]$ – интервал, в пределах которого производится построение кривой.

Шаг 2.3. Проверка гипотезы принадлежности выборки $X$ закону распределения с плотностью $f(x)$.

Для проверки гипотезы выбраны следующие критерии согласия: $\chi^2$ Пирсона, $\lambda$ Колмогорова-Смирнова, $\omega^2$ Мизеса. Вначале к ($X$, $f(x)$) применяются критерии $\chi^2$ и $\lambda$, сохраняются результаты их работы. Если логические значения результатов работы критериев $\chi^2$ и $\lambda$ эквивалентны, то на этом этап статистической проверки гипотезы оканчивается. Иначе – дополнительно применяется критерий $\omega^2$, результат работы которого принимается в качестве заключения о проверке гипотезы.

Сохраняемые результаты работы критериев $\chi^2$, $\lambda$, $\omega^2$:

1) наблюдаемые значения критериев:

$$\Delta_{\chi^2} \in R_+ \cup \{0\},\ \Delta_\lambda \in R_+ \cup \{0\},\ \Delta_{\omega^2} \in R_+ \cup \{0\};$$

2) критические значения критериев:

$$\Delta^{\text{кр}}_{\chi^2} \in R_+,\ \Delta^{\text{кр}}_{\lambda} \in R_+,\ \Delta^{\text{кр}}_{\omega^2} \in R_+;$$

3) отношение наблюдаемых значений критериев к критическим:

$$dL_1 = \Delta_{\chi^2} \Big/ \Delta^{\text{кр}}_{\chi^2};\ dL_2 = \Delta_\lambda \Big/ \Delta^{\text{кр}}_{\lambda},\ dL_3 = \Delta_{\omega^2} \Big/ \Delta^{\text{кр}}_{\omega^2}; \quad (3)$$

4) логические результаты работы критериев:

$$bL_1 = \begin{cases} true, & dL_1 < 1; \\ false, & dL_1 \geq 1, \end{cases} \quad bL_2 = \begin{cases} true, & dL_2 < 1; \\ false, & dL_2 \geq 1, \end{cases} \quad bL_3 = \begin{cases} true, & dL_3 < 1; \\ false, & dL_3 \geq 1. \end{cases} \quad (4)$$

Этап 3. В случае отклонения на этапе 2 статистическими критериями гипотезы о принадлежности выборки $X$ закону распределения с плотностью $f(x)$ производится перебор с последующей проверкой комплексом статистических критериев $\chi^2$, $\lambda$, $\omega^2$ всех основных типов кривых семейства за исключением самой $f(x)$, проверенной на этапе 2.

Этап 4. По запросу пользователя производится построение функций распределений для трех частных случаев (равномерного, нормального, экспоненциального) с последующей проверкой комплексом статистических критериев $\chi^2$, $\lambda$, $\omega^2$.

Этап 5. По результатам работы комплекса статистических критериев производится выбор кривой плотности распределения, наилучшим образом описывающей выборку.

Выбор производится на множестве тех кривых $\{f_i(x)\}$, $i=1,\dots,|f_i(x)|$, которые не отклонены статистическими критериями, то есть для результатов (4) выполняется одно из условий:

$$bL_1[i] \wedge bL_2[i] = true, \tag{5}$$

$$bL_1[i] \wedge bL_2[i] \vee bL_3[i] = true, \tag{6}$$

где $[i]$ – номер функции $f_i(x)$; $bL_1,\dots,bL_3$ – результаты проверки $f_i(x)$ по (4).

Функция плотности распределения $f^*(x)$ наилучшим образом описывает на интервале $[a,b]$ выборку $X$ тогда и только тогда когда выполняется условие:

$$(f^*(x) = f_{i_0}(x) \in \{f_i(x)\})[\rho(X, f_{i_0}(x)) = \min\{\rho(X, f_i(x)) \mid i=1,\dots,|f_i(x)|\}], \tag{7}$$

где $\rho(X, f_i(x))$ – евклидова метрика вида (8), если для всех $i=1,\dots,|f_i(x)|$ выполняется условие (5) (то есть, применялось только 2 критерия: $\chi^2$ и $\lambda$):

$$dL = \rho(X, f_i(x)) = \sqrt{(dL_1[i])^2 + (dL_2[i])^2},\ i=1,\dots,|f_i(x)|, \tag{8}$$

либо $\rho(X, f_i(x))$ – евклидова метрика вида (9), если существуют такие $f_i(x)$, для которых выполняется условие (6) и при этом не выполняется условие (5) (то есть, применялось 3 критерия: $\chi^2$, $\lambda$, $\omega^2$):

$$dL = \rho(X, f_i(x)) = \sqrt{(dL_1[i])^2 + (dL_2[i])^2 + (dL_3[i])^2},\ i=1,\dots,|f_i(x)|. \tag{9}$$

Для расчета значений евклидовой метрики (9) по результатам $dL_1,\dots,dL_3$ вида (3) необходимо произвести дополнительный расчет 3-го критерия ($\omega^2$) для всех таких функций $f_i(X)$, $i=1,\dots,|f_i(x)|$, для которых выполняется условие (5), поэтому применялось только 2 критерия.

В том случае, если все построенные кривые отклонены статистическими критериями, нет возможности построения кривой плотности распределения семейства Пирсона, описывающей данную выборку. В этом случае рекомендуется исключение выбросов и/или разделение исходной выборки на несколько однородных выборок с последующим построением на каждой из них своей функции плотности распределения.

Для построения функции плотности распределения, наилучшим образом описывающей выборку $X$ на интервале $[a,b]$, проводится эксперимент с участием программного модуля «BelSim2#.random», на вход которого подаётся файл model.xml с исходными данными, значениями параметров и списком откликов модели.

### 3. Основные теги файла model.xml

Файл модели model.xml содержит следующие основные теги: <data> – данные модели; <parameters> – параметры модели; <responses> – отклики модели.

В теге <data> файла модели model.xml сохраняются следующие данные:

1) <array> – множество данных (выборка), состоящее из отдельных элементов <element> $x_i \in X \subset R$; 2) <length> – объем выборки.

Управляющие параметры <parameters> модели «BelSim2#.random:

– <isNearestTypesForcedVerify> – признак необходимости принудительной проверки (на этапе 3) всех ближайших типов кривых к заданной кривой. Значение по умолчанию: False.

– <isNormVerify> – признак необходимости принудительной проверки (на этапе 4) нормального распределения. Значение по умолчанию: True.

– <isExpVerify> – признак необходимости принудительной проверки (на этапе 4) экспоненциального распределения. Значение по умолчанию: True.

– <isRectVerify> – признак необходимости принудительной проверки (на этапе 4) равномерного распределения. Значение по умолчанию: True.

– <confidenceLevel> – уровень значимости применяемых статистических критериев. Значение по умолчанию: 0,05.

– <integralEps> – точность, с которой вычисляется интеграл при расчете нормирующего множителя $N$ кривой $f_i(X)$ по формуле (2) численным методом Симпсона. Значение по умолчанию: 0,01.

– <isKolmogorovCrit> – Параметр выбора критерия согласия Колмогорова-Смирнова для проверки. Значение по умолчанию: False

– <isPirsonCrit> – Параметр выбора критерия согласия Пирсона для проверки. Значение по умолчанию: False

– <isMizesCrit> – Параметр выбора критерия согласия Мизеса для проверки. Значение по умолчанию: False

– <isKolmogorovPirsonMizesCrit> – Признак проверки гипотезы о распределении выборки X по закону распределения с построенной плотностью комплексом из 3-х стат критериев: Колмогорова-Смирнова, Пирсона, Мизеса. Значение по умолчанию: True.

Результаты работы программного модуля расчета кривой плотности распределения случайной составляющей в последовательности данных «BelSim2#.random» сохраняются в теге откликов <responses> файла model.xml:

– <StatPointsDataEstimations> – точечные оценки выборки.

– <FunctionByClussification> – функция $f_0$ плотности распределения выборки, построенной в соответствии с классификацией.

– <NearestFunctions> – построенные функции $f_i$ плотностей распределения выборки ближайших типов по отношению к кривой функции плотности $f_0$, построенной в соответствии с классификацией. Максимально возможное количество таких кривых $f_i$: 9 (6 типов кривых Пирсона и 3 частных случая распределений – нормальное, экспоненциальное, равномерное).

– <BestFunction> – функция плотности распределения $f^*$, наилучшим образом описывающая выборку.

Тег <function>, выводимый в теги <FunctionByClussification>, <NearestFunctions>, <BestFunction> содержит следующие сведения о функциях: <functionType> – тип функции плотности распределения; <functionForm> – функция плотности распределения; <Kolmogorov>, <Xi2>, <W2> – результаты проверки гипотезы о распределении выборки $X$ по заданному закону $f_i(X)$ критериями $\chi^2$, $\lambda$, $\omega^2$, соответственно.

## 4. Экспериментальная часть

Пусть $X$ – специальным образом заданная выборка, состоящая из 43-х равномерно распределенных на интервале [0,1] элементов, сгенерированных с использованием функции СЛЧИСЛ() в MS EXCEL.

Необходимо построить функцию плотности распределения, наилучшим образом описывающую на интервале $[0,1]$ выборку $X$.

Значения параметров «BelSim2#.random»: isNormVerify = True; isExpVerify = True; isRectVerify = True; isNearestTypesForcedVerify = False; confidenceLevel = 0,05; integralEps = 0,01, <isKolmogorovPirsonMizesCrit> = True.

Результаты работы «BelSim2#.random» приведены ниже.

Этап 1. Точечные оценки выборки $X$: $Min = 0$, $Max = 1$, $Range = 1$, $\nu_1 = 0{,}489$, $\mu_2 = 0{,}085$, $\mu_3 = 0{,}003$, $\mu_4 = 0{,}015$, $D = 0{,}085$, $\sigma = 0{,}291$, $\beta_1 = 0{,}015$, $\beta_2 = 2{,}040$, $\gamma_3 = 0{,}122$, $\gamma_4 = 5{,}040$.

Этап 2. Показатель классификации Пирсона $\aleph$ вида (1): $\aleph = -0{,}005975375043888358$. Учитывая точность классификации $\varepsilon = 0{,}1$: $\aleph = 0$, $\beta_1 = 0$, $1 < \beta_2 < 3$. В соответствии с классификацией по ($\aleph$, $\beta_1$, $\beta_2$) тип функции плотности распределения Пирсона – тип 2. Функция плотности распределения:

$$f_2(x) = 1{,}191 \cdot (1 - (x - 0{,}489)^2 / 0{,}360)\text{^}0{,}625.$$

В соответствии с результатами работы критериев $\chi^2$, $\lambda$ (таблица 1) нет оснований для отклонения гипотезы о распределении выборки $X$ по закону с плотностью $f_2(x)$, поэтому Этап 3 пропускается.

Этап 4. Построение функций плотностей равномерного $R(x)$, экспоненциального $E(x)$, нормального $N(x)$ распределений:

$$R(x) = 1{,}0;$$

$$E(x) = 2{,}043 \cdot \exp(-2{,}043 \cdot x);$$

$$N(x) = 1 / (0{,}291 \cdot \sqrt{2\pi}) \cdot \exp(-1 \cdot (\mathrm{x} - 0{,}489)^2 / 0{,}169).$$

Результаты проверки критериями $\chi^2$, $\lambda$, $\omega^2$ гипотезы о принадлежности выборки $X$ законам распределения с плотностями $R(x)$, $E(x)$, $N(x)$ приведены в таблицах 1, 2.

По результатам $\chi^2$, $\lambda$ нет оснований для отклонения гипотез о распределениях выборки $X$ в соответствии с построенными функциями распределения.

Этап 5. Для выбора кривой, описывающей выборку $X$ наилучшим образом на интервале [0,1] в соответствии с критерием (7), нет необходимости в дополнительном расчете значений критерия $\omega^2$ для построенных функций.

Таблица 1

| Параметр | $f_2(x)$ | $R(x)$ | $E(x)$ | $N(x)$ |
|---|---|---|---|---|
| $\Delta_{\chi^2}$ | 1,833 | 1,791 | 8,964 | 3,313 |
| $\Delta_{\chi^2}^{кр}$ | 5,991 | 9,488 | 9,488 | 7,815 |
| $dL_1$ | 0,306 | 0,189 | 0,944 | 0,424 |
| $bL_1$ | T | T | T | T |
| $\Delta_{\lambda}$ | 0,249 | 0,356 | 0,931 | 0,423 |
| $\Delta_{\lambda}^{кр}$ | 1,358 | 1,358 | 1,358 | 1,358 |

Таблица 2

| Параметр | $f_2(x)$ | $R(x)$ | $E(x)$ | $N(x)$ |
|---|---|---|---|---|
| $dL_2$ | 0,183 | 0,262 | 0,685 | 0,315 |
| $bL_2$ | T | T | T | T |
| Итог | T | T | T | T |
| $dL$ | 0,357 | 0,323 | 1,167 | 0,528 |
| $\min(dL)$ | – | + | – | – |
| $f^*(x)$ | – | + | – | – |

Так как для функции $R(x)$ значение $dL$ (Таблица 2) метрики (8) минимально, $f^*(x) = R(x)$. Выборка $X$ на интервале [0,1] описывается наилучшим образом равномерным распределением $R(x) = 1{,}0$ (частный случай кривых Пирсона типа 2).

## 4. Заключение

Ожидаемый результат определения функции плотности распределения подтвердился, что доказывает работоспособность процедуры, предложенной и реализованной в программном модуле.

BelSim2#.random может быть использован автономно или в составе программно-технологического комплекса имитации сложных систем на этапе эксплуатации имитационной модели.

### Литература

1. **Якимов, А. И.** Технология имитационного моделирования систем управления промышленных предприятий : монография / А. И. Якимов. – Могилев: Белорус.-Рос. ун-т, 2010. – 304 с.: ил.
2. Программный модуль расчета кривой плотности распределения случайной составляющей в последовательности данных «BelSim2#.random» : свидетельство о регистрации компьютерной программы № 306 / Е. А. Якимов, А. А. Ковалевич, Е. М. Борчик, В. В. Башаримов. – Минск: НЦИС, 2011. – Заявка № С20110024. – Дата подачи: 04.04.2011.

**Якимов Евгений Анатольевич**
Аспирант кафедры Автоматизированные системы управления
Белорусско-Российский университет, г. Могилев
Тел.: +375(222) 25-24-47
E-mail: e-soft@bk.ru
**Борчик Екатерина Михайловна**
Аспирант кафедры Автоматизированные системы управления
Белорусско-Российский университет, г. Могилев
Тел.: +375(29) 746-83-10
E-mail: katrinb15@gmail.com
**Ковалевич Александр Александрович**
Аспирант кафедры Автоматизированные системы управления
Белорусско-Российский университет, г. Могилев
Тел.: +375(29) 745-99-08
E-mail: kavalevich@tut.by
**Демиденко Олег Михайлович**
Проректор по научной работе
Гомельский государственный университет имени Франциска Скорины, г. Гомель
Тел.: +375(232) 60-30-02
E-mail: demidenko@gsu.by