УДК 519.657:519.245:519.688

АППРОКСИМАЦИЯ ЭМПИРИЧЕСКИХ ФУНКЦИЙ ПОЛИНОМАМИ ВЫСШИХ ПОРЯДКОВ

П. В. Заикин, М. А. Погореловский, В. С. Микшина
Сургутский государственный университет, zaikinpv@gmail.com

В работе рассматриваются подходы к аппроксимации экспериментальных данных, полученных в ходе измерений и являющихся случайной величиной. Обоснован выбор типа семейства кривых для аппроксимации эмпирических данных. Разработан алгоритм подбора кривой Пирсона эмпирическому распределению. Произведены расчеты параметров кривых основных типов для экспериментальных данных различной физической природы. Показана адекватность применения данного подхода.

Ключевые слова: функция плотности распределения, экспериментальные данные, ряды Грама–Шарлье, кривые Пирсона, аппроксимация.

APPROXIMATION OF EXPERIMENTAL DATA FUNCTIONS BY HIGH ORDER POLINOMIAL

P. V. Zaikin, M. A. Pogorelovsky, V. S. Mikshina
Surgut State University, zaikinpv@gmail.com

The paper considers approaches to the experimental approximation of data being a random variable and obtained during the measurement. The choice of the type of Pearson's curve for empirical data approximation is justified. The algorithm for choice of Pearson's curve fitting the empirical distribution was developed. Calculations of parameters of most common Pearson curves to the experimental data of different physical nature were made. Adequacy of the approach is demonstrated.

Keywords: density distribution function, Gram–Charlier A series, Pearson curves, approximation.

Исследователь в естественных и технических науках часто сталкивается с явлениями, поведение которых может быть описано с помощью статистических данных, т.е. сведениями о числе объектов в какой-либо более или менее обширной совокупности, обладающих теми или иными признаками. При этом ставится задача выявить и исследовать закономерности, которым подчиняются реальные процессы. Найденные закономерности имеют не только теоретическую ценность, они широко применяются на практике – в планировании, управлении и прогнозировании.

Существуют задачи, в которых обработке экспериментальных данных с целью построения моделей «сложных систем» (эмпирических зависимостей) должна предшествовать предварительная обработка, содержание которой, в основном, состоит в отсеивании грубых погрешностей измерений и в проверке соответствия распределения результатов нормальному закону. Решение отдельного класса задач требует не только проверки соответствия распределения экспериментальных данных нормальному закону, но и определения к какому же закону распределения плотности вероятности относятся эти экспериментальные данные. Это задачи, связанные с распознаванием образов, задачи диагностики в медико-биологических системах, моделированием кинетики многокомпонентных смесей химико-технологических процессов нефтепереработки [2] и т.п.

В данной работе рассмотрены методы аппроксимации экспериментальных данных полиномами высших порядков типа гипергеометрических и определения аналитического вида функции плотности распределения, а также описан алгоритм подбора параметров функции распределения на примере кривых Пирсона.

Постановка задачи аппроксимации закона распределения экспериментальных данных может быть сформулирована следующим образом:

Пусть есть выборка случайной величины $X=(x_1,x_2,x_3,\ldots,x_n)$ фиксированного размера n. Необходимо подобрать закон рапределения (вид и параметры), который в статистическом смысле соответствовал бы наблюдениям.

Решение задачи можно разбить на несколько этапов:

1) Подбор закона распределения
2) Определение параметров функции плотности распределения

3) Проверка адекватности выбора функции плотности распределения

Рассмотрим подробно каждый из шагов.

Подбор вида распределения

В общем случае для определения закона распределения используется три основных метода: "эмпирический"выбор из нескольких стандартных, использование специальных рядов и использование семейства универсальны распределений.

В случае с "эмпирическим"подходом нам требуется по гистограмме распределения "угадать"какому виду оно соответствует.

Использование рядов Тейлора и Фурье затруднительно, т.к. они не обладают необходимыми свойствами: в начале и конце вариационного ряда, для которого строится функция плотности распределения, эта функция должна стремиться к нулю, а в середине ряда она должна достигать своего максимума. Иначе говоря, скорость изменения этой функции или выражающая эту скорость первая производная $\frac{dy}{dx}$ должна равняться нулю в трех точках: в начале и конце ряда, где $y = 0$, и в средине ряда, где функция достигает максимума.

Известны ряды, основанные на полиномах Чебышева-Эрмита, например, ряды Грама-Шарлье:

$$f_A(x) = f(x) + \sum_{k=3}^{n} a_k f^{(k)}(x), \tag{1}$$

где x — нормированное значение случайой величины,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

$f^{(k)}$ – k-такая производная.

Особенностью ряда Грама – Шарлье можно назвать близость к вариациям нормального распределения. Пытаясь использовать ряд с другими видами распределений мы сталкиваемся с серьезными проблемами: ряд может вести себя нерегулярно (увеличение количества членов ряда иногда снижает точность аппроксимации); ошибки аппроксимации возрастают с удалением от центра распределения; сумма конечного числа членов ряда при большой асимметрии распределения приводит к отрицательным значениям функций, особенно на краях распределений. [3] Поэтому ряд Грама-Шарлье применяют только при весьма умеренном коэффициенте асимметрии, не превышающем 0,7. Следовательно, применение рядов тоже не обеспечивает необходимой общности решения задач аппроксимации.

Наиболее оптимальным вариантом, особенно в случае необходимости автоматической обработки значений, является использование универсальных семейств кривых. Таких как, например, семейство кривых Джонсона и семейство кривых Пирсона.

Кривые Джонсона так же, как и ряд Грама-Шарлье основаны на нормальном распределении. Вследствие чего, они обладают одними и теми же недостатками. [5]

Функция плотности распределения Джонсона задаётся следующим образом:

$$f(x) = \frac{\beta}{\sqrt{2\pi}} e^{-\frac{1}{2}(\alpha+\beta g(x))^2} g'_x(x), \tag{2}$$

где β и α – параметры, а $g(x)$ – непрерывная неограниченная монотонно возрастающая функция задающая конкретное распределение семейства.

Наиболее интересными, в смысле аппроксимации экспериментальных распределений, являются кривые Пирсона. Они описываются дифференциальным уравнением [4]:

$$\frac{dy}{y} = -\frac{x-M}{b_2 x^2 + b_1 x + b_0} dx, \tag{3}$$

где началом отсчета x служит середина ряда, мода – M.

Дифференциальное уравнение (3) выражает общие свойства функций распределения. Постоянные коэффициенты, входящие в уравнение (3), можно выразить при помощи начальных и центральных моментов [4].

$$b_0 = \frac{4\beta_2 - 3\beta_1}{10\beta_2 - 12\beta_1 - 18} \mu_2, \tag{4}$$

$$b_1=\frac{\sqrt{\mu_2}\sqrt{\beta_1}\,(\beta_2+3)}{10\beta_2-12\beta_1-18}, \tag{5}$$

$$b_2=\frac{2\beta_2-3\beta_1-6}{10\beta_2-12\beta_1-18}, \tag{6}$$

где

$$\beta_1=\frac{\mu_3^2}{\mu_2^3}, \qquad \beta_2=\frac{\mu_4}{\mu_2^2}, \tag{7}$$

а μ_2,μ_3,μ_4 соответственно центральные моменты 2, 3 и 4 порядков случайной величины x.
Вид функции плотности распределения определяется уравнением

$$b_2x^2+b_1x+b_0=0. \tag{8}$$

Введем обозначение:

$$k=\frac{b_1^2}{4b_2b_0}. \tag{9}$$

Обычно коэффициент k называют каппой Пирсона. С учетом (4) и (5), запишем:

$$k=\frac{\beta_1(\beta_2+3)^2}{4\,(2\beta_2-3\beta_1-6)\,(4\beta_2-3\beta_1)} \tag{10}$$

Из формул (4-10) видно, что кривые Пирсона определяются при помощи первых четырех моментов.

Значение k определяет тип кривой плотности распределения. В таблице 1 показаны значения k и соответствующие этим значениям кривые Пирсона [4].

Таблица 1

Типы и уравнения кривых Пирсона. y_0,a_i,γ,ν — параметры соответствующей кривой

Значение k	Тип кривой Пирсона	Общий вид уравнения кривой
$k<0$	I	$y=y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$
$0<k<1$	IV	$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m}\cdot e^{-\upsilon\arctan\frac{x}{a}}$
$k>1$	VI	$y=y_0(x-a)^{m_1}\cdot x^{-m_2}$
$k=0$, $\mu_4=3$	Нормальное распределение	$y=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2\sigma^2}}$
$k=0$, $\mu_4<3$	II	$y=y_0\left(1-\frac{x^2}{a^2}\right)^{m}$
$k=0$, $\mu_4>3$	VII	$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m}$
$k=1$	V	$y=y_0\cdot x^{-p}\cdot e^{-\frac{\gamma}{x}}$
$k=\pm\infty$	III	$y=y_0\left(1+\frac{x}{a}\right)^{\gamma a}\cdot e^{-\gamma x}$

Алгоритм определение типа кривой Пирсона соответствующей эмпирическому распределению данных заключается в следующем:

1) По статистическому ряду эмпирических данных определяют первые четыре момента распределения (начальный момент и центральные моменты).
2) Для целей нормирования, определяется разрядность ряда случайной величины c – т.е. среднюю длину отрезка между x_i,x_{i+1}.
3) Определяется медиана X_a случайной величины, как вариант на котором достигается середина частотного ряда.
4) Вычисляется выравнивающий ряд, смещенный относительно центра ряда и нормированный согласно разрядности ряда. $x_i' = (X_i - X_a)/c$
5) Рассчитывается каппа Пирсона k.
6) Определяется тип кривой по классификации Пирсона.
7) Определяются параметры кривой плотности распределений. Для каждого из типов кривых существует уникальный порядок определения параметров. Каждый параметр вычисляется согласно ранее рассчитанных моментов, а также других характеристик случайной величины.

Для расчета подбора кривой распределения соответствующей эмпирическим данным была разработана компьютерная программа на языке Python [6]

В таблице 2 представлены результаты подбора кривой на основании экспериментальных данных.

В гистограммах (табл.2) графически показаны частотные характеристики ряда, по оси x – значение нормированной случайной величины, по оси $y-$ частота. На графиках отображены кривые, являющиеся функцией плотности распределения данного ряда. $\int_{-\infty}^{+\infty} f(x)dx = 1$, а в четвертом столбце приводятся уравнения функций плотности распределения (табл.2).

Для исследования метода подбора теоретической кривой распределения использовались данные шести экспериментов: биомедицинские данные (1 - возраст научных работников, 2 – межимпульсный интервал энцефалограммы), метеорологические данные (3 – скорость ветра в Московской области), данные технической обработки деталей (4 – хрупкость эбонита, 5- остаточное удлинение болтового железа), данные материаловедения (6 – процент содержания железа в мартеновских сплавах).

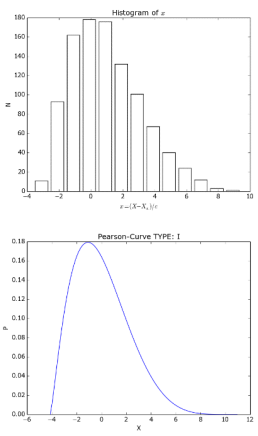
Приведенные расчеты показали, что применение кривых Пирсона для аппроксимации эмпирических данных различной природы даёт хорошие результаты. Адекватность аппроксимации, измеренная методом χ^2, колеблется от минимального значения 1,64 для медицинских данных, до 5,8 для распределения хрупкости материалов.

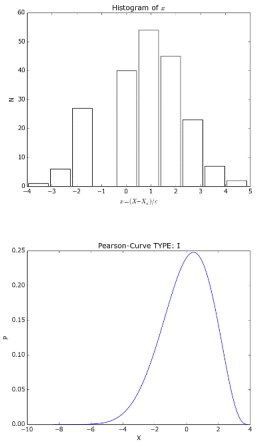
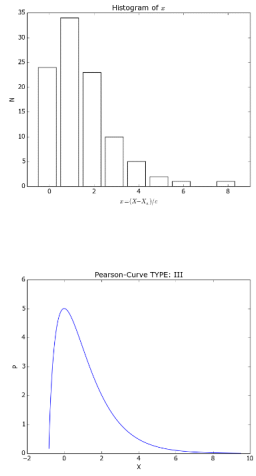
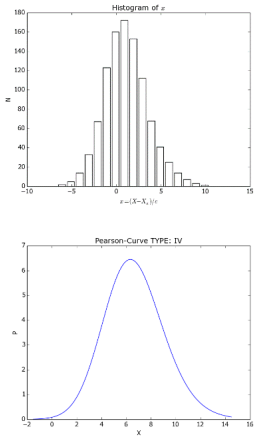
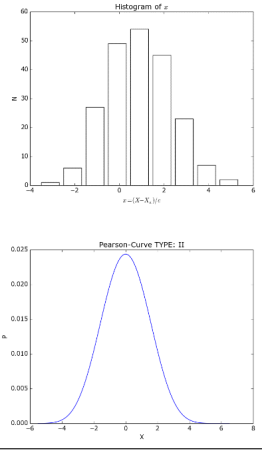
Таким образом показано, что метод подбора типа кривой экспериментальному распределению основанный на использовании универсальных кривых Пирсона пригоден для аппроксимации экспериментальных данных любой физической природы.

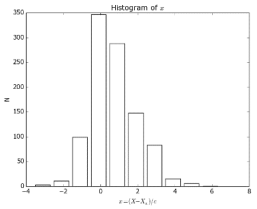
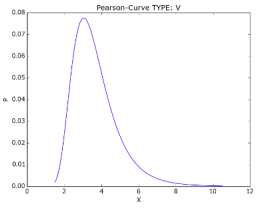
Разработан программный продукт, реализующий статистический анализ данных, подбор типа кривой и расчет её параметров. Использованные технологии разработки компьютерной программы, основанные на принципе модульности, позволяют использовать данный программный продукт для последующих исследований.

Таблица 2

Результаты расчета

№ п/п	Экспериментальные данные	Каппа Пирсона	Тип	Гистограмма и график кривой	Уравнение
1	Возраст научных работников СССР [4]	$k=$ -0.25814	I		$y =$ $0{,}164\left(1+\frac{x}{4{,}126}\right)^{1.2751}\cdot$ $\left(1-\frac{x}{10{,}959}\right)^{5{,}043}$

№ п/п	Экспериментальные данные	Каппа Пирсона	Тип	Гистограмма и график кривой	Уравнение
2	Ряд распределения межимпульсного интервала энцефалограммы человека [2]	$k = -0.16128$	I		$y = 0{,}24\left(1+\frac{x}{8{,}42}\right)^{7{,}448}\left(1-\frac{x}{3{,}848}\right)^{2{,}861}$
3	Ряд распределения скорости ветра в Московской области (в м/сек) в процентах (n = 40000) [4]	$k = -57.499$	III		$y = 5{,}001\left(1+\frac{x}{0{,}753}\right)^{0{,}753*0{,}823}\cdot e^{-0{,}823x}$
4	Ряд распределения хрупкости эбонита (в $\frac{кг^{*}см}{см^{3}}$) [4]	$k = 0.201$	IV		$y = 13{,}301\left(1+\frac{x^2}{19{,}982^2}\right)^{-13{,}301}\cdot e^{19{,}05\arctan\frac{x}{19{,}982}}$
5	Ряд распределения остаточного удлинения болтового железа (%) [4]	$k = -0.004 \approx 0$	II		$y = 0{,}024\left(1-\frac{x^2}{7{,}119^2}\right)^{10{,}317}$

№ п/п	Экспериментальные данные	Каппа Пирсона	Тип	Гистограмма и график кривой	Уравнение
6	Ряд распределения процентного содержания кремния в рельсовых мартеновских плавках, подвергнутых копровому испытанию (в ‰, n = 4850) [4]	$k = 0.9299 \approx 1$	V	 	$y = 3{,}305 \cdot x^{-11{,}574} \cdot e^{-\frac{35{,}288}{x}}$

ЛИТЕРАТУРА

1. Григоренко В. В., Лысенкова С. А., Гавриленко Т. В. Математическое моделирование ситуации возникновения критических состояний в организме человек // Вестник кибернетики. 2015. № 2. С. 106–111.
2. Заикин П. В., Микшина В. С. Непрерывный подход в моделировании кинетики реакций многокомпонентной смеси // Вестник кибернетики. 2014. № 2. С. 25–31.
3. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. М. : Наука, 1976. 736 с.
4. Митропольский А. К. Техника статистических вычислений. М. : Наука, 1971. 570 с.
5. Хан Г., Шапиро С. Статистические модели в инженерных задачах. М. : Мир, 1969. 400 с.
6. Программа подбора параметров функций распределен из семейства кривых Пирсона:свидетельство о гос. регистрации прогр. для ЭВМ № 2014611354 Российская Федерация, Заикин П.В., правообладатель ГБОУВПО СурГУ, – Зарегистрировано в Реестре программ для ЭВМ 31.01.2014, заявка N2013661220 от 04.12.2013; опубл. 20.02.2014, ОБПБТ № 2 (88).