ОБ ИНСТРУМЕНТАХ СИСТЕМЫ *MATHEMATICA* ДЛЯ РЕАЛИЗАЦИИ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИХ МЕТОДОВ

В. Б. Таранчук
Кафедра компьютерных технологий и систем, Факультет прикладной математики и информатики,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: taranchuk@bsu.by

Изложение содержит обзор основных инструментов статистической обработки данных системы компьютерной алгебры Wolfram Mathematica.

Введение

Одним из наиболее интенсивно развивающихся секторов рынка приложений информационных технологий (ИТ) является разработка компьютерных программ для анализа статистических данных с целью выявления закономерностей, сравнения вероятных альтернатив выбора, построения прогнозов развития событий, обнаружения связей между явлениями, процессами, создания компьютерных моделей. Распространяемые программы, пакеты, системы постоянно совершенствуются в направлениях ускорения обработки данных, увеличения числа встроенных статистических процедур и функций, улучшения графической визуализации результатов анализа, повышения удобства интерфейса, совершенствования справочной системы. Потребность в средствах статистического анализа данных в различных областях деятельности, особенно в науке и образовании, очень велика. По официальным данным Международного статистического института число различных наименований распространяемых на рынке статистических программных продуктов приближается к тысяче. Поэтому для разработчиков и пользователей программных средств важно сориентироваться в разнообразии, иметь рекомендации, как выбрать подходящий продукт. Методология сравнительного анализа статистических программных продуктов дана в работах С.А. Айвазяна и В.С. Степанова (см. например [1]); предлагаемые авторами классификация и оценки программных продуктов по-прежнему актуальны. С современным состоянием, достоинствами основных статистических пакетов, их функциональными возможностями можно ознакомиться в [2].

Целью настоящей работы является привлечение внимания специалистов к кардинальному изменению ситуации на рынке предложений в обсуждаемом секторе ИТ, которое произошло в 2010 г.. Тогда была выпущена система компьютерной алгебры Wolfram *Mathematica* 8 ([3, 4]) с существенно расширенными функциональными возможностями реализации вероятностно-статистических методов; а также предложен Wolfram группой формат вычисляемых документов CDF ([5]). В этом формате можно распространять приложения, работающие локально на компьютерах любого класса во всех операционных системах или в Интернет под любым браузером (нужна инсталляция бесплатного CDF проигрывателя). CDF документы можно создавать с инструментами интерактивности (меню, кнопок, указателей, бегунков, динамических локаторов), с возможностями представления результатов в математической нотации, визуализации шагов вычислений и иллюстрирования графиками всех типов (1D, 2D, 3D, анимация), импорта и экспорта результатов во все общепринятые форматы данных и графики ([6]). Отдельно заметим, что изучение методов теории вероятностей и математической статистики предполагает наличие теоретических знаний и практических навыков математического анализа, линейной алгебры, дифференциальных уравнений – именно в системах компьютерной алгебры (СКА) решение задач названных дисциплин реализуется в математической нотации, СКА дают максимальные возможности аналитического исследования.

Основные понятия теории вероятностей, соответствующие математические выражения и преобразования, функции манипулирования и графической визуализации данных реализованы практически во всех системах компьютерной алгебры ([4]). Ниже отмечены функции и реализованные в *Mathematica* алгоритмы вероятностно-статистических методов, которые целесообразно использовать при создании профессиональных и специальных статистических программных приложений, изучении методов.

I. Характеристики и решатели для теории вероятностей и статистики.

Для поддержки статистического моделирования и анализа с помощью распределений система *Mathematica*, начиная с версии 8, предлагает коллекцию встроенных распределений вероятностей, для каждого из которых поддерживается несколько десятков свойств, таких как функции распределения, моменты, квантили и генерирующие функции. В частности, пользователям доступны ([3]): символьные, приближен-

ные вычисления вероятностей и условных вероятностей событий, заданных в форме логических комбинаций равенств и неравенств; символьные и численные вычисления ожиданий и условных ожиданий выражений; генерирование выборки из распределений, оценивание параметров распределения и проверка согласованности данных и распределений; поддержка функций распределений, среди которых плотность вероятностей, кумулятивная функция распределения, функция надёжности, плотность отказов, обратная кумулятивная функция и обратная функция надёжности; вычисление разных типов моментов; поддержка всех производящих функций, связанных с моментами; переход между моментами разных типов; вычисление стандартных и несмещённых точечных оценок моментов.

II. Параметрические, непараметрические, вторичные и формульные распределения вероятностей

Основываясь на наибольшей в мире коллекции специальных функций и двух десятилетиях развития символьных и численных алгоритмов, система *Mathematica*, начиная с версии 8, предлагает беспрецедентный уровень поддержки параметрических распределений [7]. Основные составляющие параметрического моделирования и анализа в СКА: коллекция параметрических распределений вероятностей; обширная поддержка распределений с длинным хвостом; поддержка распределений вероятностей экстремальных значений; несколько десятков рассчитываемых характеристик для распределений вероятностей; автоматизированное оценивание параметров и проверка статистических гипотез; обширная документация, содержащая свойства распределений и их соотношения с другими распределениями вероятностей.

В системе *Mathematica* 8 был введен принципиально новый подход к моделированию распределений вероятностей [7]. Первое – это понятие непараметрического распределения, автоматизирующее и обобщающее целый ряд непараметрических методов, используемых для вычисления определённых свойств распределения вероятностей. Второе – понятие вторичного распределения вероятностей, создаваемого из произвольного распределения вероятностей с помощью функциональных преобразований, операций усечения, смешивания и т.п. Третье – это понятие распределения вероятностей, заданного формулой плотности вероятности, кумулятивной функции распределения или функции надёжности. Различные типы распределений вероятностей слаженно работают, образуя оболочку для моделирования и анализа. Основные составляющие, инструменты ([3]): одномерные и многомерные преобразования случайных переменных; одномерные и совместные распределения вероятностей порядковых статистик произвольного распределения; компонентное смешивание произвольных совместимых распределений вероятностей; параметрическое смешивание распределений вероятностей с использованием дискретных или непрерывных весовых распределений вероятностей; непараметрические распределения, включающие эмпирическое, гистограммное, по ядерному сглаживанию и другие; оценивание методом ядерного сглаживания с автоматическим выбором фиксированной или адаптивной ширины окна; оптимизированные одномерные и многомерные эмпирические распределения вероятностей; усечение, цензурирование распределений вероятностей в произвольном измерении, для непрерывных или дискретных случайных величин; вторичные распределения вероятностей, полученные в результате функциональных преобразований, операций усечения, смешивания; непараметрическое оценивание цензурированных данных методом максимального правдоподобия; эффективное моделирование выживаемости и надёжности с помощью усечённых и цензурированных распределений; копулы с разными семействами ядер и произвольными маргинальными распределениями вероятностей; маргинальные распределения вероятностей с произвольной размерностью произвольного распределения вероятностей в пространстве большей размерности; распределения вероятностей, заданные формулами для распределения плотности вероятности (PDF), функции накопленного распределения, надёжности.

III. Оценивание параметров распределений и проверка статистических гипотез

Используя в системе тесную интеграцию символьных и численных возможностей, *Mathematica*, начиная с версии 8, предлагает автоматизированные и производительные оценивания параметров распределений, проверку статистических гипотез для более 100 встроенных параметрических распределений и конструкторов вторичных распределений вероятностей. Автоматизация выбора процедур решения уравнений и оптимизации позволяет пользователю уделять больше времени вопросу, на который нужно ответить, а не алгоритмическим подробностям. При этом имеется возможность установки уточняющих опций, которые позволяют экспертам контролировать процесс оптимизации и подробности конкретных методов проверки статистических гипотез. Для оценивания параметров или проверки гипотезы необходимо несколько коротких команд, что позволяет быстро переходить в их работе к этапу принятия решений и построения отчётов. Основные методики ([3, 7]): оценивание параметров параметрических и вторичных распределений вероятностей;

оценка параметров распределений вероятностей методом максимального правдоподобия или методом моментов; критерии согласия для одномерных, многомерных, дискретных и непрерывных данных; непосредственный доступ к критериям согласия Колмогорова-Смирнова, Пирсона, Андресона-Дарлинга и др.; автоматизированные тесты соотношений коэффициентов сдвига и масштаба для произвольного количества наборов данных; полный комплект именных тестов соотношений коэффициентов сдвига и масштаба; полный набор результатов тестирования, включая значения тестовых статистик.

IV. Алгоритмы теории групп

В версии 8 добавлены новые функции и алгоритмы для работы с перестановками и группами перестановок, пользователям *Mathematica* предоставляется систематизированный доступ к множеству групп, порожденных произведениями множества перестановок. В частности: поддержка записи перестановки в виде произведения непересекающихся циклов; поддержка групп, порожденных множествами перестановок. Предварительно сохраненные генераторы бесконечных семейств групп и спорадических групп; визуализация групп в виде таблиц умножения или графов Кэли; вычисления орбит, стабилизаторов, централизаторов, представителей классов смежности и др.

V. Случайные процессы

В *Mathematica* 9, 10 добавлены функции, которые моделируют системы, изменяющиеся во времени случайным, а не детерминированным образом, но в которых состояния, в близкие моменты времени являются зависимыми ([8]). Доступны: построение скалярных и векторных процессов скользящего среднего, авторегрессивных, комбинированных процессов; работа с регулярными или нерегулярными, скалярными или векторными данными временного ряда; оценивание параметров процесса по данным временного ряда; реализация моделей временных рядов с полиномиальными или сезонными трендами, моделей временных рядов с долгой памятью; работа с одним или несколькими временными срезами процессов, как с распределениями вероятностей; нахождение статистик временных срезов таких, как функции среднего значения, медианы, вариации и пр.; непосредственное вычисление ковариационной, корреляционной и абсолютной корреляционной функций; большое число параметрических случайных процессов, конечных цепей Маркова с дискретным и с непрерывным временем; процессов массового обслуживания, включая очереди и сети очередей общего положения; решение параметрических стохастических дифференциальных уравнений (СДУ) таких, как диффузионный процесс Кокса-Ингерсолла-Росса; автоматическое преобразование параметрических диффузионных процессов в соответствующие процессы Ито или Стратоновича; разные методы построения случайных реализаций решений СДУ, в том числе метод Эйлера-Маруямы, стохастические методы Рунге-Кутта и др.

VI. Прикладные области

Система *Mathematica* содержит инструментарий для решения задач в классическом и современном финансовом деле (встроенные финансовые вычисления), а также предоставляет доступ к большим массивам финансовых и экономических данных, содержит богатую функциональность финансового импорта и экспорта для работы с внешними данными. В том числе можно использовать ([3, 8]): полный спектр вычислений по финансовым деривативам; поддержка опционов европейских и американских стилей погашения, а также экзотических финансовых деривативов; вычисление греческих коэффициентов и подразумеваемой волатильности; символьные и численные вычисления временной стоимости денег; функции оценивания разовых сумм, аннуитетов, непрерывных и дискретных денежных потоков; вычисление эффективной процентной ставки для непрерывных и дискретных временных процессов и кривых доходности; набор инструментов для расчётов по облигациям и их оценивание; вычисления мер восприимчивости облигации, начисленного процента и календарных измерений; поддержка непрерывных и дискретных во времени процессов для купонных и процентных ставок; гладкая интеграция финансовых функций с подсистемами визуализации и статистики системы *Mathematica*.

VII. Возможности использования в собственных разработках

В информационных ресурсах Wolfram, системе *Mathematica* в Центр документации (Documentation Center), Навигатор по функциям (Function Navigator), Виртуальный учебник (Virtual Book) по всем методам доступны упражнения-иллюстрации [9]. Например, по блоку характеристики и решатели для теории вероятностей и статистики можно использовать упражнения с исходными кодами: нахождение вероятности событий и среднего значения выражений для параметрических, непараметрических, вторичных и формульных распределений вероятностей; определение вероятностей и средних значений при заданных условиях для произвольных распределений; вычисление двусторонней вероятности для стандартного нормального распределения; вычисление и визуализация функций распределения для одномерных и многомерных непрерывных или дискретных случайных величин; использование центральных

моментов по данным для построения разложения Эджворта, аппроксимирующего распределение генеральной совокупности; пояснения теоремы Гливенко-Кантелли, демонстрация, что выборочная функция распределения при увеличении объёма выборки стремится к её теоретическому аналогу (путём сравнения графиков выборочной и теоретической функций распределения); расчёт вероятности события, соответствующего участку комплексной плоскости; меры вариации и сдвига для параметрических распределений; использование разных масштабов для рассмотрения характерных деталей и общей формы плотностей распределений двух различных распределений вероятностей, обладающих идентичными последовательностями моментов, и сравнение первых членов последовательности моментов в форме таблицы; нахождение многомерных полуинвариантов по моментам: значение кумулянты, Cumulant, трехмерной случайной величины, выраженное в терминах её моментов, Moment, используя команду MomentConvert; нахождение несмещённых точечных оценок для произведения центральных моментов двумерной случайной величины в терминах формальных симметричных многочленов сумм степеней; переход между произвольными типами формальных моментов (Moment, CentralMoment, FactorialMoment, Cumulant); нахождение математического ожидания многочленов стандартных точечных оценок: команда MomentConvert позволяет найти математическое ожидание формальных стандартных точечных оценок относительно выборочного распределения в терминах формальных моментов; сравнение плотности вероятности согласованного распределения с гистограммой данных, и сводная таблица значений статистик различных критериев соответствия.

Отметим, что компанией Wolfram Research создан и регулярно обновляется систематизированный каталог онлайновых интерактивных демонстраций [10], свободно распространяемых в форматах NB и CDF интерактивных программных приложений-проектов. По состоянию на сентябрь 2015 г. в каталоге размещены и доступны посетителям сайта более 10295 демонстраций по разным разделам науки, техники, жизни (Mathematics, Computation, Physical Sciences, Life Sciences, Business & Social Systems, Engineering & Technology, Systems, Models & Methods, Our World, Creative Arts, Kids & Fun, Mathematica Functionality, Browse by US Common).

Приведём несколько проектов из каталога [10], которые в классе статистических программных продуктов демонстрируют опущенные и не обсуждаемые в настоящем изложении уникальные возможности интерактивности приложений и графической визуализации в СКА *Mathematica*: “Illustrating the Use of Discrete Distributions”, “Stock Price Probability with Stable Distributions”, “Stock Price Simulation Using Stable Random Variables”, “Minimal Model of Simulating Prices of Financial Securities Using an Iterated Finite Automaton”, “Cumulative Sums and Visual Change Detection between Two Random Processes”, “Cluster Analysis”, “Exploring Multivariate Data”, “Regression toward the Mean”, “Comparing Regression Models with and without Data Transformation”, “k-Nearest Neighbor (kNN) Classifier”, “Simulating a Catastrophe Insurer”, “Credit Risk”, “Constant Risk Aversion Utility Functions”.

Заключение

Упомянутые выше реализованные в Wolfram *Mathematica* методы и функции только частично отражают широкий спектр возможностей этой СКА. Приведены лишь основные, но даже такой список убедительно свидетельствует о достоинствах *Mathematica*, целесообразности использования в системе высшего образования, науке, технике, жизни.

1. Айвазян, С. А. Инструменты статистического анализа данных / С. А. Айвазян, В. С. Степанов// Мир ПК. – 1997. – № 8. – С. 33-–41.
2. Введение в программные системы и их разработку. Лекция 11: Статистическая обработка данных. [Электронный ресурс] / Режим доступа: http://www.intuit.ru/studies/courses/3632/874/lecture/14309. – Дата доступа: 07.09.2015.
3. What’s New in Mathematica 8. [Электронный ресурс] / Режим доступа: http://www.wolfram.com/mathematica/new-in-8/index.en.html?footer=lang. – Дата доступа: 8.09.2015.
4. Таранчук, В. Б. Основные функции систем компьютерной алгебры : пособие для студентов фак. прикладной математики и информатики / В. Б. Таранчук // – Минск : БГУ, 2013. – 59 с.
5. CDF. Формат вычисляемых документов – Документы оживают благодаря возможностям вычислений. [Электронный ресурс] / Режим доступа: http://www.wolfram.com/cdf. – Дата доступа: 8.09.2015.
6. Таранчук, В. Б. О создании интерактивных образовательных ресурсов с использованием технологий Wolfram. / В. Б. Таранчук// Информатизация образования. – 2014. –№ 1. – С. 78–89.
7. Statistical Data Analysis. [Электронный ресурс] / Режим доступа: http://reference.wolfram.com/language/guide/Statistics.html. – Дата доступа: 8.09.2015.
8. What’s New in Mathematica 9. [Электронный ресурс] / Режим доступа: http://www.wolfram.com/mathematica/new-in-9/index.en.html?footer=lang. – Дата доступа: 8.09.2015.
9. Wolfram *MATHEMATICA*. Наиболее полная система для современных технических вычислений в мире [Электронный ресурс] / Режим доступа: http://www.wolfram.com/mathematica. – Дата доступа: 8.09.2015.
10. Wolfram Demonstrations Project. [Электронный ресурс] / Режим доступа: http://http://demonstrations.wolfram.com. – Дата доступа: 8.09.2015.