УДК 519.8

В.М. Сафонов[1]

МОДЕЛИ ОГРАНИЧЕННЫХ СЛУЧАЙНЫХ ВЕЛИЧИН В ЗАДАЧАХ ИДЕНТИФИКАЦИИ КЛАВИАТУРНОГО ПОЧЕРКА

Предлагается универсальная модель для аппроксимации законов распределений ограниченных случайных величин, используемая в исследовании для описания временных параметров клавиатурного почерка.

В статье предложено описание элементов клавиатурного почерка (продолжительность нажатия клавиши, период времени между нажатиями клавиш, период времени между отпусканием и нажатием следующей клавиши и т.д.), которые являются случайными величинами. Поскольку данные величины являются ограниченными, обоснована их аппроксимация бета-распределениями. Особенностью данного распределения является то, что оно описывает непрерывные случайные величины на ограниченном интервале.

Ключевые слова: *клавиатурный почерк, ограниченные случайные величины, закон распределения случайной величины, бета-распределение.*

V.M. Safonov

MODELS OF BOUNDED RANDOM VARIABLES IN PROBLEMS OF IDENTIFICATION OF HANDWRITING KEYBOARD

A universal model for the approximation of laws of distributions bounded random variables used in the study to describe the timing of keyboard handwriting is proposed

The article describes the elements of handwriting keyboard (pressing duration, time between keystrokes, time between releasing and pressing the next key, etc.), which are random variables. Since these values are limited, their approximation by beta distributions are justified. The peculiarity of this distribution is that it describes the continuous random variables in a limited range.

Keywords: *handwriting keyboard, limited random variables, law of random variable, beta distribution.*

Введение

В настоящее время актуальны задачи совершенствования систем аутентификации пользователей информационных систем (ИС), а также повышения результативности расследования компьютерных инцидентов. Одним из способов их решения является идентификация клавиатурного почерка (КП) пользователя ИС. Клавиатурный почерк – это поведенческие закономерности ввода конкретным оператором текста. КП фиксируется значениями биометрических характеристик. Такими характеристиками являются: продолжительность нажатия клавиши, период времени между нажатиями клавиш, период времени между отпусканием и нажатием следующей клавиши, среднее значение продолжительности нажатия клавиши, среднее квадратическое отклонение продолжительности нажатия клавиш и т.п. Совокупность значений этих характери-

[1] Адъюнкт кафедры математического и программного обеспечения Военно-космической академии им. А.Ф. Можайского.

стик, полученная при вводе текста оператором, является реализацией случайного вектора, а его распределение зависит от индивидуальных особенностей личности.

Как правило, большинство элементов рассматриваемого случайного вектора являются независимыми, а следовательно плотность распределения случайного вектора может быть представлена произведением плотностей распределения отдельных его элементов. Для описания закона распределения такого случайного вектора необходимо использовать законы распределения элементов вектора, которые могут быть получены на основе обработки зарегистрированных характеристик клавиатурного почерка. Носители распределения указанных характеристик принципиально ограничены, поэтому представляется целесообразным искать аппроксимирующие распределения в классе распределений ограниченных случайных величин (СВел).

Анализ форм кривых бета-распределения привел к выводу о том, что именно с его помощью удобно аппроксимировать большинство законов распределения (ЗакРас) ограниченных СВел.

Плотность $\phi_{\hat{x}}(x)$ и функция $F_{\hat{x}}(x)$ четырёхпараметрического бета-распределения определяются следующими выражениями [1]:

$$\phi_{\hat{x}}(x) = \phi_{\hat{x}}^{[B]}(x;a,b,\alpha,\beta) = \frac{(x-a)^{\alpha}(b-x)^{\beta}}{B(\alpha+1,\beta+1)(b-a)^{\alpha+\beta+1}}\Pi(x;a,b), \quad (1)$$

$$F_{\hat{x}}(x) = F_{\hat{x}}^{[B]}(x;a,b,\alpha,\beta) = \int_{-\infty}^{x}\phi_{\hat{x}}^{[B]}(x';a,b,\alpha,\beta)dx' = \int_{-\infty}^{x}\frac{(x'-a)^{\alpha}(b-x')^{\beta}dx'}{B(\alpha+1,\beta+1)(b-a)^{\alpha+\beta+1}}, \quad (2)$$

где $\hat{x}$ – случайная величина ($\wedge$ – символ случайного объекта); $a < b$; $a,\ b \in (-\infty, \infty)$ – параметры положения распределения (минимальное и максимальное значения СВел); $\alpha \geq -1$, $\beta \geq -1$ – параметры формы распределения; $B(\alpha+1,\beta+1) = \int_0^1 v^{\alpha}(1-v)^{\beta}dv = \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}$ – интеграл Эйлера первого рода; $\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt$ – интеграл Эйлера второго рода (гамма-функция) при z, являющимся натуральным числом $\Gamma(z) = (z-1)!$;

$$\Pi(x;a,b) = \begin{cases} 0, \text{ при } & x < a \text{ или } x \geq b \\ 1, & \text{при } a \leq x < b \end{cases}$$ – селектор интервала.

При $a = 0$; $b = 1 \Rightarrow b - a = 1 \Rightarrow$

$$\Rightarrow \phi_{\hat{x}_0}(x) = \phi_{\hat{x}}^{[B]}(x;0,1,\alpha,\beta) = \frac{x^{\alpha}(1-x)^{\beta}}{B(\alpha+1,\beta+1)}\Pi(x;0,1), \quad (3)$$

$$F_{\hat{x}_0}(x) = F_{\hat{x}}^{[B]}(x;0,1,\alpha,\beta) = \int_{-\infty}^{x}\phi_{\hat{x}}^{[B]}(x';0,1,\alpha,\beta)dx' \cdot \Pi(x;0,1) + \Delta(x-1) = \int_0^x \frac{x'^{\alpha}(1-x')^{\beta}dx'}{B(\alpha+1,\beta+1)} \cdot \Pi(x;0,1) + \Delta(x-1), \quad (4)$$

где $\hat{x}_0$ – нормированная по величине $(b-a)$ размаха выборки СВел $\hat{x}$;

$$\Delta(x-d) = \begin{cases} 0, \text{ при } & x < d \\ 1, \text{ при } & x \geq d \end{cases}$$ – селектор луча.

Бета-распределение с плотностью (3) и функцией (4) называется каноническим [2]. Примеры графиков его плотностей для различных значений α и β приведены на рис. 1.

Приведенные графики плотностей распределений показывают широкие возможности предлагаемой модели по аппроксимации ЗакРас, наблюдаемых в природе СВел. Как видно из приведенных примеров, универсальность бета-распределения позволяет аппроксимировать ЗакРас практически всех наблюдаемых в природе СВел, которые являются ограниченными в силу закономерностей процесса их измерения исследователем. Таким образом, в автоматизированных системах, использующих законы распределения, возможно описывать законы распределения только моделью бета-распределения.

Параметры ЗакРас бета-распределения легко определяются на основе следующих выражений [3]:

$$\tilde{\alpha} = \tilde{M}[\hat{x}_0]\left[\frac{\tilde{M}[\hat{x}_0](1-\tilde{M}[\hat{x}_0])}{\tilde{M}[\hat{x}_0^2]-\tilde{M}^2[\hat{x}_0]}-1\right]-1, \quad (5)$$

$$\tilde{\beta} = (1-\tilde{M}[\hat{x}_0])\left[\frac{\tilde{M}[\hat{x}_0](1-\tilde{M}[\hat{x}_0])}{\tilde{M}[\hat{x}_0^2]-\tilde{M}^2[\hat{x}_0]}-1\right]-1, \quad (6)$$

где $\tilde{M}[\hat{x}_0]$; $\tilde{M}[\hat{x}_0^2]$ – оценки первого и второго начальных моментов распределения СВел $\hat{x}_0$, связанной с СВел $\hat{x}$ следующим соотношением:

$$\hat{x} = (b-a)\hat{x}_0 + a. \quad (7)$$

Отметим, что СВел $\hat{x}_0$ подчинена каноническому бета-распределению на интервале $[0,1)$. В программах Mathcad, Microsoft Excel приведены встроенные плотности и функции распределе-

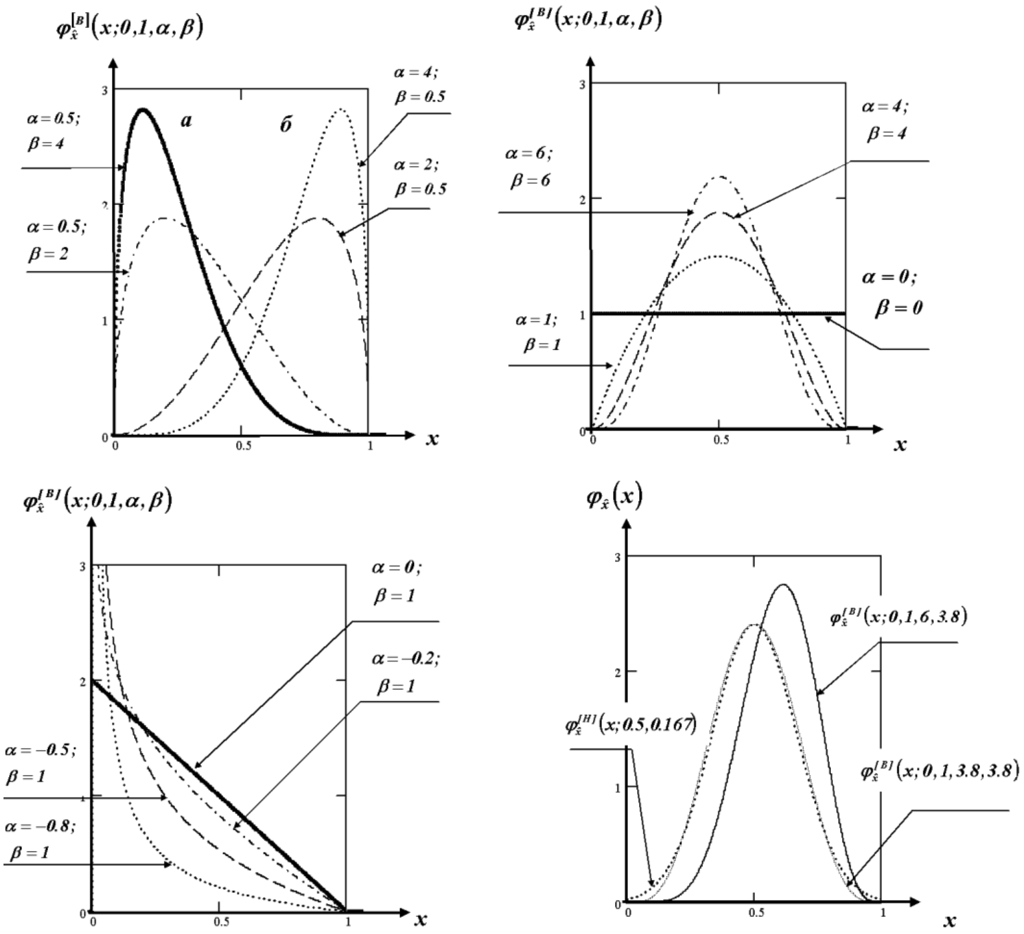

Рис. 1. Семейства кривых канонического бета-распределения, определяемых сочетаниями параметров форм α и β

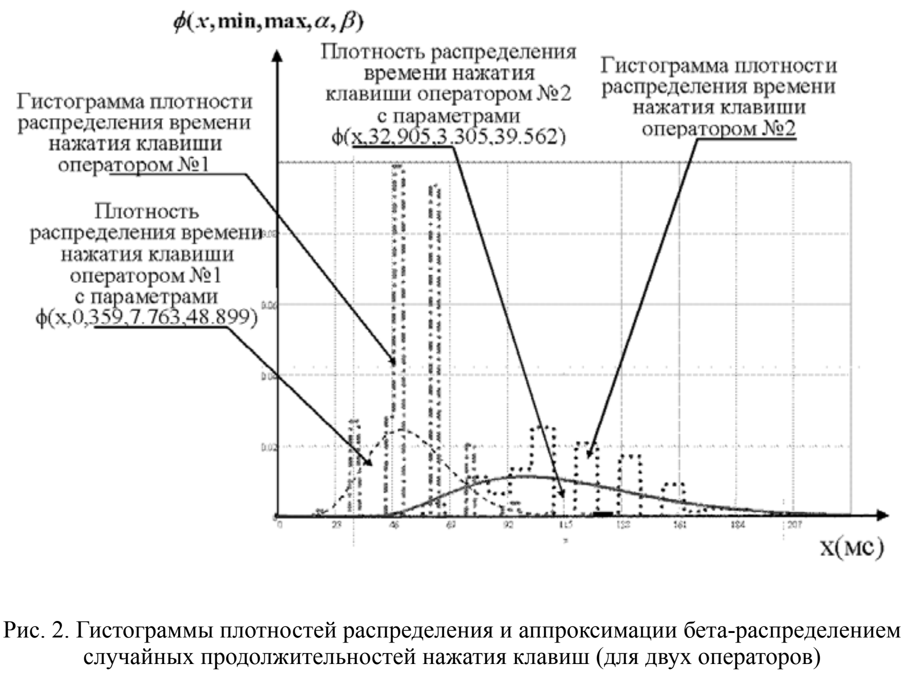

Рис. 2. Гистограммы плотностей распределения и аппроксимации бета-распределением случайных продолжительностей нажатия клавиш (для двух операторов)

ния именно канонического бета-распределения. Следует учитывать, что коэффициенты (α_1, β_1) форм в данных программах связаны с коэффициентами (α, β) форм (в данной статье) следующими соотношениями:

$$\alpha_1 = \alpha + 1, \quad \beta_1 = \beta + 1. \tag{8}$$

Выражение, связывающее плотности произвольно бета-распределенной СВел $\hat{x}$ и связанной с ней канонической СВел $\hat{x}_0$, имеет вид

$$\varphi_{\hat{x}}(x; a, b, \alpha, \beta) = \frac{1}{b-a} \varphi_{\hat{x}_0}\left(\frac{x-a}{b-a}; 0, 1, \alpha, \beta\right) =$$

$$= \frac{1}{b-a} \varphi_{\hat{x}_0}^{[Mathcad]}\left(\frac{x-a}{b-a}; \alpha+1, \beta+1\right). \tag{9}$$

На рис. 2 приведены графики гистограмм плотностей распределений СВел продолжительностей нажатия клавиш для двух различных операторов. Там же приведены графики аппроксимаций бета-распределениями. Эффект «отдельно стоящих столбов гистограмм» вызван дискретностью сообщений таймера. Поскольку продолжительность нажатия клавиши является непрерывной СВел, применение бета-распределения будет более адекватно описывать данную СВел.

Заметим, что при накоплении статистики в течение длительного срока нет необходимости хранить всю выборку. Целесообразно хранить текущие параметры бета-распределения и объём выборки [4]. Так при появлении нового члена x_{m+1} исходной выборки несложно откорректировать параметры (a, b, α, β) бета-распределения по следующей схеме. Определяются значения $\tilde{M}[\hat{x}_0]; \tilde{M}[\hat{x}_0^2]$, соответствующие объему выборки m. Уточняются параметры положения

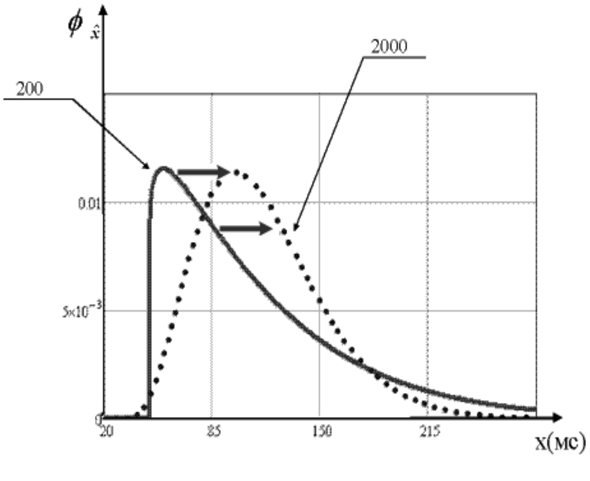

Рис. 3. Уточнение плотностей распределения СВел при продолжительности нажатия клавиши при фиксации 200 и 2000 нажатий

распределения по тому, попадает ли x_{m+1} в интервал $[a, b]$. Затем уточняются оценки упомянутых математических ожиданий для объема выборки $m + 1$. Далее с использованием выражений (5)–(6) определяются оценки (α, β), соответствующие объему выборки $m + 1$.

На рис. 3 приведены графики аппроксимаций бета-распределением при различных объемах выборки при ее накоплении.

На рис. 4 приведены совмещенные двумерные плотности распределения случайных векторов $\langle \hat{x}, \hat{y} \rangle$ разных операторов, где СВел $\hat{x}$ – это продолжительность нажатия клавиши, а $\hat{y}$ – это период времени между нажатиями клавиш. Поскольку коэффициент корреляции между $\hat{x}$ и $\hat{y}$ близок к нулю, следовательно, эти СВел можно считать независимыми, поэтому аппроксимация двумерного бета-распределения для оператора № 1 будет иметь вид

$$\varphi_{\langle \hat{x}, \hat{y} \rangle}(x, y) = \varphi_{\hat{x}}\left(\frac{x-a}{b-a}; 0, 1, \alpha, \beta\right) \times$$

$$\times \varphi_{\hat{y}}\left(\frac{y-a}{b-a}; 0, 1, \alpha, \beta\right). \tag{10}$$

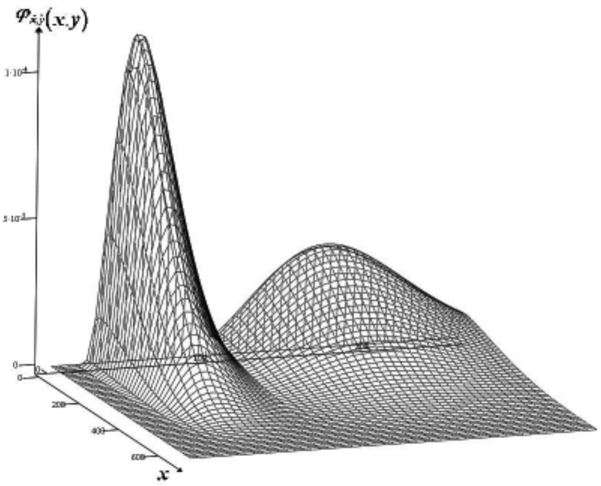

Рис. 4. Графики двумерных ЗакРас случайного вектора $\langle \hat{x}, \hat{y} \rangle$ различных операторов

Области расположения двумерного признака для приведенного случая различаются существенно, что позволяет легко идентифицировать этих операторов с использованием методов теории статистического принятия решения. Однако на практике, в условиях большого числа операторов, решение задачи идентификации пользователя по клавиатурному почерку сопряжено с большими трудностями. Но описание решения этой задачи не является целью данной статьи.

Заключение

На основании вышесказанного можно сделать следующие выводы.

1. Универсальность бета-распределения позволяет аппроксимировать законы распределения практически всех наблюдаемых в природе случайных величин.

2. В системах идентификации клавиатурного почерка успешно апробирована технология аппроксимации законов распределения четырёхпараметрического бета-распределения.

3. Использование программного модуля идентификации параметров закона распределения случайных величин в классе бета-распределений позволит унифицировать задачу аппроксимации закона распределения в программных средствах, а в случае адаптации закона к изменяющейся исходной выборке значительно сократит необходимый объем памяти для ее хранения.

Литература

1. Айвазян С.А. и др. Прикладная статистика: основы моделирования и первичная обработка данных : справочное пособие. – М. : Финансы и статистика, 1983. – 471 с.

2. Петухов Г.Б., Девяткин А.М., Якунин В.И. Идентифицирование законов распределений ограниченных случайных величин. – СПб. : ВКА, 2005. – 30 с.

3. Петухов Г.Б., Якунин В.И. Моделирование случайных ситуационных характеристик на авиационных тренажёрах / Г.Б. Петухов, В.И. Якунин // Изв. ВУЗов. Приборостроение. – 2007. – № 3 – С. 7–11.

4. Маков А.Б., Суворов С.С., Кулешов Ю.В. Динамико-стохастический подход в задачах адаптации методов прогнозирования опасных явлений / А.Б. Маков, С.С. Суворов, Ю.В. Кулешов // Вестник Санкт-Петербургского университета. – СПб. : СПБГУ, 2008. – Сер. № 7. – № 4.