УДК 519.2 (075.8)

Б.Ю. Лемешко, С.Н. Постовалов

СТАТИСТИЧЕСКИЙ АНАЛИЗ СМЕСЕЙ РАСПРЕДЕЛЕНИЙ ПО ЧАСТИЧНО ГРУППИРОВАННЫМ ДАННЫМ

В статье рассмотрены вопросы формирования вероятностных моделей в виде смесей одномерных непрерывных распределений, оценивания параметров смеси законов и моделирования. Приведены результаты программной реализации.

При статистическом анализе одномерных наблюдений случайных величин возникает необходимость в решении совокупности следующих задач: выбор или формирование вероятностной модели; оценивание параметров модели; проверка согласия с выборкой. В технических приложениях достаточно важной оказывается и задача моделирования (имитации) случайных величин по заданной модели, тесно связанная с перечисленными выше.

Вероятностная модель, используемая для описания наблюдений, может быть известна a priori исходя из теоретических предпосылок, характеризующих случайную величину. Однако чаще всего такая модель оказывается неизвестной. В этом случае перебирают различные модели и выбирают ту, которая лучше всего согласуется с выборочными данными.

Как правило, на практике ограничиваются рассмотрением нескольких простых распределений, таких как нормальное, экспоненциальное, гамма , Вейбулла-Гнеденко, Коши и др. Существенно увеличить количество моделей, используемых для описания реальных случайных величин, можно за счет применения операций над распределениями. Некоторые операции приведены в таблице.

Параметры сдвига и масштаба явно включены в большинство распределений. Усечение целесообразно применять и применяют тогда, когда известны физические ограничения на область определения случайной величины. Смесь законов распределений может образоваться, например, при объединении выборок, распределенных по разным законам, или когда наблюдаемая величина может являться следствием различных причин.

Операции над распределениями

Операция	Число дополнит. параметров	Функция распределения	Функция плотности распределения
Сдвиг	1	$F(x-\mu)$	$f(x-\mu)$
Масштаб	1	$F(x/\sigma)$	$\dfrac{f(x/\sigma)}{\sigma}$
Усечение слева	1	$\begin{cases} 0, & x<a \\ \dfrac{F(x)-F(a)}{1-F(a)}, & x\ge a \end{cases}$	$\begin{cases} 0, & x<a \\ \dfrac{f(x)}{1-F(a)}, & x\ge a \end{cases}$
Усечение справа	1	$\begin{cases} \dfrac{F(x)}{F(b)}, & x\le b \\ 0, & x>b \end{cases}$	$\begin{cases} \dfrac{f(x)}{F(b)}, & x\le b \\ 0, & x>b \end{cases}$
Двустороннее усечение	2	$\begin{cases} 0, & x<a \\ \dfrac{f(x)}{F(b)-F(a)}, & x\in[a,b] \\ 1, & x>b \end{cases}$	$\begin{cases} 0, & x<a \\ \dfrac{F(x)-F(a)}{F(b)-F(a)}, & x\in[a,b] \\ 1, & x>b \end{cases}$
Смесь	1	$wF_1(x,\theta_1)+(1-w)F_2(x,\theta_2)$	$wf_1(x,\theta_1)+(1-w)f_2(x,\theta_2)$

Использование усечённых законов распределения и смесей законов существенно расширяет множество вероятностных моделей, применяемых для описания реальных данных, увеличивая соответственно сложности, в том числе вычислительного характера. Например, для усечённых законов распределения решение задачи асимптотически оптимального группирования [1,2], имеющее принципиально важное значение для качества статистических выводов, оказывается зависящим от параметров усечения. Это означает, что таблицы асимптотически оптимального группирования для использования в критериях согласия могут формироваться либо для конкретных значений параметров усечения, либо, что более предпочтительно, соответствующая задача должна решаться непосредственно перед проверкой гипотезы о согласии. Ряд принципиальных моментов возникает и при

анализе смесей распределений.

Рассмотрим, что представляет собой область определения параметра смеси. Параметр w выбирают таким образом, чтобы функция плотности была неотрицательной:

$$f(x,w) = wf_1(x) + (1-w)f_2(x) \geq 0,\ \forall w \in \Omega,\ \forall x \in X. \qquad (1)$$

Условие (1) можно преобразовать к виду:

$$w\left[f_2(x) - f_1(x)\right] \leq f_2(x).$$

Пусть $X = A \bigcup B \bigcup C$, где

$$A = \left\{x \in X\colon f_2(x) - f_1(x) < 0\right\};$$
$$B = \left\{x \in X\colon f_2(x) - f_1(x) > 0\right\};$$
$$C = \left\{x \in X\colon f_2(x) - f_1(x) = 0\right\}.$$

Тогда $\Omega = \Omega_A \bigcap \Omega_B \bigcap \Omega_C$, где $\Omega_C = R$, а

$$\Omega_A = \left\{w \in R\colon w \geq \frac{f_2(x)}{f_2(x) - f_1(x)}, \forall x \in A\right\};$$

$$\Omega_B = \left\{w \in R\colon w \leq \frac{f_2(x)}{f_2(x) - f_1(x)}, \forall x \in B\right\}.$$

Обозначим

$$a = \max_{x \in A} \frac{f_2(x)}{f_2(x) - f_1(x)} \leq 0;$$

$$b = \min_{x \in B} \frac{f_2(x)}{f_2(x) - f_1(x)} \geq 1.$$

В результате получим, что область определения параметра смеси имеет вид: $\Omega = \Omega_A \bigcap \Omega_B \bigcap \Omega_C = [a,b] \supseteq [0,1]$.

Когда параметр смеси принадлежит интервалу $[0,1]$, мы имеем классическую смесь, получаемую, например, объединением выборок. Если же $w \notin [0,1]$, то одно из распределений входит в смесь со знаком минус и, таким образом, вычитается из другого распределения.

Обобщением смеси двух распределений является смесь из s распределений. Функция распределения в этом случае имеет вид

$$F(x) = \sum_{i=1}^{s} w_i F_i(x, \theta_i),$$

где s - число распределений в смеси; w_i - параметры смеси; F_i - i-я функция распределения; θ_i - вектор её параметров. Параметры смеси удовлетворяют условию нормировки:

$$\sum_{i=1}^{s} w_i = 1.$$

Характер выборочных данных, по которым осуществляется анализ смесей распределений, может быть различным. Наиболее общим случаем является частично группированная выборка [1]. Выборка является *негруппированной*, если выборочные значения представляют собой индивидуальные значения наблюдений из области определения случайной величины. Выборка является *группированной*, если область определения случайной величины разбита на k непересекающихся интервалов граничными точками

$$x_0 < x_1 < \ldots < x_{k-1} < x_k,$$

где x_0 - нижняя грань области определения случайной величины X; x_k - верхняя грань области определения случайной величины X, и зафиксированы количества наблюдений n_i, попавших в i -й интервал значений. Выборка является *частично группированной*, если имеющаяся в нашем распоряжении информация связана с множеством непересекающихся интервалов, которые делят область определения случайной величины так, что каждый интервал принадлежит к одному из двух типов:

а) i -й интервал принадлежит к первому типу, если число n_i известно, но индивидуальные значения x_{ij}, $j = \overline{1, n_i}$, неизвестны;

б) i -й интервал принадлежит ко второму типу, если известно не только число n_i, но и все индивидуальные значения x_{ij}, $j = \overline{1, n_i}$.

Оценки параметров смеси распределений по частично группированной выборке находятся по методу максимального правдоподобия. Функция правдоподобия для частично группированной выборки имеет вид:

$$L(\theta) = \prod_{(1)} P_i^{n_i}(\theta) \prod_{(2)} \prod_{j=1}^{n_i} f(x_{ij}, \theta),$$

где $f(x, \theta)$ - функция плотности случайной величины; $P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x, \theta) dx$ - вероятность попадания наблюдения в i -й интервал значений; (1) и (2) означают, что умножение осуществляется по интервалам с группированными и негруппированными данными соответственно.

Смесь является *идентифицируемой*, если для $\forall w_1, w_2 \in \Omega$: $w_1 \neq w_2$ следует, что $\exists x \in X$: $f(x, w_1) \neq f(x, w_2)$. Очевидно, что смесь (1) неидентифицируема, если $f_1 \equiv f_2$. Численная реализация метода максимального правдоподобия показала, что функция $L(\theta)$ для смеси распределений зачастую является многоэкстремальной и поэтому получаемые оценки параметров зависят от начального приближения. Качество оценки параметра смеси заметно ухудшается, когда входящие в смесь распределения достаточно близки по форме.

Смоделировать выборку размера n по закону смеси распределений можно обычным методом обратных функций. Если же $0 \leq w \leq 1$, то можно сгенерировать две выборки: размера $n_1 = wn$ по закону $F_1(x)$ и размера $n_2 = (1-w)n$ по закону $F_2(x)$, а затем объединить их. Нетрудно показать, что эмпирическая функция распределения $F_n(x)$ полученной выборки будет смесью эмпирических функций распределения $F_{n_1}(x)$ и $F_{n_2}(x)$ первой и второй выборок соответственно с параметром смеси $w = n_1/(n_1 + n_2)$.

Программное обеспечение, реализующее решение задачи выбора закона распределения или *смеси* законов, наиболее хорошо описывающих выборочные данные, является дальнейшим развитием программной системы "Статистический анализ одномерных наблюдений случайных величин" [2]. Программное обеспечение позволяет находить оценки максимального правдоподобия параметров *для смесей* любых из 26 наиболее часто используемых в приложениях распределений.

При анализе осуществляется проверка гипотез о согласии по критериям: χ^2- Пирсона, отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса. Использование совокупности критериев дает возможность принимать более обоснованное решение, а при противоречивости выводов по отдельным критериям - формировать компромиссный критерий и делать окончательный вывод с учетом его.

В качестве примера на рис.1 приведены результаты статистического анализа негруппированной выборки (оценки параметров и результаты проверки гипотез о согласии по различным критериям), полученной объединением двух выборок: первая распределена по экспоненциальному закону, вторая - по нормальному; каждая объемом по 100 наблюдений (t_0 - параметр сдвига, t_1 - параметр масштаба).

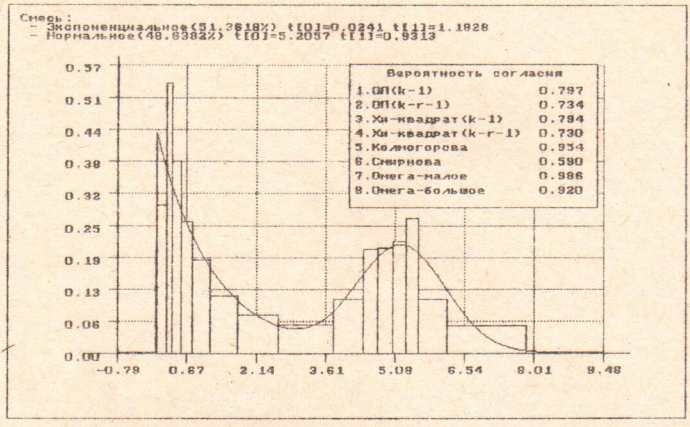

Рис. 1. Оценивание параметров смеси экспоненциального и нормального распределений

На рис. 2 приведены результаты статистического анализа выборки, полученной объединением двух выборок, распределённых по нормальному закону: первая объемом 100 наблюдений генерировалась с параметрами (0, 1/2), вторая объемом 900 наблюдений - с параметрами (0, 5).

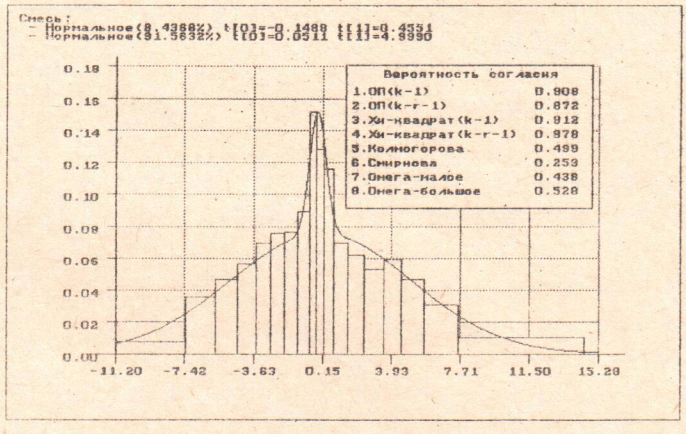

Рис. 2. Оценивание параметров смеси двух нормальных распределений

В приведенных примерах мы явно смешивали выборки, распределённые в соответствии с известными законами. Очевидно, что любые попытки

описать сформированную смесь с использованием какого-либо одного закона в данной ситуации заведомо обречены на неудачу, так как смешивались пары существенно отличающиеся видом или параметрами законов.

В реальных условиях, когда исследуемая выборка действительно является смесью наблюдений двух различных, но с близкими законами, величин, либо, например, наблюдаемая величина по истечении времени несколько меняет свои параметры, а выборка сформирована из наблюдений до и после изменения, при подборе закона распределения иногда удаётся остановиться на каком-то одном законе. Однако уровень значимости α (вероятность ошибки первого рода), при котором гипотеза о согласии не отвергается, оказывается чрезвычайно малым, за границей разумного. В то же время использование в качестве искомого закона смеси распределений даёт хорошие результаты. Более того, иногда это позволяет подсказать, что наблюдается не одна, а несколько случайных величин. На рис.3 приведены результаты анализа геодезических измерений, представляющие собой смесь трех законов.

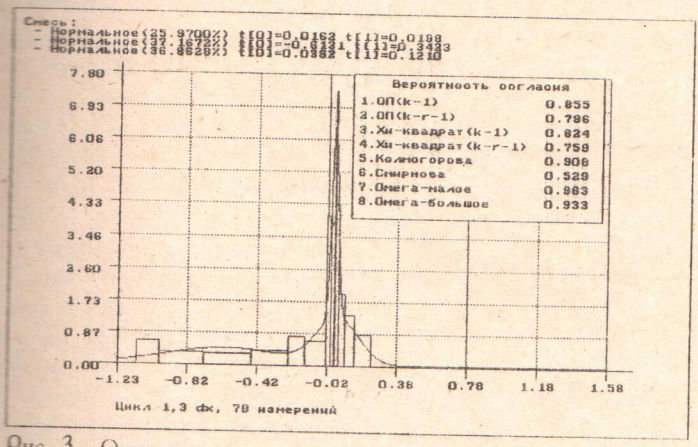

Рис. 3. Оценивание параметров смеси трех нормальных распределений

1. Куллдорф Г. Введение в теорию оценивания по группированным и частично группированным выборкам. - М.: Наука, 1966. - 176 с.
2. Денисов В.И., Лемешко Б.Ю., Цой Е.Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов. - Новосибирск, 1993. - 347 с.