Материалы региональной конференции (с участием иностранных ученых) "Вероятностные идеи в науке и философии". – Новосибирск: Ин-т философии и права СО РАН / НГУ. 2003. – С. 110-113.

Компьютерное моделирование как способ познания статистических закономерностей в технике, экономике, естествознании[1]

Лемешко Б.Ю., Постовалов С.Н.
НГТУ, Новосибирск. E-mail: headrd@fpm.ami.nstu.ru

Вероятностные идеи и методы в той или иной мере используются во всех сферах человеческой деятельности. Степень их использования зависит от сложности системы и уровня знаний об объекте исследования. Хотя все, в конце концов, познаваемо, неопределенность в знаниях о структуре сложной системы, о детерминированных закономерностях ее функционирования, о множестве внешних факторов и характере их воздействия на систему всегда существует. Знание вероятностных (статистических) закономерностей процессов, происходящих в системе, дополняют известные физические (детерминированные) законы, определяющие функционирование системы.

Множество вероятностных законов (закономерностей), реально наблюдаемых в различных приложения, вообще говоря, бесконечно. И в практике статистического анализа возникает неизмеримо больше различных постановок задач, чем предлагается решений в классической математической статистике. Разнообразие статистических гипотез, выдвигаемых в процессе статистического анализа в различных приложениях, оказывается существенно шире предлагаемого классическим аппаратом. В частности аппарат математической статистики включает в себя ограниченный перечень задач проверки статистических гипотез, для которых найдены предельные распределения статистик, используемых в соответствующих критериях. Классические результаты оказываются применимыми при достаточно строгих предположениях, которые на практике очень часто не выполняются.

Например, классический аппарат проверки гипотез в корреляционном анализе многомерных случайных величин опирается на многомерный нормальный закон. Проверка адекватности регрессионных моделей базируется на нормальности ошибок наблюдений отклика. В классическом дисперсионном анализе также опираются на нормальность ошибок. В приложениях это не всегда справедливо.

Немаловажен и такой аспект. Большинство наиболее весомых результатов в математической статистике имеет асимптотический характер. На практике же, как правило, приходится оперировать с выборками наблюдений ограниченного объема. И свойства используемых статистик в таких ситуациях порой существенно отличаются от асимптотических.

Таким образом, можно говорить о наличии в математической статистике множества "белых пятен", которые не позволяют эффективно применять статистические методы в различных приложениях. Такие "белые пятна" чаще всего оказываются связанными с проверкой статистических гипотез. Вопрос, как правило, упирается в необходимость нахождения предельного распределения статистики построенного критерия. Нахождение предельного закона для статистики конкретного критерия аналитическими методами оказывается чрезвычайно сложной задачей, а задач, требующих разрешения, – слишком много. На настоящем этапе развития математической (и прикладной) статистики можно констатировать, что количество и уровень сложности задач, выдвигаемых практикой, возрастают настолько быстро, что ресурсы человеческого интеллекта, его производительность просто не в состоянии обеспечить решение такого множества задач без создания и использования соответствующих вычислительных технологий.

Отсутствие теоретического обоснования решения вообще или получения оптимального решения, отсутствие четко очерченных границ применения методов анализа приводит на практике к неэффективным или, более того, некорректным статистическим выводам. Сказанное касается и разделов математической статистики, наиболее широко используемых в приложениях.

[1] Работа выполнена при поддержке Минобразования РФ (проект № Т02-3.3-3356)

Все это подчеркивает необходимость (а практика уже показывает возможность) развития компьютерных методов исследования статистических закономерностей, компьютерных методов исследования свойств оценок и статистик различных критериев проверки статистических гипотез, построения вероятностных моделей для исследуемых закономерностей. Компьютерные (вычислительные) методы позволяют с меньшими интеллектуальными затратами получать фундаментальные знания в области математической статистики, и, следовательно, осуществлять корректные статистические выводы при анализе данных в различных прикладных областях.

В последние годы при исследовании некоторых задач математической и прикладной статистики нами получен ряд результатов, связанных как с реализацией вычислительных алгоритмов, обеспечивающих построение оценок параметров с наилучшими свойствами при различной форме регистрации наблюдений, так и с исследованием статистических закономерностей методами компьютерного моделирования. Накопленный опыт показал, что с использованием методов статистического моделирования и последующего анализа можно получать результаты по точности не уступающие аналитическим. Были построены достаточно простые модели законов распределений статистик различных критериев для целого множества проверяемых сложных гипотез. Для ряда задач методами компьютерного моделирования и анализа были найдены достаточно точные решения, которые десятилетиями не удавалось получить аналитическими средствами. Появилась обоснованная уверенность, что с использованием данного подхода можно закрывать многие существующие в статистике “белые пятна”, применяя относительно простой вычислительный и математический аппарат.

При исследовании различных методов оценивания и свойств оценок было показано, что высокой устойчивостью к различным отклонениям от предположений и к наличию аномальных наблюдений обладают оценки максимального правдоподобия (ОМП) по группированным данным. Группирование при оценивании позволяет получать устойчивые оценки. Повышению качества таких оценок способствует применение асимптотически оптимального группирования, минимизирующего потери в информации Фишера. Предложены и исследованы оптимальные L-оценки параметров сдвига и масштаба, построенные на оценках квантилей, соответствующих асимптотически оптимальному группированию. Для ряда параметрических моделей законов распределений, наиболее часто используемых в приложениях, построены таблицы коэффициентов для таких оценок (для 15 моделей законов при различном числе используемых квантилей), исследованы их свойства. Показана эффективность параметрической отбраковки аномальных наблюдений с использованием предложенных робастных методов оценивания.

При использовании критериев согласия типа χ^2 неоднозначность при построении и вычислении статистик бывает связана с выбором числа интервалов и тем, каким образом область определения случайной величины разбивается на интервалы. Такой произвол отражается на статистических свойствах применяемых критериев, в частности, на их мощности при различении близких конкурирующих гипотез. Очевидно, что выбор числа интервалов и способа разбиения на интервалы следует осуществлять с позиций обеспечения максимальной мощности критерия по отношению к близким альтернативам.

Способ группирования оказывает особенно сильное влияние на мощность критериев типа χ^2. Нами показано, что критерии согласия χ^2 Пирсона и отношения правдоподобия при проверке как простых, так и сложных гипотез имеют максимальную мощность против близких альтернатив, если использовать такое разбиение области определения случайной величины на интервалы, при котором потери в информации Фишера о параметрах закона, соответствующего проверяемой гипотезе H_0, минимальны (асимптотически оптимальное группирование).

За всю историю применения критериев типа χ^2 была предложена не одна формула для выбора числа интервалов, но ни одна из представленных в различных рекомендациях не выводилась с позиций максимальной мощности применяемого критерия, а, в основном, исходя из близости плотности к ее непараметрической оценке, гистограмме. Исследование мощности критериев типа χ^2 как функции от объема выборки n и числа интервалов k показали, во-первых, что действительно с ростом k происходит падение мощности, что согласуется с результатами Д.М. Чибисова и А.А. Боровкова. Во-вторых, для любой пары альтернатив и объема выборки существует оптимальное k, при котором мощность максимальна. Оптимальное число интервалов k зависит от объема выборки n и от конкретной пары конкурирующих гипотез H_0 и H_1. Чаще всего оптимальное k оказывается существенно меньше значений, рекомендуемых различными регламентирующими документами и задаваемых множеством эмпирических формул.

Результаты исследований мощности критериев типа χ^2 в зависимости от способа группирования и числа интервалов, полученные таблицы асимптотически оптимального группирования составили основу рекомендаций по стандартизации Р 50.1.033-2001.

К наиболее используемым критериям согласия относятся непараметрические критерии типа Колмогорова, типа ω^2 (Крамера-Мизеса-Смирнова) и Ω^2 (Андерсона-Дарлинга) Мизеса.

В случае простых гипотез предельные распределения статистик непараметрических критериев типа Колмогорова, ω^2 и Ω^2 Мизеса известны давно и не зависят от вида наблюдаемого закона распределения и значений его параметров. Говорят, что они являются "свободными от распределения". Это достоинство предопределило широкое использование данных критериев в приложениях.

При проверке сложных гипотез, когда по той же самой выборке оцениваются параметры наблюдаемого закона $F(x,\theta)$, непараметрические критерии согласия теряют свойство "свободы от распределения". Различия в предельных распределениях тех же самых статистик при проверке простых и сложных гипотез очень существенны. При проверке сложных гипотез на условный закон распределения статистики $G(S|H_0)$ влияет целый ряд факторов, определяющих "сложность" гипотезы: вид наблюдаемого закона $F(x,\theta)$, соответствующего истинной гипотезе H_0; тип оцениваемого параметра и количество оцениваемых параметров; в некоторых ситуациях конкретное значение параметра (например, в случае гамма-распределения); используемый метод оценивания параметров и точность вычисления оценок.

Исходной точкой для исследований непараметрических критериев согласия при сложных гипотезах послужила работа Каса-Кифера-Вольфовица. В литературных источниках изложен ряд подходов к использованию непараметрических критериев согласия в случае проверки сложных гипотез. При достаточно большом объеме выборки ее можно разбить на две части и по одной из них оценивать параметры, а по другой проверять согласие. В некоторых частных случаях предельные распределения статистик исследовались аналитическими методами, процентные точки распределений строились методами статистического моделирования. Для приближенного вычисления вероятностей "согласия" вида $P\{S > S^*\}$ (достигаемого уровня значимости) строились формулы, дающие достаточно хорошие приближения при малых значениях соответствующих вероятностей.

В наших работах исследование распределений статистик непараметрических критериев согласия и построение моделей этих распределений осуществлялось с использованием методики компьютерного анализа статистических закономерностей. Были построены модели распределений статистик при проверке согласия с параметрическими моделями. Исследовалась возможность применения непараметрических критериев согласия для проверки адекватности непараметрических моделей. Построенные в результате применения методики модели предельных распределений статистик рассматриваемых критериев при проверке различных сложных гипотез и таблицы процентных точек послужили основой рекомендаций по стандартизации Р 50.1.037-2002.

С задачей обработки цензурированных выборок, когда наблюдению оказывается доступной только часть области определения случайной величины, а для выборочных значений, попавших левее и/или правее этой области, фиксируется лишь сам факт этого попадания, приходится сталкиваться в различных приложениях. Особенно часто с цензурированными выборками встречаются в задачах надежности при оценивании продолжительности жизни. В такой неполной (цензурированной) выборке содержится меньше информации, чем в полной. Потеря части информации отражается на точности оценивания параметров аппроксимирующего закона распределения. При цензурировании наблюдений снижается способность критериев согласия различать близкие законы распределения.

Проведенные нами исследования потерь в информации Фишера в зависимости от степени цензурирования для различных законов распределения показали, что в некоторых случаях, когда доступной наблюдению оказывается лишь незначительная область определения случайной величины, в цензурированной выборке сохраняется достаточно много информации.

ОМП параметров распределений по цензурированным наблюдениям являются асимптотически эффективными. Однако при ограниченных объемах выборок и значительной степени цензурирования законы распределения ОМП весьма далеки от асимптотически нормального и, более того, оказываются *асимметричными*, а сами оценки *смещенными*. С уменьшением объемов выборок n и увеличением степени цензурирования увеличивается асимметрия закона распределения оценок. С использованием методики компьютерного моделирования и анализа статистических закономерностей были исследованы величины смещения ОМП параметров некоторых законов в зависимости от объема всей выборки n и величины наблюдаемой её части.

Для проверки согласия при цензурированных наблюдениях и простых гипотезах могут использоваться критерии типа Реньи, которые в этой ситуации являются "свободными от

распределения”. Но даже при проверке простых гипотез отмечена сильная зависимость распределений статистик типа Реньи от объема выборки n, что резко ограничивает возможность применения критерия при конечных объемах выборок.

Проведенные нами в исследования распределений ряда статистик корреляционного анализа в случае многомерных законов, отличающихся от нормального в достаточно широких пределах (более островершинных или более плосковершинных, но симметричных), показали, что значимого изменения предельных распределений статистик не происходит. Эмпирические распределения данных статистик по-прежнему хорошо описываются предельными законами, полученными в классическом корреляционном анализе в предположении о нормальности наблюдаемого вектора. Это существенно расширяет сферу корректного применения методов классического корреляционного анализа в приложениях. Данные выводы не касаются задач проверки гипотез о ковариационных матрицах многомерного закона. Есть основания полагать, что предельные распределения статистик, используемых при проверке таких гипотез, существенно зависят от наблюдаемого многомерного закона.

В классических регрессионном и дисперсионном анализах аппарат проверки статистических гипотез базируется на предположении нормальности закона ошибок наблюдений. Нарушение данного предположения по-разному отражается на распределениях статистик используемых критериев проверки гипотез. Предельные распределения статистик критериев могут зависеть от закона распределения ошибок и применяемого метода оценивания параметров. Проведенные численные исследования распределений статистик, используемых при проверке гипотез об адекватности линейных регрессионных моделей, в случае ошибок наблюдений отклика, подчиняющихся законам распределения, не совпадающим с нормальным, и при применении для оценивания параметров регрессии метода максимального правдоподобия показали, что в этом случае распределения статистик уже не подчиняются *F*-распределению Фишера. В то же время эмпирические распределения статистик хорошо описываются бета-распределениями II рода, частным случаем которого является *F*-распределение Фишера. Аналогичные исследования распределений статистик в дисперсионном анализе показали их существенную зависимость от закона распределения ошибок и перспективность применения развиваемой методики.

В целом, опираясь на результаты многочисленных исследований, можно утверждать, что методы компьютерного моделирования и статистического анализа являются наиболее эффективным аппаратом исследования вероятностных закономерностей, возникающих в технике, экономике, естествознании.