УДК 519.25

ПРОБЛЕМЫ ИДЕНТИФИКАЦИИ МОДЕЛЕЙ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН С ПРИМЕНЕНИЕМ СОВРЕМЕННОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Бахрушин В.Е.

Классический приватный университет, Запорожье, e-mail: Vladimir.Bakhrushin@zhu.edu.ua

Рассмотрены некоторые проблемы идентификации моделей распределения данных, при использовании современного математического аппарата для решения этой задачи. Показано, что использование методов нелинейной оптимизации для идентификации моделей приводит к улучшению результатов идентификации, но одновременно, изменяет формальную постановку задачи. Выделено три группы проблем, связанных с выбором критериев согласия, их критических значений и проверкой адекватности получаемых моделей. Проанализированы возможные подходы к решению этих проблем.

Ключевые слова: распределение данных, модель, идентификация, оптимизация, критерий согласия, адекватность

SOME PROBLEMS OF IDENTIFICATION OF RANDOM VALUES DISTRIBUTION MODELS WHEN USING THE MODERN SOFTWARE

Bakhrushin V.E.

Classic Private University, Zaporozhye, e-mail: Vladimir.Bakhrushin@zhu.edu.ua

Some problems of identification of data distribution models, which arise at using of modern mathematical tools, are discussed. It is shown, that the application of non-linear optimization methods for models identification improves it results but at the same time it changes the formal problem definition. Three groups of problems, related to fitting criterion choice, their critical values estimation and received models verification, are marked out.

Keywords: data distribution, model, identification, optimization, fitting criterion, validity

Задача идентификации моделей распределения выборок часто встречается в различных прикладных исследованиях. В частности, большинство параметрических методов статистического анализа данных предполагает предварительную проверку гипотезы о нормальности закона распределения исследуемых данных [1]. Еще одним примером являются параметрические методы классификации без обучения, которые исходят из того, что распределение данных можно представить в виде смеси распределений известного типа (как правило, нормальных) и включают этап идентификации функции распределения по исходным данным [2]. Знание закона распределения часто бывает необходимым при построении имитационных моделей сложных систем [3], при разработке статистических методов контроля качества на производстве [4], для создания методик обработки данных [5] и т.п.

Традиционные методы идентификации были разработаны в первой половине XX в. и не предполагают необходимости использования современной вычислительной техники. Ее появление и широкое использование в статистических исследованиях не только позволило существенно ускорить и облегчить процедуру идентификации, но и создало потенциальную возможность использовать для решения этой задачи более сложные математические методы, в частности методы решения задач нелинейной оптимизации. Однако их применение может изменять формальную постановку задачи идентификации, что создает ряд новых проблем.

Параллельно с развитием методов идентификации моделей распределения развивались общие методы идентификации математических и регрессионных моделей. При этом сложилась ситуация, когда однотипные задачи в разных областях решаются по-разному.

Целью данной работы является формулировка некоторых проблем, возникающих при применении современных математических методов для решения задачи идентификации законов распределения случайных величин, и анализ возможных путей их решения.

Традиционные методы идентификации законов распределения

Обычная процедура идентификации законов распределения случайных величин предполагает два основных этапа исследования – выдвижение гипотез о законе распределения и их проверку на основе тех или иных статистических критериев [1, 6]. При этом формальная постановка задачи на втором этапе может быть различной. В статистике ее обычно формулируют как проверку нулевой гипотезы о том, что имеющиеся данные соответствуют некоторому полностью определенному закону распределения либо распределению, принадлежащему некоторому параметрически заданному семейству, параметры которого необходимо

оценить в процессе идентификации (простая и сложная гипотезы) [7]. Для решения этой задачи по имеющимся эмпирическим данным вычисляют значение соответствующего критерия и сравнивают его с критической величиной для заданного уровня значимости. При этом возможны ошибки принятия неправильной нулевой гипотезы или отклонения правильной. Разрабатывая критерии, эти ошибки стремятся минимизировать, но сделать их равными нулю принципиально невозможно. К тому же снижение вероятности одной из ошибок ведет увеличению вероятности другой. Наиболее часто используют критерии типа омега-квадрат, Колмогорова–Смирнова и хи-квадрат.

Критерий омега-квадрат был предложен в 1928–1930 г. Х. Крамером и Р. фон Мизесом, и на сегодняшний день он является наиболее мощным из непараметрических критериев согласия [6]. Его расчетное значение определяют [1] по формуле:

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

где $F(x_i)$ – значение теоретической функции распределения в точке x_i; n – объем выборки; i – индекс, используемый для нумерации значений ее элементов, упорядоченных в порядке возрастания. При $n > 40$ критические значения критерия можно определить по специальным таблицам.

Критерий Колмогорова–Смирнова был разработан в 1930-х г. А.Н. Колмогоровым и Н.В. Смирновым. Его расчетное значение для двусторонней гипотезы определяют [1] по формулам:

$$D_n = \max_{1\le i\le n}\left\{D_n^{(1)}, D_n^{(2)}\right\};$$

$$D_n^{(1)} = \max_{1\le i\le n}\left\{\frac{i}{n} - F\left(x_i\right)\right\}; \quad (2)$$

$$D_n^{(2)} = \max_{1\le i\le n}\left\{F\left(x_i\right) - \frac{i-1}{n}\right\},$$

а для односторонней $D_n = D_n^{(1)}$.

Критерий Колмогорова–Смирнова несколько уступает по мощности критерию омега-квадрат [6], однако его преимуществом является то, что при $n > 35$ критические значения можно определять не по таблицам, а рассчитывать по асимптотической формуле:

$$D_{n\alpha} \approx \sqrt{-\frac{\ln\frac{\alpha}{2}}{2n}}, \quad (3)$$

где α – уровень значимости.

В обоих случаях критические значения зависят от выбранного вида теоретической функции распределения и способа оценивания ее параметров. Формула (3) и таблицы дают критические значения для случая, когда параметры определяются независимо от исследуемой выборки. Если же их рассчитывают непосредственно по выборке (например, определяют как выборочные среднее арифметическое и стандартное отклонение для нормального закона распределения), то критические значения должны быть существенно уменьшены.

Критерий хи-квадрат предложен в 1900 г. К. Пирсоном. В отличие от двух предыдущих, для его использования производят предварительную группировку данных по интервалам равной ширины. Значение критерия рассчитывают [1] по формуле:

$$\chi^2 = \sum_{i=1}^{k}\frac{\left(\nu_i - np_i'\right)^2}{np_i'}, \quad (4)$$

где ν_i – абсолютные частоты для k классов; p_i' – теоретические вероятности попадания данных в соответствующий интервал для выбранного распределения; n – общее число наблюдений. Число степеней свободы берут равным $k - r - 1$, где r – число параметров модели распределения. В частности, при расчете параметров модели по интервальному вариационному ряду число степеней свободы берут равным $k - 2$ для биномиального $i\ k - 3$ – для нормального распределения.

Наряду с этим возможен другой подход к формальной постановке задачи на втором этапе идентификации. В соответствии с общей методологией идентификации математических моделей она может быть сформулирована как подбор модели, которая в некотором смысле наилучшим образом описывает имеющийся набор эмпирических данных. Для решения этой задачи необходимо задать тип модели и подобрать ее параметры минимизацией заданного целевого функционала. В качестве функционала обычно используют сумму квадратов остатков модели, сумму их модулей или максимальный по модулю остаток. При таком подходе к идентификации моделей используются другие критерии адекватности, которые будут рассмотрены ниже. Во многих реальных задачах такая постановка задачи может быть более корректной, чем традиционная, поскольку предполагается, что любая модель лишь приближенно отображает реальный объект исследования. Поэтому не ставится вопрос о ее правильности, а проверяется лишь ее адекватность, т.е. возможность использования анализиру-

емой модели для решения некоторой конкретной задачи.

Проблема определения критических значений

Первая из проблем связана с возможностью существенного уменьшения расчетных значений для критериев типа омега-квадрат и Колмогорова–Смирнова за счет оптимизации параметров подбираемых моделей распределения путем решения задачи минимизации критериального показателя.

Из (1) видно, что критическое значение критерия омега-квадрат по смыслу является максимально допустимой (при заданном уровне значимости) суммой квадратов отклонений теоретической функции распределения (т.е. получаемой в результате идентификации модели распределения) от эмпирической. Аналогично, из (2) следует, что критическое значение критерия Колмогорова–Смирнова является предельно допустимым значением максимального отклонения теоретической функции распределения от эмпирической.

Практика применения рассматриваемого подхода для идентификации моделей распределения различных типов данных показывает, что решение задачи минимизации критериальных показателей (1, 2, 4) во многих случаях позволяет существенно снизить их значения по сравнению с моделями, параметры которых определяются непосредственным расчетом по выборочным данным. В качестве примера на рис. 1 показаны результаты подбора модели нормального распределения для показателей рейтинга ТОП-200 вузов Украины – 2011.

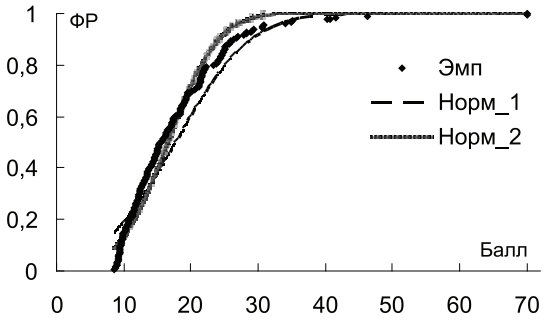

Рис. 1. Функция распределения рейтинга ТОП-200 вузов Украины

Модель Норм_1 была получена при использовании в качестве параметров распределения выборочных среднего арифметического и стандартного отклонения, а модель Норм_2 – минимизацией максимального по модулю остатка модели при использовании параметров Норм_1 в качестве начального приближения.

В результате оптимизации расчетное значение критерия Колмогорова – Смирнова удалось снизить с 2,06 до 1,19, а максимальный остаток с 0,146 до 0,084. При этом оценка математического ожидания изменилась от 16,6 до 17,8, а оценка стандартного отклонения – от 5,9 до 8,8. Следует отметить, что полученное для оптимизированной модели расчетное значение критерия меньше, чем критическое значение для простой гипотезы, но превышает критическое значение (0,895 при уровне значимости 0,95 [6]) для случая, когда в качестве оценок параметров нормального распределения берут выборочные среднее арифметическое и стандартное отклонение. Поэтому в рассматриваемом случае в качестве более адекватной была выбрана модель однородного логнормального распределения, для которой расчетное значение критерия было близко к 0,3.

Как указывалось выше, критические значения рассмотренных критериев зависят от способа задания параметров модели распределения. Это связано с тем, что формально мы переходим от проверки простой гипотезы (соответствия выборки заданному закону распределения) к сложной (соответствия выборки параметрически заданному закону распределения, параметры которого необходимо определить в ходе этой проверки). В этом случае изменяется распределение статистики используемого критерия [8, 9], которая зависит не только от способа оценивания параметров, но и от выбора модели распределения. Поэтому можно ожидать, что в случае, когда параметры определяют не прямым расчетом по выборочным данным, а путем минимизации некоторого целевого функционала, критические значения могут оказаться меньшими, чем величины, рекомендуемые при определении параметров по выборке, и, тем более чем значения, рекомендуемые для проверки простой гипотезы. Дополнительные проблемы могут быть связаны с неустойчивостью процедуры минимизации, что характерно для сложных моделей распределения, и с возможностью выбора различных алгоритмов и начальных приближений для этой процедуры.

В связи с этим иногда делается вывод, что развитие методов оценивания согласия эмпирических выборок с параметрическими семействами распределений относится к тупиковым направлениям, поскольку ни одна реальная выборка не может в точности соответствовать никакому параметрическому семейству [8]. Однако такой вывод, на наш взгляд, является излишне категоричным, поскольку идентификация моделей распределений реальных данных, как правило, является не самоцелью, а частью решения более сложных прикладных проблем. Теоретические законы (модели) распределения всегда являются следствием некоторых содержательных предположений. Подтверждение или отклонение гипотез о соответствии имеющихся данных той или иной модели распределения одновременно можно рассматривать, как подтверждение правильности или ложности этих исходных предположений. Во многих случаях такой вывод является более важным, чем наличие несущественных отклонений от полученной модели.

Проблема выбора критерия согласия

При решении задачи идентификации законов распределения обычно считают, что модель адекватна при заданном уровне значимости, если расчетное значение одного из критериев (1, 2, 4) не превышает соответствующего критического значения. При этом возникают две группы проблем – выбор критерия согласия и соответствие полученной модели общим критериям адекватности математических моделей.

Формально, ответ на первый вопрос дает анализ мощности рассматриваемых критериев, по результатам которого сделан вывод, что она убывает в ряду $1 \rightarrow 2 \rightarrow 4$ [6]. Соответственно, рекомендуется выбирать для использования наиболее мощный из критериев, приемлемых для соответствующего набора данных.

Вместе с тем, этот вывод нуждается в некотором уточнении. Критерии типа омега-квадрат и Колмогорова – Смирнова базируются на сравнении эмпирической функции распределения с теоретической моделью. Одним из свойств функции распределения F(x) является [1, 7] то, что:

$$\lim_{x\to-\infty} F(x) = 0; \quad \lim_{x\to+\infty} F(x) = 1. \qquad (5)$$

В силу этого рассматриваемые критерии значительно более чувствительны к отклонениям от теоретического закона вблизи центра распределения, чем к отклонениям вдали от него. Поэтому можно ожидать, что критерии типа омега-квадрат и Колмогорова–Смирнова будут более мощными в тех случаях, когда различие функций распределения обусловлено, главным образом, различием моментов низких порядков (математического ожидания и дисперсии).

Однако при анализе некоторых типов реальных данных, в частности при подборе моделей неоднородных распределений, важным является наличие отклонений на всем интервале вариации данных. Критерий хи-квадрат более чувствителен к отклонениям на краях области вариации данных и, соответственно, к различию моментов высоких порядков. Поэтому в ситуациях, когда такие отклонения важны для решения конкретной задачи, целесообразно проводить проверку с совместным использованием как одного из критериев (1, 2), так и критерия (4).

В частности, в работе [10] нами было показано, что при идентификации распределения результатов Единого государственного экзамена Российской Федерации по многим дисциплинам с использованием критерия Колмогорова–Смирнова можно подобрать модель однородного нормального распределения, которая удовлетворяет этому критерию, несмотря на высокие значения коэффициентов асимметрии, а в отдельных случаях и отчетливо видимую на гистограмме неоднородность выборки. В то же время расчетное значение критерия хи-квадрат при этом существенно превышает критическое для той же модели распределения. Совместное использование этих критериев позволяет получить более адекватную модель распределения в виде суммы нормально распределенных компонент.

Проблема адекватности модели

Общий подход к анализу адекватности математических и регрессионных моделей [1] предполагает необходимость использования критериев, проверяющих совокупность свойств остатков модели, которые должны подчиняться нормальному закону распределения с нулевым средним арифметическим и быть независимыми друг от друга. В то же время при идентификации моделей законов распределения свойство независимости остатков часто нарушается.

Эта проблема становится особенно актуальной при уменьшении объема исследуемых выборок, что иллюстрирует рис. 2. На нем приведены графики зависимости остатков моделей нормального распределения для двух выборок объемом 50 и 100 элементов, сгенерированных в соответствии со стандартным нормальным распределением.

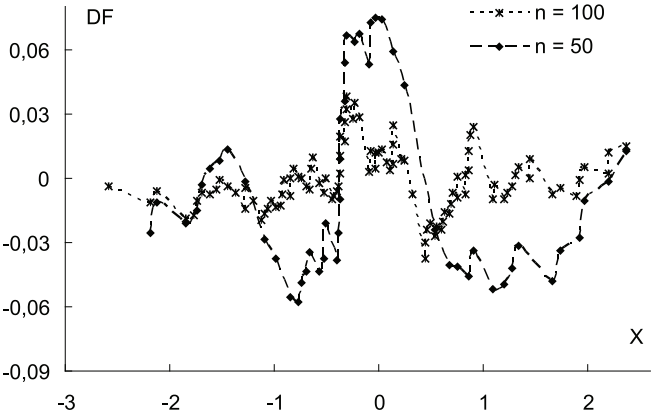

Рис. 2. Графики остатков моделей нормального распределения для выборок разного объема

Параметры моделей были определены минимизацией максимального по модулю остатка. Из рис. 2 видно, что остатки не являются независимыми. Значение критерия Дарбина–Уотсона в рассмотренных случаях составляет 0,058 для выборки объемом 50 элементов и 0,35 для выборки из 100 элементов. Это подтверждает значимую корреляцию остатков моделей распределения. При этом следует отметить, что расчетные значения критерия (2) для найденных моделей существенно меньше критических, определяемых по стандартной методике, и находятся в пределах 0,2–0,5.

С формальной точки зрения возможна ситуация, когда при удовлетворении всех трех критериев (1, 2, 4) все остатки будут иметь один знак, т.е. эмпирическая выборка будет иметь незначительный сдвиг относительно рассматриваемой модели. Несмотря на выполнение критериев согласия, понятно, что эти модели (для простых законов распределения) могут быть легко улучшены изменением их параметров. Однако вопрос о существенности таких отклонений должен решаться отдельно для каждой конкретной задачи.

По нашему мнению, вопрос о целесообразности использования и области применимости общих критериев адекватности моделей в задачах идентификации законов распределения нуждается в дальнейшем исследовании. При этом в соответствии с общей методологией проверки адекватности моделей, следует исходить из способности достижения конкретных задач исследования с помощью рассматриваемой модели.

Выводы

Проведенный анализ показывает, что применение современных компьютерных технологий для решения задач идентификации законов распределения данных порождает ряд проблем, в числе которых большое значение имеют:

– неопределенность критических значений используемых статистик в случае оценивания параметров распределений минимизацией критериальных показателей;

– возможность совместного использования нескольких критериев согласия в случаях, когда возможно различие моментов высоких порядков теоретической модели и исследуемой выборки;

– вопрос о целесообразности использования и области применимости общих критериев адекватности моделей в задачах идентификации законов распределения.

Список литературы

1. Бахрушин В.Є. Методи аналізу даних. – Запоріжжя: КПУ, 2011. –268 с.

2. Айвазян С.А., Буштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерностей. – М.: Финансы и статистика, 1989. – 607 с.

3. Емельянов А.А., Власова Е.А., Дума Р.В. Имитационное моделирование экономических процессов. – М.: Финансы и статистика, 2002. – 368 с.

4. Логанина В.И., Федосеев А.А. Статистические методы контроля и управления качеством продукции. – М.: Феникс, 2007. – 219 с.

5. Румшиский Л.З. Математическая обработка результатов эксперимента. – М.: Наука. ФИЗМАТЛИТ, 1971 – 192 с.

6. Орлов А.И. Прикладная статистика. – М.: Экзамен, 2006. – 672 с.

7. Боровков А.А. Математическая статистика. Оценка параметров. Проверка гипотез. – М.: Наука, 1984. – 472 с.

8. Орлов А.И. О критериях согласия с параметрическим семейством // Заводская лаборатория. – 1997. – Т. 63, № 5. – С. 49–50.

9. Durbin, J. *Distribution theory for tests based on the sample distribution function*. –SIAM, 1973. – 74 p.; Мартынов Г.В. Статистические критерии, основанные на эмпирических процессах, и связанные с ними вопросы // Итоги науки и техники: Сер. Теория вероятностей. Математическая статистика. Теоретическая кибернетика. – М.: ВИНИТИ, 1992. – Т. 30. – С. 3 – 112; Лемешко Б.Ю. Об ошибках, совершаемых при использовании непараметрических критериев согласия // Измерительная техника. – 2004. – № 2. – С. 15–20.

10. Бахрушин В.Е., Журавель С.В., Игнахина М.А. Эмпирические функции распределения результатов тестирования выпускников школ // Управляющие системы и машины. – 2009. – № 2. – С. 82–84.