# Программное обеспечение статистического анализа смесей случайных величин, представленных частично группированными и интервальными выборками

Б.Ю. Лемешко, С.Н. Постовалов

Кафедра прикладной математики, НГТУ, Новосибирск, Россия

**Аннотация**

Рассматривается система статистического анализа, обеспечивающая обработку как частично группированных данных, так и интервальных наблюдений с использованием в качестве моделей реальных законов распределения смесей усеченных и неусеченных законов распределений. Рассматриваются достоинства объектно-ориентированного подхода к реализации задач статистики. Приведена иерархия классов статистического анализа и пример использования.

Разработанная в Новосибирском государственном техническом университете программная система предназначенная для статистического анализа одномерных непрерывных случайных величин [1]. Система ориентирована на представление исходных наблюдений в виде частично группированных данных. Это может быть негруппированная, группированная, или цензурированная выборка. Системой охвачено 26 законов и семейств непрерывных законов распределения. Система позволяет получать оценки параметров непрерывных законов распределения, проверять гипотезы о согласии по критериям $\chi^2$-Пирсона, отношения правдоподобия, Колмогорова, Смирнова, $\omega^2$ и $\Omega^2$ Мизеса, идентифицировать закон распределения, моделировать случайные величины, подчиняющиеся различным законам распределения. Оригинальные алгоритмы позволяют получать робастные оценки, гарантируют максимальную мощность критериев согласия отношения правдоподобия и $\chi^2$-Пирсона, отбраковывать аномальные наблюдения.

Многообразие законов распределения обычно используемых для описания реально наблюдаемых величин к глубокому сожалению довольно ограничено. Без достаточного основания наиболее часто используют нормальное распределение. Однако в большинстве случаев такой выбор не является обоснованным [2]. Очень часто даже того набора законов распределения, который охвачен программной системой [1], оказывается недостаточным для более или менее приемлемого описания регистрируемых измерений. Это связано с тем, что, во-первых, область определения реальных случайных величин бывает ограничена физическими условиями. И, следовательно, на самом деле наблюдаем усеченные случайные величины. Во-вторых, достаточно часто наблюдаемые законы распределения имеют двухмодальный и многомодальный характер. Наиболее типична такая ситуация для распределений погрешностей измерительных приборов [3]. В этом случае определенный выход при описании таких законов может быть найден в использовании смесей законов распределения. Усеченные законы и смеси распределений значительно расширяет множество моделей, используемых для описания наблюдений, регистрируемых во многих приложениях.

В последнее время в связи с распространением идей нечеткой и интервальной математики, стали предлагаться другие способы учета ошибок измерений. Суть этого подхода состоит в том, что ошибки наблюдений рассматриваются либо как неслучайные величины, либо как случайные величины, но с заведомо

неизвестным распределением. В этом случае выборочное наблюдение задаётся не одним числом, а интервалом, определяющим его верхнюю и нижнюю границу. Например, если известна погрешность $\Delta$ прибора, с помощью которого делаются измерения, то интервальное представление измеренного значения $x_i$ будет иметь вид $[x_i - \Delta, x_i + \Delta]$ [4].

В этой связи вновь разрабатываемая программная система, в которой сохраняются все достоинства предыдущей версии [1] (робастность оценок, асимптотически оптимальное группирование данных, возможность исключения аномальных ошибок наблюдений и т.д), обеспечивает обработку не только частично группированных данных, но и интервальных наблюдений, а также использование для идентификации наблюдаемого закона моделей усеченных законов и смесей законов распределения.

Применение смесей и усеченных законов существенно увеличивает сложности вычислительного характера вследствие роста размерности задач и ряда новых особенностей задач оценивания. Одна из них связана с тем, что параметр смеси не обязательно принадлежит интервалу $[0,1]$. В общем случае параметр смеси $w$ двух распределений с функцией плотности $f(x) = wf_1(x) + (1-w)f_2(x)$, принадлежит интервалу $[a,b]$, где

$$a = \max_{x \in A} \frac{f_2(x)}{f_2(x) - f_1(x)} \le 0, \quad b = \min_{x \in B} \frac{f_2(x)}{f_2(x) - f_1(x)} \ge 1,$$

$$A = \{x | f_2(x) < f_1(x)\}, \quad B = \{x | f_2(x) > f_1(x)\},$$

где $a = 0$, если $\exists x$: $f_2(x) = 0 \wedge f_1(x) \neq 0$, и $b = 1$, если $\exists x$: $f_1(x) = 0 \wedge f_2(x) \neq 0$ [5]. В системе предусмотрена возможность использования смесей распределений как с параметром $w \in [0,1]$, так и $w \notin [0,1]$.

Усеченные распределения имеют плотность вида

$$f(x) = \begin{cases} f_0(x)/(F_0(\alpha) - F_0(\beta)), & x \in [\alpha, \beta], \\ 0, & x \notin [\alpha, \beta], \end{cases}$$

где $\alpha$ и $\beta$ - параметры усечения, $f_0(x)$ - исходная функция плотности. Параметры усечения могут задаваться пользователем или оцениваться по порядковым статистикам, в том числе с применением оптимальных оценок, использующих асимптотически оптимальное группирование.

На рис. 1 в качестве примера приведены результаты статистического анализа смеси двух усеченных нормальных распределений.

Система реализуется с использованием возможностей объектно-ориентированного программирования (ООП). Во-первых, это увеличивает надежность программного комплекса. Все данные локализованы в классах, доступ к данным и проверка их корректности выполняется соответствующими процедурами. Во-вторых, улучшается структура программ, она становится более наглядной и читабельной. Достаточно простым становится введение дополнительных параметров, образование смешанных и усеченных распределений. В-третьих, ООП упрощает расширение и добавление новых функций в программу и, таким образом, позволяет задействовать большие коллективы разработчиков программного обеспечения. Иерархия классов системы показана на рис.2. Кратко рассмотрим возможности, заложенные в эти классы.

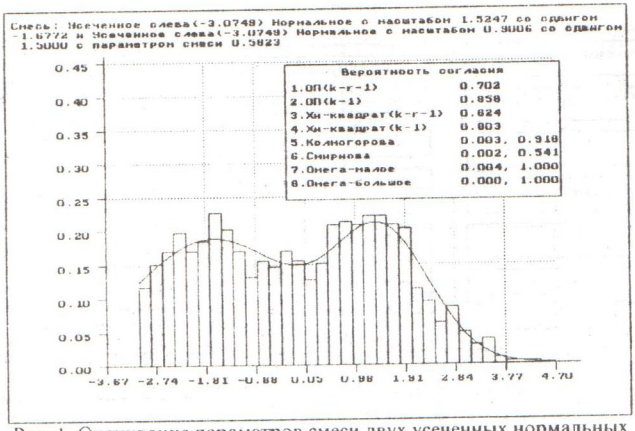


Рис. 1. Оценивание параметров смеси двух усеченных нормальных распределений по группированной выборке

**Класс "Наблюдение" ("Observation")**. Описывает единичное наблюдение. Рассматривается самый общий случай, когда наблюдение задано интервалом. *Данные*: порядковый номер, левая граница, правая граница, количество наблюдений, принадлежащих этому интервалу, тип интервала (вырожденный, бесконечный слева, бесконечный справа).

**Класс "Выборка" ("Sample")**. Описывает одномерную выборку, содержащую множество наблюдений. *Данные*: количество наблюдений в выборке, массив наблюдений, название выборки, входной файл, файл результатов, тип выборки (негруппированная, группированная, частично группированная, интервальная). *Методы*: чтение выборки из файла, запись выборки в файл, сортировка выборки, вычисление числовых характеристик выборки, получение наблюдения по номеру или по порядку возрастания левой или правой границ интервалов.

**Класс "Распределение" ("Distribution").** Описывает абстрактное распределение. *Данные*: количество параметров у распределения. *Методы*: вычисление обратной функции распределения, генерирование псевдослучайного числа, процедура равновероятного группирования, процедура асимптотически оптимального группирования, вычисление максимума функции плотности. *Виртуальные методы*: получить значение параметра, задать значение параметра, получить индикатор параметра, задать индикатор параметра, проверить область определения параметра. *Чисто виртуальные методы*: вычисление функции распределения, плотности распределения, их производных по $x$ и по параметрам до второго порядка включительно. Этот класс является абстрактным, так как в нем определены чисто виртуальные функции. За счет того, что не определен явно вид функции распределения, имеется возможность работать с произвольным распределением. В качестве производных от класса "Распределение" реализованы классы - операции над распределениями: "Сдвиг", "Масштаб", "Смесь", "Усечение слева", "Усечение справа", "Двустороннее усечение", а также классы - конкретные распределения, такие как "Нормальное", "Экспоненциальное", "Парето", "Вейбулла", "Коши", "Лапласа" и др.

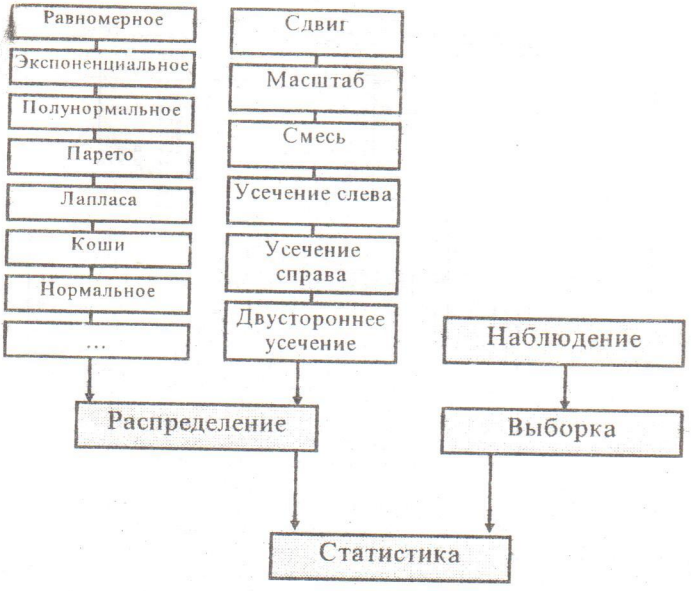


Рис. 2. Структура классов статистического анализа

**Класс "Статистика" ("Stat").** Связывает конкретную выборку и распределение. *Данные*: указатель на распределение, указатель на выборку, массив граничных точек для группирования. *Методы*: оценивание параметров распределения по методу максимального правдоподобия, проверка гипотез о согласии по критериям $\chi^2$, Колмогорова, Смирнова, $\omega^2$ и $\Omega^2$ Мизеса, графическое изображение функции распределения и гистограммы.

За счет использования возможностей ООП, можно строить сколь угодно сложные комбинации распределений на основе стандарных.

**Литература**


1. Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. - Новосибирск: Изд-во НГТУ, 1995. - 125 с.
2. Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. 1991. №7.
3. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. - Л.: Энергоатомиздат, 1991. - 303 с.
4. Орлов А.И. О влиянии погрешностей наблюдений на свойства статистических процедур (на примере гамма-распределения) // Статистические методы оценивания и проверки гипотез: Межвуз. сб. науч. тр. / Перм. ун-т. Пермь, 1988.
5. Лемешко Б.Ю., Постовалов С.Н. Статистический анализ смесей распределений по частично группированным данным. // Сб. научных трудов НГТУ. - 1995. - №1. С. 25-31.