УДК 519.2

МОДЕЛИРОВАНИЕ УСЕЧЕННЫХ РАСПРЕДЕЛЕНИЙ

М.С.Токмачёв, П.П.Рязанцев

Институт электронных и информационных систем НовГУ, tokm@mail.natm.ru

Разработан способ моделирования нового класса вероятностных распределений на основе базового нормального распределения. Приведены формулы расчета функции распределения, характеристической функции и числовых характеристик.

Ключевые слова: нормальное распределение, плотность распределения, функция распределения, характеристическая функция, усеченное распределение

The method of modeling of a density distribution new class is formulated based on a standard normal distribution. The formulae of calculating of the distribution function, the characteristic function and the numerical characteristics are given.

Keywords: normal distribution, density function, distribution function, characteristic function, truncated distribution

Введение

Многие классические вероятностные распределения предполагают распределение вероятностей по бесконечным интервалам значений случайных величин. Использование таких распределений для реальных задач, в которых данные рассредоточены на конечных интервалах, всегда сопряжено с проблемой «хвостов распределений»: либо игнорированием малых значений вероятностей, либо их учетом тем или иным образом. В частности, принято строить усеченные распределения. В классическом случае при усечении распределения на интервал $[a,b]$ плотность распределения вероятностей $f(x)$ корректируется умножением ее на соответствующий нормирующий множитель:

$$g_0(x)=\begin{cases}\dfrac{f(x)}{\int\limits_a^b f(x)dx} & \text{при } x\in[a,b],\\ 0 & \text{при } x<a \text{ или } x>b.\end{cases}$$

Разумеется, усечение распределений может быть и односторонним.

Б.Ф.Кирьяновым [1] предложена следующая модификация: смещение оси абсцисс при классическом усечении распределения так, чтобы кривая плотности в одной из границ, *a* или *b* (при симметричном распределении и симметричном усечении — в обеих), совпала с точкой на этой оси. Такое распределение вероятностей при $f_{\text{ус}}(a)=0$ или $f_{\text{ус}}(b)=0$ более удобно для исследования и интерпретации.

Можно использовать и другие методики. Например, при усечении на $[a,b]$ сдвиг относительно оси ординат кривой плотности $f(x)$ на величину $h=\dfrac{S_1+S_2}{b-a}$, где S_1, S_2 — площади усекаемых криволинейных трапеций («хвостов распределения»). Это преобразование увеличивает значение функции $f(x)$ в каждой точке на одну и ту же величину *h* и тем самым сохраняет форму усеченной кривой.

Очевидно, что существуют и иные способы обеспечения справедливости условия нормировки для полученного усеченного распределения [2]. В представленной работе рассмотрена методика моделирования новых усеченных вероятностных распределений на основе исходного теоретического распределения. В заданных интервалах производится умножение функции плотности $f(x)$ на линейные множители. Введенная операция позволяет подбирать нестандартные теоретические распределения в соответствии с реальными данными. Для базового нормального распределения выведены формулы расчета параметров и числовых характеристик моделируемых распределений, найдены их функции распределения и характеристические функции.

Основные результаты

Пусть задана случайная величина *X* с плотностью распределения вероятностей $f(x)$. Область значений случайной величины *X* включает в себя интервал $[a,b]$. Разобьем $[a,b]$ на два интервала $[a,x_0]$, $[x_0,b]$, в каждом из которых умножим функцию плотности исходного распределения на свой линейный множитель. Тогда функция плотности нового распределения имеет вид

$$g(x)=\begin{cases}(k_1x+c_1)f(x) & \text{при } x\in[a,x_0],\\ (k_2x+c_2)f(x) & \text{при } x\in[x_0,b],\\ 0 & \text{при } x<a \text{ или } x>b.\end{cases} \qquad (1)$$

Чтобы функция $g(x)$ вида (1) являлась плотностью некоторого распределения, достаточно, чтобы она была неотрицательна и удовлетворяла условию нормировки $\int\limits_a^b g(x)dx=1$.

Для однозначного определения коэффициентов k_1, c_1, k_2, c_2 рассмотрим систему четырех уравнений

$$\begin{cases}\int\limits_a^{x_0}(k_1x+c_1)f(x)dx=\alpha,\\ \int\limits_{x_0}^{b}(k_2x+c_2)f(x)dx=\beta,\\ (k_1x_0+c_1)f(x_0)=g(x_0),\\ (k_2x_0+c_2)f(x_0)=g(x_0).\end{cases} \qquad (2)$$

Значения α, β, $g(x_0)$ выбираются исследователем, причем $g(x_0)\ge 0$ (в частности, $g(x_0)=f(x_0)$), вероятности α и β связаны соотношением $\alpha+\beta=1$ (в частности, значения вероятностей α и β можно взять соответственно пропорциональными площадям криволинейных трапеций $\int_a^{x_0} f(x)dx$, $\int_{x_0}^{b} f(x)dx$ в исходном распределении). Последние два уравнения в системе (2) следуют из непрерывности функции $g(x)$ в точке x_0. Варьирование значениями $x_0, g(x_0), \alpha$ позволяет подбирать подходящее распределение в соответствии с реальными данными. Таким образом, числа $a, b, x_0, g(x_0), \alpha$ — параметры моделируемого усеченного распределения при ограничениях $x_0\in[a,b]$, $g(x_0)\ge 0$, $0<\alpha<1$. Для обеспечения неотрицательности функции $g(x)$ — в противном случае она не может быть плотностью никакого распределения — достаточно ограничений $g(a)\ge 0$, $g(b)\ge 0$.

Безусловно, решение системы (2) зависит от типа исходного (базового) распределения с функцией плотности $f(x)$, которая явно присутствует в каждом из уравнений системы.

Теорема 1. Пусть $f(x)$ — плотность гауссовской случайной величины $X: X\sim N(m;\sigma)$ и выполнены условия на параметры: $x_0\in[a,b]$, $g(x_0)\ge 0$, $0<\alpha<1$, $\beta=1-\alpha$. Тогда коэффициенты функции плотности $g(x)$ моделируемого распределения вида (1) удовлетворяют соотношениям

$$k_1=\frac{\alpha-\frac{g(x_0)}{f(x_0)}\left[\Phi\left(\frac{x_0-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right)\right]}{\sigma\left[\varphi\left(\frac{a-m}{\sigma}\right)-\varphi\left(\frac{x_0-m}{\sigma}\right)\right]+(m-x_0)\left[\Phi\left(\frac{x_0-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right)\right]};$$

$$k_2=\frac{\beta-\frac{g(x_0)}{f(x_0)}\left[\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{x_0-m}{\sigma}\right)\right]}{\sigma\left[\varphi\left(\frac{x_0-m}{\sigma}\right)-\varphi\left(\frac{b-m}{\sigma}\right)\right]+(m-x_0)\left[\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{x_0-m}{\sigma}\right)\right]};$$

$$c_1=\frac{g(x_0)}{f(x_0)}-k_1x_0;\quad c_2=\frac{g(x_0)}{f(x_0)}-k_2x_0,$$

$$k_1a+c_1\ge 0,\quad k_2b+c_2\ge 0,$$

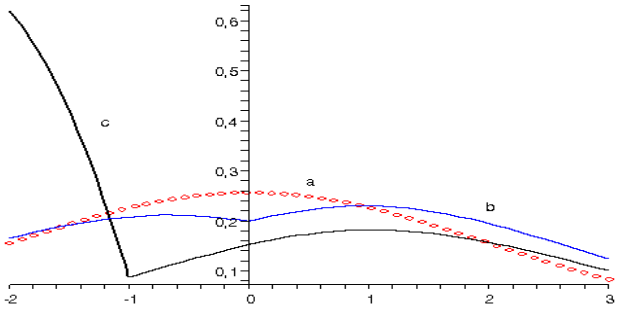
где $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{x^2}{2}}$ — плотность вероятностей стандартного нормального распределения, $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{\frac{x^2}{2}}dx$ — функция Лапласа.

В частности, из теоремы 1 при $X\sim N(0;1)$ и $g(x_0)=f(x_0)$, $a<0$, $b>0$, $x_0=0$ следуют более простые выражения для коэффициентов $g(x)$:

$$k_1=\frac{\sqrt{2\pi}(\alpha-0{,}5+\Phi(a))}{e^{-0{,}5a^2}-1};\quad k_2=\frac{\sqrt{2\pi}(\Phi(b)-\beta-0{,}5)}{e^{-0{,}5b^2}-1};$$

$$c_1=c_2=1.$$

На рис. представлены графики функции $g(x)$ при усечении на одном и том же интервале, как в классическом случае, так и в случаях соотношения (1), при различных значениях параметров.



Графики функции $g(x)$ при различном усечении нормального распределения $N(0;2)$ на интервал [–2, 3]: a) классическое усечение; b) $x_0=0$; $g(x_0)=f(x_0)$; $\alpha=0{,}4$; c) $x_0=-1$; $g(x_0)=2f(x_0)$; $\alpha=0{,}4$.

Как легко заметить, график плотности в модельном распределении может иметь не единственную вершину (кривая b); вершину, смещенную по сравнению с вершиной базового распределения, что определяет концентрацию вероятности на отдельных участках значений (кривая c), и т.д.

Представим и другие характеристики усеченных распределений с функцией плотности типа (1).

Теорема 2. Пусть $f(x)$ — плотность гауссовской случайной величины $X: X\sim N(0;1)$. Плотность усеченного распределения $g(x)$ удовлетворяет соотношению (1). Тогда функция распределения имеет вид

$$G(x)=\begin{cases}0 \text{ при } x\le a,\\ k_1[\varphi(a)-\varphi(x)]+c_1[\Phi(x)-\Phi(a)] \text{ при } a<x\le x_0,\\ \alpha+k_2[\varphi(x_0)-\varphi(x)]+c_2[\Phi(x)-\Phi(x_0)] \text{ при } x_0<x\le b,\\ 1 \text{ при } x>b.\end{cases}$$

Теорема 3. Пусть $f(x)$ — плотность гауссовской случайной величины $X: X\sim N(0;1)$, плотность усеченного распределения $g(x)$ удовлетворяет соотношению (1). Тогда характеристическая функция этого усеченного распределения имеет вид

$$\chi(t)=k_1[\varphi(a)e^{iat}-\varphi(x_0)e^{ix_0t}]+$$
$$+e^{-\frac{t^2}{2}}(ik_1t+c_1)[\Phi(x_0-it)-\Phi(a-it)]+$$
$$+k_2[\varphi(x_0)e^{ix_0t}-\varphi(b)e^{ibt}]+$$
$$+e^{-\frac{t^2}{2}}(ik_2t+c_2)[\Phi(b-it)-\Phi(x_0-it)].$$

Зная характеристическую функцию распределения, по формуле $\chi^{(n)}(0)=i^nM(X^n)$, связывающей производные характеристической функции с соответствующими начальными моментами, можно вычис-

лить, в частности, математическое ожидание и дисперсию.

Теорема 4. Пусть $f(x)$ — плотность гауссовской случайной величины $X: X \sim N(0;1)$, плотность усеченного распределения $g(x)$ удовлетворяет соотношению (1). Тогда математическое ожидание и дисперсия усеченного распределения подчиняются соотношениям

$$M(X) = k_1[a\varphi(a) - x_0\varphi(x_0)] + k_1[\Phi(x_0) - \Phi(a)] - c_1[\varphi(x_0) - \varphi(a)] + k_2[x_0\varphi(x_0) - b\varphi(b)] + k_2[\Phi(b) - \Phi(x_0)] - c_2[\varphi(b) - \varphi(x_0)];$$

$$D(X) = k_1[a^2\varphi(a) - x_0{}^2\varphi(x_0)] + c_1[\Phi(x_0) - \Phi(a)] + c_1[\varphi(a) - \varphi(x_0)] + k_2[x_0{}^2\varphi(x_0) - b^2\varphi(b)] + c_2[\Phi(b) - \Phi(x_0)] + c_2[\varphi(x_0) - \varphi(b)] - M^2(X).$$

Аналогично, путем дальнейшего дифференцирования характеристической функции (третья и четвертая производные в нуле), можно рассчитать коэффициенты асимметрии и эксцесса.

Использование в качестве базовой функции $f(x)$ плотности именно стандартного нормального закона не принципиально. Для произвольных m и σ расчетные формулы в модельном распределении отличаются лишь большей громоздкостью. Также $f(x)$ может иметь и другой тип распределения, который технически влияет только на сложность вычисления интегралов при использовании $g(x)$ вида (1) и расширяет множество модельных распределений.

Заключение

Представленный способ моделирования определяет широкий класс вероятностных распределений, в основу формирования которого положены следующие шаги: выбор базового распределения (функция $f(x)$), усечение (выбор параметров a и b), разбиение на интервалы (выбор параметра x_0), изменение вероятностей на полученных интервалах (выбор параметров $g(x_0)$ и α), расчет коэффициентов в соотношении (1).

Варьирование тремя параметрами группировки (a, b, x_0) и двумя параметрами концентрации ($g(x_0)$ и α) приводит к большому разнообразию распределений, в том числе бимодальных с заданными координатами вершин, что весьма важно для практических исследований. Компьютерная реализация методики позволяет легко строить распределения указанного типа при различных базовой функции и параметрах, а также находить численные характеристики распределений и осуществлять подгонку модели по реальным данным.

1. Кирьянов Б.Ф., Токмачев М.С. Математические модели в здравоохранении. В. Новгород: НовГУ, 2009. 350 с.
2. Токмачев М.С., Рязанцев П.П. Преобразование усеченных распределений // Математика в вузе: Тр. XXI Междунар. науч.-практ. конф. СПб: Петербургский гос. ун-т путей сообщения, 2009. С. 123-124.