

УДК 519.234.33, 519.676, 519.254

Т.В. ВАСИЛЬЕВА

АВТОМАТИЗАЦИЯ ИДЕНТИФИКАЦИИ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ НА ОСНОВЕ АППРОКСИМАТИВНОГО АНАЛИЗА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Предложен алгоритм автоматизации идентификации законов распределения на основе аппроксимативного анализа экспериментальных данных. Метод основан на разложении эмпирических данных в ряд Фурье в пространстве L_2 через полиномы Лежандра. Приведены два необходимых и достаточных критерия для автоматического выдвижения гипотезы о законе распределения. Изложенный метод рассмотрен на подробном примере.

Ключевые слова: автоматизация статистического анализа, аппроксимация, закон распределения, доказательство гипотез, критерии согласия.

Введение. Статистический анализ получил свое новое развитие с появлением ЭВМ. Компьютеризация статистических исследований, возможности анализа и хранения информации с помощью ЭВМ сделали статистический анализ доступным для широкого практического применения в любых прикладных отраслях, имеющих дело с большими объемами данных. Все это привело к появлению большого количества статистических пакетов на рынке программ. Наилучший выбор статистического пакета зависит от характера решаемых задач, объема, специфики обрабатываемых данных, квалификации пользователей и т.д. Каждый пакет содержит набор основных статистических методов, изучаемых в курсах статистического анализа и эконометрики. Специализированные пакеты наряду с основными методами реализуют новые методы статистической обработки данных или дают более глубокий анализ одного или нескольких методов. Актуальной задачей остается автоматизация новых статистических методов, популяризация их для массового применения, приведение интерфейсов статистических пакетов к новому уровню удобства использования конечным пользователем, разработка алгоритмов, способных реализовать введение интеллектуальной составляющей для анализа статистической информации. Одной из таких нерешенных задач является автоматизация выдвижения гипотезы о виде закона распределения статистических данных. До сих пор эта задача решается на уровне визуального анализа гистограммы. В данной статье предлагается методика автоматизации, и приводится компьютерная реализация выдвижения гипотезы о законе распределения на основе проведенного исследования в рамках диссертации автора, ранее результаты и основные выводы по предложенной методике можно найти в публикациях [1,2].

Обзор методов идентификации функции распределения. Выбор подходящего аналитического выражения с неизвестными параметрами является самым сложным этапом процедуры сглаживания. Это объясняется тем, что задача плохо формализована, всегда существует опасность, что рассмотрены не все варианты законов распределения, решения необходимо принимать по ограниченному набору данных.

На практике задачу идентификации решают на основе анализа:

- 1) гистограммы и ее соответствия кривой теоретического распределения;
- 2) числовых характеристик и их соответствия значениям теоретических числовых характеристик;
- 3) фазовых портретов (структурных функций) [9].

С точки зрения автоматизации этот процесс полностью полагается на мнение исследователя. Легко можно автоматизировать второй пункт, но на основе соответствия только числовых характеристик сложно сделать выводы о законе распределения. Анализ же первого и второго пунктов не автоматизирован. Третий пункт также полагается на визуальное сравнение и выводы

исследователя. Этот пробел в автоматизации предварительного анализа закона распределения эмпирических данных мы попытались исправить с помощью разработанного в рамках представленной работы метода.

Метод идентификации законов распределения на основе аппроксимативного анализа экспериментальных данных. Приведем алгоритм метода идентификации законов распределения на основе аппроксимации статистических данных в ряд Фурье в пространстве L_2 через полиномы Лежандра.

Алгоритм идентификации функции распределения.

1. Рассмотрим эмпирические данные, объем выборки n варианта $x \in (a, b)$.
2. Перейдем от отрезка (a, b) к отрезку $(0, 1)$ с помощью формулы:

$$t = \frac{x - a}{b - a}. \quad (1)$$

3. Частным случаем разложения функции в гильбертовом пространстве является разложение функции в ряд Фурье в пространстве $L_2 [-1, 1]$ через многочлены Лежандра [8]. В нашем случае будем рассматривать разложение в ряд Фурье через сдвинутые многочлены Лежандра в пространстве $L_2 [0, 1]$. Аналитическое представление эмпирической функции распределения, построенной на отрезке $(0, 1)$, выражается формулой:

$$F^*(t) = \sum_{i=0}^{\infty} c_i^* \cdot P_i(t), \quad (2)$$

где многочлены Лежандра находим по формуле:

$$P_i(t) = \frac{1}{i! 2^i} \frac{d^i}{d(2t-1)^i} ((2t-1)^2 - 1)^i. \quad (3)$$

Многочлены Лежандра программно вычисляем с помощью рекуррентного соотношения [10]:

$$(i+1)P_{i+1}(2t-1) = (2i+1)(2i-1)P_i(2t-1) - iP_{i-1}(2t-1);$$

$$P_i(1-2t) = (-1)^i P_n(2t-1), \quad P_i(1) = 1, \quad P_i(-1) = (-1)^i.$$

Коэффициенты c_i^* находим по формуле

$$c_i^* = (2i+1) \cdot \int_0^1 F^*(t) \cdot P_i(t) dt. \quad (4)$$

И в силу того, что

$$\int_0^1 t^k F^*(t) dt = \frac{1}{k+1} (1 - \mu_{k+1}),$$

каждый коэффициент $c_i^* = c_i^*(\mu)$ выражается через эмпирические начальные моменты распределения. Приведем пять первых коэффициентов разложения:

$$c_0^* = (1 - \mu_1);$$

$$c_1^* = 3(\mu_1 - \mu_2);$$

$$c_2^* = 5(-2\mu_3 + 3\mu_2 - \mu_1);$$

$$c_3^* = 7(-5\mu_4 + 10\mu_3 - 6\mu_2 + \mu_1);$$

$$c_4^* = -9(14\mu_5 - 35\mu_4 + 30\mu_3 - 10\mu_2 + \mu_1).$$

4. Очевидно, следующей нашей задачей является оценка погрешности, возникающей при использовании этого метода, и нахождение необходимого количества элементов ряда Фурье. Дадим определение *верхних и нижних сумм Дарбу*. Пусть G_i и g_i – соответственно верхняя и нижняя границы изменения $F(x)$ в элементарном интервале $[x_i, x_{i+1}]$ для разбиения Z отрезка $[a, b]$. Числа $S(Z) = \sum_{i=1}^n G_i(x_{i+1} - x_i)$ и $s(Z) = \sum_{i=1}^n g_i(x_{i+1} - x_i)$ называются соответственно *верхней и нижней суммой Дарбу* разбиения Z [4. С.390].

В силу того, что система полиномов Лежандра ортогональна, но не нормирована, для применения равенство Парсеваля [8] коэффициенты c_i^* необходимо умножить на $\frac{1}{\sqrt{2i+1}}$.

Обозначим $\tilde{c}_i^* = \frac{c_i^*}{\sqrt{2i+1}}$.

Получаем равенство Парсеваля:

$$\sum_{i=0}^k \tilde{c}_i^{*2} + \sum_{i=k+1}^{\infty} \tilde{c}_i^{*2} = \|F^*\|^2.$$

Чтобы оценить количество коэффициентов c_i^* , необходимых при построении эмпирической функции распределения $F^*(t)$, найдем следующую разность:

$$\|F^*\|^2 - \sum_{i=0}^k \tilde{c}_i^{*2} = \sum_{i=k+1}^{\infty} \tilde{c}_i^{*2} < \varepsilon.$$

Из сходимости ряда Фурье при $k \rightarrow \infty$ следует $\varepsilon \rightarrow 0$. Таким образом, задавая требуемое $\varepsilon > 0$, ε – малое, мы можем оценить количество коэффициентов c_i^* . Функция распределения монотонно возрастающая, следовательно, верхняя и нижняя границы изменения $F^*(t)$ в элементарном интервале $[t_i, t_{i+1}]$ соответственно равны:

$$G_i = \omega_{i+1}; \quad g_i = \omega_i,$$

где ω_n – относительная накопленная частота.

В нашей задаче для эмпирической функции $F^*(t)$ численное значение нормы необходимо оценить через верхние и нижние интегральные суммы Дарбу:

$$\sum_{i=1}^{n-1} \omega_i^2(t_{i+1} - t_i) < \|F^*\|^2 = \int_0^1 (F^*)^2 dt < \sum_{i=1}^{n-1} \omega_{i+1}^2(t_{i+1} - t_i).$$

Таким образом, получаем двустороннюю интервальную оценку, дающую погрешность метода

$$\sum_{i=1}^{n-1} \omega_i^2(t_{i+1} - t_i) - \sum_{i=0}^k \tilde{c}_i^{*2} < \|F^*\|^2 - \sum_{i=0}^k \tilde{c}_i^{*2} < \sum_{i=1}^{n-1} \omega_{i+1}^2(t_{i+1} - t_i) - \sum_{i=0}^k \tilde{c}_i^{*2}.$$

Откуда следует выражение для нахождения верхней интервальной оценки:

$$\varepsilon = \sum_{i=1}^{n-1} \omega_{i+1}^2(t_{i+1} - t_i) - \sum_{i=0}^k \frac{c_i^{*2}}{2i+1}. \quad (5)$$

Таким образом, задавая необходимое ε , можно рассчитать количество членов разложения в ряд Фурье. Количество коэффициентов разложения теоретической функции распределения будем брать равным количеству коэффициентов разложения эмпирической функции.

Получаем разложение эмпирической функции распределения через k начальных моментов распределения:

$$F^*(t) = \sum_{i=0}^k c_i^*(\mu) \cdot P_i(t) + o(t^{k+1}). \quad (6)$$

5. Проверим качество аппроксимации критерием ω^2 –Мизеса при заданном уровне значимости α [7]. В общем случае критерий Мизеса является критерием согласия, т.е. критерием проверки близости теоретического и эмпирического распределения. Выдвигается нулевая гипотеза при заданном уровне значимости α

$$H_0: F^*(x) = F(x).$$

В качестве меры различия теоретической функции распределения $F(x)$ и эмпирической $F^*(x)$ по критерию Мизеса (критерию ω^2) выступает средний квадрат отклонений по всем значениям аргумента x

$$\omega_n^2 = \int_{-\infty}^{\infty} [F^*(x) - F(x)]^2 dF(x).$$

Статистику критерия находим по следующей формуле:

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^n \left[F(x_i) - \frac{i-0,5}{n} \right]^2.$$

При неограниченном увеличении n существует предельное распределение статистики $n\omega_n^2$. Задав значение вероятности α , можно определить критические значения $n\omega_n^2(\alpha)$. Проверка гипотезы о законе распределения осуществляется обычным образом: если фактическое значение $n\omega_n^2$ окажется больше критического или равно ему, то согласно критерию Мизеса с уровнем значимости α гипотеза H_0 о том, что закон распределения генеральной совокупности соответствует $F(x)$, должна быть отвергнута.

Для проверки качества аппроксимации будем придерживаться следующего алгоритма:

1) построим расчетную таблицу, состоящую из следующих столбцов:

i	x_i	t_i	$F(t_i)$	F_i^*	Δ_i
-----	-------	-------	----------	---------	------------

В таблице: i – порядковый номер эмпирических данных; x_i – исходные упорядоченные варианты; t_i – пересчитанные варианты по формуле (1). В этой формуле $a = \min(x_i)$, $b = \max(x_i)$; $F(t_i)$ – значения теоретической функции распределения, подсчитанные в каждой точке t_i . В качестве теоретического распределения $F(t)$ выступает аналитическое разложение эмпирических данных в ряд Фурье в пространстве L_2 , представленное формулой (5); F_i^* – значения эмпирической функции распределения. В качестве значений эмпирического распределения принимаются значения, вычисленные по следующей формуле $F_i^* = \frac{i-0,5}{n}$; Δ_i – расчетные данные наблюдаемого значения статистики, находятся по формуле $\Delta_i = (F(t_i) - F_i^*)^2$;

2) наблюдаемое значение статистики $n\omega^2$ находим по формуле: $n \cdot \omega^2 = \frac{1}{12 \cdot n} + \sum_{i=1}^n \Delta_i$;

3) критическое значение находим в зависимости от уровня значимости α в статистических таблицах, представленных, например, в [7]. Критическое значение при уровне значимости $\alpha = 0,05$ равно: $n \cdot \omega^2(0,05) = 0,461$;

4) если $n\omega^2 < n\omega^2(\alpha)$, делаем вывод о том, что полученная аппроксимация не противоречит эмпирическим данным.

6. Аналогично изложенному методу разложения эмпирических данных в ряд Фурье выполняем разложение нескольких теоретических функций. Параметры распределения определяем с помощью статистических оценок методом наибольшего правдоподобия или методом моментов.

Будем рассматривать $F(x)$ не на всей числовой прямой $(-\infty; +\infty)$, а на отрезке $[a, b]$, полагая, что при малом $\gamma > 0$ $F(a) \geq \gamma$ и $1 - F(b) \geq \gamma$.

Приведем основные формулы для разложения в ряд Фурье теоретической функции распределения на основе начальных моментов распределения:

$$t = \frac{x - a}{b - a};$$

$$F(t) = \sum_{i=0}^{\infty} c_i \cdot P_i(t);$$

$$c_i = (2i + 1) \cdot \int_0^1 F(t) \cdot P_i(t) dt.$$

В силу того, что

$$\int_0^1 t^k F(t) dt = \frac{1}{k+1} (1 - M_{k+1}),$$

каждый коэффициент c_i выражается через начальные теоретические моменты M_k распределения.

7. Составим нормированный вектор $C^*(i)$ на основе коэффициентов c_i^* . При подсчете коэффициентов c_i^* используем эмпирические начальные моменты на основе введенных данных.

8. Составим эталонную матрицу $C(i, j)$ на основе теоретических начальных моментов для нескольких, интересующих нас, распределений. Эталонная матрица пересчитывается автоматически на основе статистических оценок параметров распределения.

9. Введем понятие ε – близких функций.

Пусть $F(x)$ и $F^*(x)$ – монотонно возрастающие функции с областью значений $[0, 1]$, назовем $F(x)$ ε – близкой функцией $F^*(x)$ в пространстве L_2 , если можно подобрать такое ε – малое, $\varepsilon \geq 0$, то для $\forall x$ выполняется равенство

$$\|F(x) - F^*(x)\| = \varepsilon.$$

10. Воспользуемся одним из двух критериев ε – близости формы кривой эмпирической функции распределения к кривой теоретического распределения.

Критерий 1. Пусть $F(t)$ и $F^*(t)$ – соответственно теоретическая и эмпирическая функции распределения, разложенные в ряд Фурье на отрезке $[0, 1]$ через многочлены Лежандра в пространстве L_2 ; $\bar{c}$ и $\bar{c}^*$ – вектора коэффициентов, соответствующие этим разложениям. Тогда, если форма кривой $F^*(t)$ ε близка к форме кривой $F(t)$, то необходимо и достаточно, чтобы $\rho(\bar{c}, \bar{c}^*) = \varepsilon$, $\varepsilon \rightarrow 0$, $\varepsilon \geq 0$.

Доказательство. Расстояние между двумя функциями определяется выражением $\rho(F, F^*) = \|F - F^*\|$. Согласно равенству Парсеваля имеем:

$$\sum_{i=1}^{\infty} (c_i - c_i^*)^2 = \|F - F^*\|^2 \Leftrightarrow \rho(F, F^*) = \|F - F^*\| = \sqrt{\sum_{i=1}^{\infty} (c_i - c_i^*)^2} = \rho(c, c^*) = \varepsilon.$$

Что и требовалось доказать.

Критерий 2. Пусть $F(t)$ и $F^*(t)$ – соответственно теоретическая и эмпирическая функции распределения, разложенные в ряд Фурье на отрезке $[0,1]$ через многочлены Лежандра в пространстве L_2 ; $\bar{c}$ и $\bar{c}^*$ – нормированные вектора коэффициентов, соответствующие этим разложениям. Тогда, если форма кривой $F^*(t)$ близка к форме кривой $F(t)$, то необходимо и достаточно, чтобы $|\langle \bar{c}, \bar{c}^* \rangle| = \delta$, $\delta \leq 1$, $\delta \rightarrow 1$.

Доказательство. Пусть F является ε – близкой функцией F^* , тогда $\|F - F^*\| = \varepsilon$. Покажем, что в этом случае $\exists \delta \leq 1$, $\delta \rightarrow 1$, $|\langle \bar{c}, \bar{c}^* \rangle| = \delta$.

Согласно равенству Парсеваля имеем:

$$\sum_{i=1}^{\infty} (c_i - c_i^*)^2 = \|F - F^*\|^2 = \varepsilon^2.$$

Тогда и только тогда получаем:

$$\begin{aligned} (c - c^*, c - c^*) = \varepsilon^2 &\Leftrightarrow (c^*, c^*) - (c^*, c) - (c, c^*) + (c, c) = \varepsilon^2 \Leftrightarrow \\ \|c^*\|^2 - 2 \cdot (c^*, c) + \|c\|^2 = \varepsilon^2 &\Leftrightarrow \frac{\|c^*\|^2 - \varepsilon^2 + \|c\|^2}{2} = (c^*, c) \cdot \frac{\|c\| \|c^*\|}{\|c\| \|c^*\|} \Leftrightarrow \\ (\bar{c}^*, \bar{c}) = \frac{\|c^*\|^2 - \varepsilon^2 + \|c\|^2}{2 \|c\| \|c^*\|}. \end{aligned}$$

Обозначим через δ выражение, стоящее в правой части полученного равенства, $\frac{\|c^*\|^2 - \varepsilon^2 + \|c\|^2}{2 \|c\| \|c^*\|} = \delta$. Покажем, что $\delta \leq 1$. Согласно неравенству Коши-Буняковского,

$$|\langle \bar{c}^*, \bar{c} \rangle| \leq \|\bar{c}^*\| \cdot \|\bar{c}\|, \text{ так как } \|\bar{c}^*\| = 1, \|\bar{c}\| = 1, \text{ получаем } |\langle \bar{c}^*, \bar{c} \rangle| \leq 1, \text{ следовательно, и } \delta \leq 1.$$

Покажем, $\delta \rightarrow 1$. В силу утверждения 2 имеем, $\rho(c, c^*) = \varepsilon$ при ε – малое, $\varepsilon \geq 0$.

Таким образом,

$$\left| \|\bar{c}^*\| - \|\bar{c}\| \right| \leq \|c - c^*\| \rightarrow \varepsilon \Rightarrow \|c\| \rightarrow \|c^*\|.$$

Найдем чему равно предельное соотношение

$$\lim_{\|c^*\| \rightarrow \|c\|, \varepsilon \rightarrow 0} \frac{\|c^*\|^2 - \varepsilon^2 + \|c\|^2}{2 \|c\| \|c^*\|} = \frac{1}{2} + \frac{1}{2} = 1.$$

Таким образом, мы показали, что если F является ε – близкой функцией F^* , тогда $\exists \delta \leq I$, $\delta \rightarrow I$, $\left| (\bar{c}, \bar{c}^*) \right| = \delta$.

Что и требовалось доказать.

10. Выдвинуть гипотезу о законе распределения на основании критерия 1 или 2.

11. Проверить гипотезу через критерий согласия ω^2 – Мизеса при заданном уровне значимости α .

Численный пример применения метода. Предположим, что выборочная совокупность эмпирических данных дана в табл.1. Эмпирические данные представляют собой типовое непрерывное распределение. С помощью метода эталонной функции выдвинуть гипотезу о законе распределения. Точность разложения принять равной $\varepsilon=0,05$. Проверить гипотезу о согласованности аппроксимации эмпирическим данным при уровне значимости $\alpha=0,05$ с помощью критерия Мизеса (ω^2). Проверить выдвинутую гипотезу при уровне значимости $\alpha=0,05$ с помощью критерия Мизеса (ω^2).

Таблица 1

Выборка эмпирических данных

x_i	x_i	x_i	x_i	x_i
0,843	-0,327	0,727	-0,023	0,392
0,354	0,277	0,752	0,351	1,59
-0,449	1,02	1,03	1,44	-0,657
0,657	0,645	0,746	0,348	0,37
0,141	0,879	1,21	0,953	0,172
0,643	0,0526	1,38	0,447	-0,314
0,201	0,436	0,485	0,133	1,26
-0,139	-0,46	0,684	0,565	0,385
1,57	-0,209	1,36	0,852	1,26
0,262	1,32	1,01	0,586	0,514

Перейдем от переменных x_i к переменным t_i по формуле:

$$t = \frac{x + 0,657}{1,59 + 0,657} = \frac{x + 0,657}{2,247}.$$

Подсчитаем пять коэффициентов c_k через эмпирические моменты по формуле (4):

$$C^* = (0,460842; 0,565035; 0,040092; -0,10971; -0,00572).$$

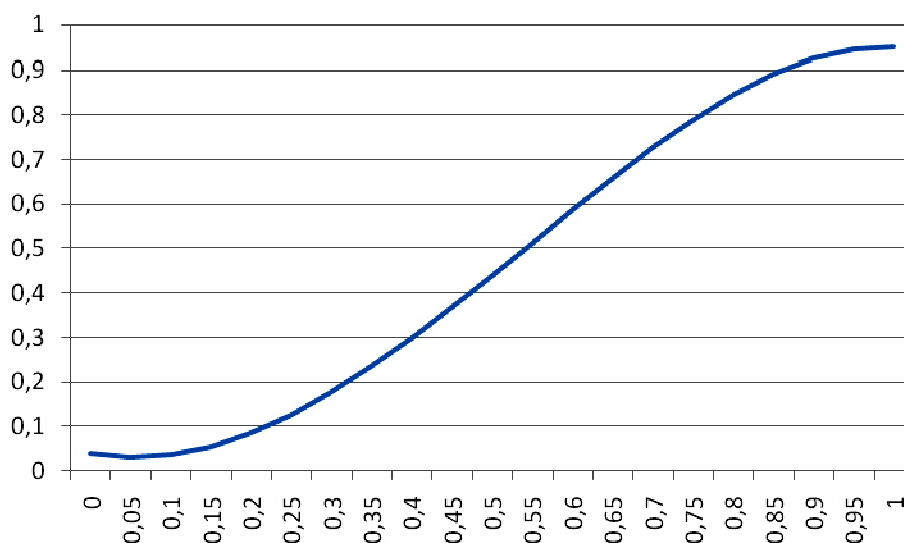
Проверим, является ли достаточным пять первых членов разложения в ряд Фурье для достижения точности $\varepsilon=0,05$.

По формуле (6) получаем $\varepsilon=0,019$, из чего делаем вывод о достаточности представления эмпирической функции распределения через пять первых членов разложения.

Получена аппроксимация эмпирической функции распределения:

$$F^*(t) = 0,460842 + 0,565035(2t-1) + 0,040092\left(\frac{3}{2}(2t-1)^2 - \frac{1}{2}\right) - 0,10971\left(\frac{5}{2}(2t-1)^3 - \frac{3}{2}(2t-1)\right) - 0,00572\left(\frac{35}{8}(2t-1)^4 - \frac{15}{4}(2t-1)^2 + \frac{3}{8}\right) + o(t^5).$$

На рисунке приведен график, построенный на основе аналитического разложения эмпирических данных в ряд Фурье.



Аппроксимация эмпирической функции распределения в ряд Фурье

Проверим гипотезу о согласованности аппроксимации эмпирическим данным при уровне значимости $\alpha=0,05$ с помощью критерия Мизеса (ω^2). Вычисления приведены в табл.2.

Таблица 2

Исходные данные и результаты вычислений

i	x_i	t_i	$F(t_i)$	F_i^*	Δ_i
1	2	3	4	5	6
1	-0,657	0	0,039895	0,01	0,000894
2	-0,46	0,087672	0,034710	0,03	2,22E-05
3	-0,449	0,092568	0,035669	0,05	0,000205
4	-0,327	0,146862	0,054451	0,07	0,000242
5	-0,314	0,152648	0,057296	0,09	0,001069
6	-0,209	0,199377	0,085816	0,11	0,000585
7	-0,139	0,23053	0,109958	0,13	0,000402
8	-0,0232	0,282065	0,157947	0,15	6,32E-05
9	0,0526	0,315799	0,194188	0,17	0,000585
10	0,133	0,35158	0,236244	0,19	0,002139
11	0,141	0,35514	0,240612	0,21	0,000937
12	0,172	0,368936	0,25783	0,23	0,000775
13	0,201	0,381842	0,274332	0,25	0,000592
14	0,262	0,40899	0,310178	0,27	0,001614
15	0,277	0,415665	0,319207	0,29	0,000853
16	0,348	0,447263	0,362924	0,31	0,002801
17	0,351	0,448598	0,364803	0,33	0,001211
18	0,354	0,449933	0,366685	0,35	0,000278
19	0,37	0,457054	0,376759	0,37	4,57E-05
20	0,385	0,463729	0,386259	0,39	1,4E-05
21	0,392	0,466845	0,390709	0,41	0,000372
22	0,436	0,486426	0,418911	0,43	0,000123

Окончание табл.2

1	2	3	4	5	6
23	0,447	0,491322	0,426015	0,45	0,000575
24	0,485	0,508233	0,450687	0,47	0,000373
25	0,514	0,521139	0,469624	0,49	0,000415
26	0,565	0,543836	0,503059	0,51	4,82E-05
27	0,586	0,553182	0,516845	0,53	0,000173
28	0,643	0,578549	0,554206	0,55	1,77E-05
29	0,645	0,579439	0,555514	0,57	0,00021
30	0,657	0,58478	0,563351	0,59	0,00071
31	0,684	0,596796	0,580929	0,61	0,000845
32	0,727	0,615932	0,608716	0,63	0,000453
33	0,746	0,624388	0,620890	0,65	0,000847
34	0,752	0,627058	0,624719	0,67	0,00205
35	0,843	0,667557	0,681676	0,69	6,93E-05
36	0,852	0,671562	0,687174	0,71	0,000521
37	0,879	0,683578	0,703498	0,73	0,000702
38	0,953	0,716511	0,746733	0,75	1,07E-05
39	1,01	0,741878	0,778272	0,77	6,84E-05
40	1,02	0,746328	0,783626	0,79	4,06E-05
41	1,03	0,750779	0,788923	0,81	0,000444
42	1,21	0,830886	0,872920	0,83	0,001842
43	1,26	0,853138	0,891790	0,85	0,001746
44	1,26	0,853138	0,891790	0,87	0,000475
45	1,32	0,87984	0,911414	0,89	0,000459
46	1,36	0,897641	0,922520	0,91	0,000157
47	1,38	0,906542	0,927444	0,93	6,53E-06
48	1,44	0,933244	0,939556	0,95	0,000109
49	1,57	0,991099	0,950793	0,97	0,000369
50	1,59	1	0,950537	0,99	0,001557

В табл.2: $F_i^*=(i-0,5)/50$ – значения эмпирической функции распределения; $F(t_i)$ – значения функции распределения, подсчитанные с помощью аппроксимации в ряд Фурье; $\Delta_i=(F(t_i)-F_i^*)^2$ – расчетные данные наблюдаемого значения статистики, $\sum \Delta_i=0,031118$.

Наблюдаемое значение статистики $n\omega^2$ находим по формуле:

$$n \cdot \omega^2 = \frac{1}{12 \cdot 50} + \sum_{i=1}^{50} \Delta_i = 0,03278.$$

Критическое значение при уровне значимости $\alpha=0,05$ равно $n \cdot \omega^2(0,05)=0,461$.

Наблюдаемое значение $n\omega^2$ меньше критического, следовательно, полученная аппроксимация не противоречит эмпирическим данным. На основе предложенных критериев выдвинем гипотезу о принадлежности эмпирических данных одному из четырех типовых распределений.

Критерий №1. Найдем минимальное расстояние между эмпирической и теоретической функциями. Эталонная матрица для этого примера принимает вид: первая строка – равномерный закон, вторая – экспоненциальный закон, третья – закон Лапласа, четвертая – нормальный закон:

$$\sqrt{\begin{pmatrix} 0,5 & 0,5 & 0 & 0 & 0 \\ 0,4608 & -0,1267 & -3,3788 & -25,79 & -256,45 \\ 0,4608 & 1,3624 & -1,4204 & -1,9439 & 22,1227 \\ 0,4608 & 0,5650 & 0,0267 & -0,1891 & -0,0837 \end{pmatrix} - \begin{pmatrix} 0,460842 \\ 0,565035 \\ 0,040092 \\ -0,10971 \\ -0,00572 \end{pmatrix}^T}^2 = \begin{pmatrix} 0,139427 \\ 257,7609 \\ 22,26656 \\ 0,112084 \end{pmatrix}.$$

В силу критерия № 1 минимальное расстояние между векторами находится в четвертой строке, что дает право выдвинуть гипотезу, что выборка взята из нормальной генеральной совокупности.

Критерий №2. Эталонная матрица для этого примера принимает вид: первая строка – равномерный закон, вторая – экспоненциальный закон, третья – закон Лапласа, четвертая – нормальный закон:

$$\begin{pmatrix} 0,707107 & 0,707107 & 0 & 0 & 0 \\ 0,0018 & -0,0005 & -0,0131 & -0,1001 & -0,9949 \\ 0,0207 & 0,0611 & -0,0637 & -0,0872 & 0,9921 \\ 0,6077 & 0,7451 & 0,0352 & -0,2494 & -0,1104 \end{pmatrix} \begin{pmatrix} 0,6241 \\ 0,7652 \\ 0,0543 \\ -0,1486 \\ -0,0077 \end{pmatrix} = \begin{pmatrix} 0,982317 \\ 0,022604 \\ 0,061486 \\ 0,989181 \end{pmatrix}$$

Полученный вектор дает право выдвинуть гипотезу в силу критерия № 2 о том, что выборка взята из нормальной генеральной совокупности, так как максимально близкое значение к 1 находится в четвертой строке, которая соответствует нормальному закону распределения.

Проверим гипотезу о нормальном распределении с помощью критерия Мизеса при уровне значимости $\alpha=0,05$.

Проведя аналогичные вычисления приведенным в табл.2, получим наблюдаемое значение критерия, равное $n\omega^2=0,0235$, критическое значение при уровне значимости $\alpha=0,05$ равно $n\omega^2=0,461$. Следовательно, гипотеза H_0 не противоречит имеющимся данным.

Заключение. Таким образом, предложенный метод идентификации законов распределения на основе аппроксимативного анализа экспериментальных данных легко автоматизируется. Исходные данные не нуждаются в группировке. Метод позволяет выдвигать гипотезы о законе распределения на основе аппроксимации эмпирических данных, а не на основе визуального анализа гистограммы исследователем.

Метод программно реализован на базе платформы 1С Предприятие 8.1. в рамках статистического пакета «Математическая статистика», а также в программном продукте: «Метод эталонной функции для определения типа распределения выборки» от 20.10.2007, на который автор статьи получено свидетельство об отраслевой регистрации разработки № 9309.

Библиографический список

1. Васильева Т.В. Метод эталонной функции для определения дискретных и непрерывных законов распределения. / Т.В. Васильева // Информационные и математические технологии в науке и управлении. Ч.1. - Иркутск: ИСЭМ СО РАН, 2008. – С.152-158.
2. Васильева Т.В. Метод эталонной функции для нахождения функции распределения. / Т.В. Васильева // Вестник ТГТУ. – 2007. - Ч.4.
3. Айвазян С.А. Прикладная статистика. Основы моделирования и первичная обработка данных. / С.А. Айвазян, И.С. Енюков, Е.Д. Мешалкин. – М.: Финансы и статистика, 1983.
4. Бронштейн И.Н. Справочник по математике / И.Н. Бронштейн, К.А. Семендяев. – М.: Наука, 1980.
5. Прохоров С.А. Аппроксимативный анализ случайных процессов. / С.А. Прохоров – Самара: ГАУ, 2001. - 329 с.

6. Васильева Т.В. Свидетельство об отраслевой регистрации разработки №9309 "Метод эталонной функции для определения типа распределения выборки" от 20.10.2007
7. Ходасевич Г.Б. Обработка экспериментальных данных на ЭВМ. [Электрон. ресурс]. Режим доступа: http://www.dvo.sut.ru/libr/opds/i130hodo_part1/index.htm.
8. Колмогоров А.Н. Элементы теории функций и функционального анализа / А.Н. Колмогоров, С.В. Фомин. – М.: Наука, 1976.
9. Прохоров С.А. Аппроксимативный анализ случайных процессов / С.А. Прохоров. – Самара: ГАУ, 2001. – 329 с.
10. Математическая энциклопедия. // Советская энциклопедия. – 1984. – Т.1-5.

Материал поступил в редакцию 06.10.09.

T.V. VASILYEVA

IDENTIFICATION AUTOMATION OF THE DISTRIBUTION LAWS ON THE BASIS OF APPROXIMATE ANALYSIS OF THE EXPERIMENTAL DATA

An algorithm of identification automation of the distribution laws on the basis of approximate analysis experimental data is given in the article. The method is founded on the decomposition statistics in Fourier series in space L_2 . There are two necessary and sufficient criteria for automatic setting up of the hypothesis about the distribution law. The method is considered on a detailed example.

Key words: automation of the statistics analysis, approximation, distribution law, hypotheses proving, fitting criteria

ВАСИЛЬЕВА Татьяна Валерьевна, старший преподаватель кафедры «Автоматизированные системы» Иркутского государственного технического университета. Окончила Иркутский государственный университет (1996) по специальности «Прикладная математика», аспирантуру (2008). Область научных интересов: теория вероятностей, теория информации, математическая статистика, программирование.

Автор 7 научных статей и 6 свидетельств о регистрации программ.

vtv74@inbox.ru