УДК 519.216+519.224

Шепель В.Н., Акимов С.С.
Оренбургский Государственный Университет
E-mail: elite17@yandex.ru

ИСПОЛЬЗОВАНИЕ ОЦЕНКИ ХИЛЛА ДЛЯ РАЗЛИЧЕНИЯ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ

Рассматривается проблема использование оценки Хилла для восстановления плотности распределения вероятности по выборке. Определен пороговый индекс для проверки распределения на симметричность, а также для грубой предварительной проверки на нормальность
Ключевые слова: оценка Хилла, индекс тяжести, закон распределения вероятности

Одна из основных задач математической статистики – восстановить закон распределения вероятностей случайной величины на основе конкретных результатов проведенных наблюдений или экспериментов. Эта задача объясняется тем, что в практике статистического анализа и моделирования точный вид закона распределения анализируемой генеральной совокупности, как правило, бывает неизвестен. Исследователь располагает лишь выборкой из интересующей его генеральной совокупности [4].

Существует достаточно большое количество процедур восстановления плотности вероятности по выборке – интегральная оценка, дескриптивное приближение сплайнами, проекционный метод, стохастическая регуляризация, рекуррентная ядерная оценка, корневая оценка, гистограммный метод и ряд других.

Несмотря на многообразие методов, достаточно часто их результаты являются недостоверными или значительно искаженными. На это влияет множество причин, одна из которых – многообразие законов распределения. Законы распределения вероятностей по своей природе настолько разнообразны, что единый подход к их оценке конкретным методом является несостоятельным. Решением подобной проблемы может стать классификация законов распределения по какому-либо признаку, основанная на легкопроверяемых характеристиках изучаемого массива данных; рассмотрение ряда характеристик массивов данных [4], [12], [13], [14], [15] позволяет сделать вывод о том, что ключевой характеристикой может выступать тяжесть хвоста закона распределения.

Большинство применяемых процедур для оценки плотности вероятности справедливы для распределений с легкими хвостами.

Природа же законов распределения вероятностей случайных величин с тяжелыми хвостами такова, что стандартные процедуры оценивания их параметров, зачастую, несостоятельны. Рассмотрение и статистический анализ характеристик случайных массивов данных, распределение которых описывается при помощи тяжелохвостовых моделей, требует специальных, неклассических методов статистики. Например, гистограммный метод достаточно точно оценивают легкохвостовые плотности распределения вероятностей, но для массивов с тяжелыми хвостами, в условиях ограниченности выборки, дают вводящие в заблуждение результаты. Помимо гистограммной оценки, данная погрешность справедлива и для многих других оценок: сплайн оценок, проекционных оценок, ядерных оценок. Об этом подробно написано в работе Маркович Н. М.[5].

Основная причина несостоятельности стандартных процедур оценивания вытекает из нарушения условия Крамера касательно существования производящей функции моментов. Для тяжелохвостых законов распределения случайной величины моменты попросту отсутствуют (в качестве примера очень удачным является распределение Коши). Соответственно нарушается и теорема Крамера о сходимости хвостов, согласно которой хвосты распределения конечной суммы случайных величин, независимых от какого-либо стороннего фактора, сходятся к хвосту нормального распределения.

В той же работе [5] показаны недостатки, как классических методов, так и достаточно оригинальных, в том числе и различных непараметрических оценок.

Цель исследования: определить, к какому виду распределений (с легкими или с тяжелыми хвостами) относится исследуемая выборка данных из генеральной совокупности с неизвестной плотностью распределения.

Задачи исследования:

– определить способ расчета оценки тяжести хвоста;

– определить пороговое значение оценки, исходя из которой распределение можно отнести к легкохвостовому или тяжелохвостовому;

– произвести расчеты оценки для распределений с заведомо тяжелыми и легкими хвостами, с целью проверки адекватности модели.

Исследования выполнены в лабораторных условиях кафедры управления и информатики в технических системах Оренбургского государственного университета с применением генератора случайных чисел программы Mathcad 15. Были сгенерированы 1000 массивов данных с различающимся количеством исследований (от 10 до 10 000). Для подтверждения опыта практическими примерами, использовались данные приведенные в работах [1], [3], [9], [10]. На описанных массивах данных производилась оценка тяжести хвоста [15].

Для исследования тяжести хвоста была использована оценка Хилла. Оценка Хилла определялась по формуле:

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \ln \frac{x_{(i)}}{x_{(k-1)}}, k < n \ , \qquad (1)$$

Данная формула не дает численный результат, иными словами, необходимы преобразования, которые позволили бы получить определяющее значение оценки. Кроме того, среди исследователей, занимающихся проблемой распределения вероятностей случайных величин нет единого мнения о том, каким должно быть значение k [2], [8].

В проведенном исследовании данный параметр находился для каждой квантили и были построены так называемые «квантильно-квантильные графики» [1]. Далее использовался метод оценки величины индекса тяжести хвоста, представленный в работе А.Н. Гуда, М.А. Бутаковой, Н.А. Москат [1]. Он заключается в построении касательной к полученному квантильно-квантильному графику.

Для описания подобного графика использовалось уравнение регрессии методом наименьших квадратов с использованием полинома второй степени, соответственно, касательная, согласно основному свойству нахождения производной, описана простейшим уравнением прямой.

Далее, определялся тангенс угла в пересечении графика касательной с осью абсцисс. Величина тангенса угла составляет искомый индекс α .

Оценка критического значения тяжести хвоста проводилась эмпирически, основываясь на результатах исследований перечисленных выше авторов. В работе [1] упоминается, что к тяжелохвостым распределениям относятся распределения, хвост которых тяжелее, чем хвост экспоненциального распределения. Поэтому за основу определения критического значения тяжести хвоста было использовано именно это распределение.

В результате построения оценок Хилла для значительного количества экспоненциальных распределений выявлено, что наиболее информативным является поведение хвоста распределения на промежутке между третьим и четвертым квартилем.

При нахождении касательной к графику регрессии на заданном промежутке использовалась медиана заданного промежутка; также выявлено, что выбор точки в окрестностях медианы влияет на наклон графика касательной к оси ординат незначительно.

При увеличении параметра λ экспоненциального распределения оценка Хилла $H_{k,n} \to 0$, тем самым облегчая хвост распределения. Справедливо и обратное утверждение, снижение данного параметра увеличивает оценку.

Для оценки критического значения оценки тяжести хвоста использовалось экспоненциальное распределение с различными параметрами. Было установлено, что тангенс угла касательной, построенной по центральной точке промежутка между третьим и четвертым квартилем в приведенной оценке Хилла стремится к значению 0,55 (исследовались значения λ от 0,5 до 5).

Данное значение использовалось как критическое для оценки других распределений.

По приведенному выше методу, оценивались другие виды распределений. Согласно приведенному исследованию получились следующие данные.

Наименьшее значение оценки распределения Коши с параметрами $(x = 0; \gamma = 1)$ равно 1,049, наибольшее, с параметрами $(x = 10; \gamma = 50)$ равняется 3,124.

Наименьшее значение оценки распределения Вейбулла с параметрами $(k = 5; \lambda = 10)$ равно 0,562, наибольшее, с параметрами $(k = 0; \lambda = 1)$ равняется 1,146.

Наименьшее значение оценки логнормального распределения с параметрами $(\mu = 1; \sigma^2 = 1)$ равно 0,571, наибольшее, с параметрами $(\mu = 5; \sigma^2 = 10)$ равняется 4,867.

Перечисленные выше распределения относятся к распределениям с тяжелыми хвостами и полученный коэффициент превышает критическое значение 0,55.

Приведем некоторые полученные значения для распределений с легкими хвостами.

Наименьшее значение оценки нормального распределения с параметрами $(\mu = 0; \sigma^2 = 1)$ равно 0,295, наибольшее, с параметрами $(\mu = 20; \sigma^2 = 20)$ равняется 0,372.

Наименьшее значение оценки логистического распределения с параметрами $(\mu = 4; s = 1)$ равно 0,322, наибольшее, с параметрами $(\mu = 50; s = 10)$ равняется 0,474.

Наименьшее значение оценки распределения Рэлея с параметром $(\sigma = 2)$ равно 0,098, наибольшее, с параметрами $(\sigma = 15)$ равняется 0,52.

Наименьшее значение оценки гамма-распределения с параметрами $(\lambda = 2; \theta = 5)$ равно 0,389, наибольшее, с параметрами $(\lambda = 5; \theta = 20)$ равняется 0,535.

Справедливость полученной оценки подтверждается, также работами авторов А.Н. Гуда, М.А. Бутакова, Н.А. Москат, Х. Беврани, К.Аничкин [1], [9]. В работе [1] при оценке распределения Парето (распределение с тяжелым хвостом, параметры не указаны) получена оценка 0,87. В работе [9] рассмотрено распределение Стьюдента (также, относящееся к распределению с тяжелыми хвостами) с параметром (t) равным трем. Оценка приведенным выше способом в самой представленной работе не проводилась, однако использование этих данных дало результат, равный 0,593.

Кроме приведенных выше исследовались, также, и другие законы распределения, в частности, бета-распределение, равномерное распределение, геометрическое, гипергеометрическое, биноминальное, Пуассона, Бернулли. Однако индекс 0,55 для дискретных видов распределений не играет столь значительной роли, и при определенном наборе параметров легкохвостые распределения могут превышать критическое значение. Таким образом, для дискретных значений данная оценка не является состоятельной.

Необходимо сделать еще одно замечание. Распределение Бернулли, большинство частных случаев гипергеометрического распределения, а также некоторые другие дискретные распределения при определенном подборе параметров могут задаваться только двумя значениями 0 и 1. Расчет оценки Хилла для такого ограниченного количества значений невозможен. Потому такие случаи заслуживают отдельного рассмотрения.

Таким образом, показано новое применение оценки Хилла в качестве способа отделения легкохвостовых распределений от тяжелохвостовых. Предварительная обработка исходного массива данных согласно приведенной оценке позволяет выявить распределения с легкими хвостами, для восстановления плотности вероятности которых справедливы различные параметрические процедуры, и распределения с тяжелыми хвостами, для которых эти процедуры дают несколько искаженный результат.

Заключение

В работе приведен способ расчета оценки тяжести хвоста, который заключается в построении оценки Хилла. Выявлено, что наиболее информативным в данной оценке является промежуток между третьим и четвертым квартилем. Касательная строится к графику приведенного уравнения регрессии (полином второй степени). Тангенс угла наклона определяет искомую величину индекса.

Также определено пороговое (критическое) значение, равное 0,55; массивы данных, имеющие значения индекса ниже критического, относятся к распределениям с легкими хвостами, выше критического – к распределениям с тяжелыми.

Выявленная оценка прошла проверку на ряде распределений, сгенерированных с помощью программы Mathcad 15. Для повышения достоверности использовались, также, данные, приведенные в статьях других исследователей.

Установлено, что приведенная оценка 0,55 несостоятельна для дискретных распределений, а для некоторых из них (Бернулли, часть гипергеометрических) вообще не может быть произведена.

16.11.2013

Список литературы:
1. Гуда, А.Н. Модели оценки параметров телекоммуникационного трафика в автоматизированных информационно-управляющих системах / А.Н. Гуда, М.А. Бутакова, Н.А. Москат // Вопросы современной науки и практики. Ун-т им. В.И. Вернадского. – 2010. – №4–6(29). – С. 71–87.
2. Ажгиреева, Р. А. Оценка квантилей годовых агрегированных операционных потерь / Р. А. Ажгиреева, О. П. Волобуева // Вестник КазНТУ. – 2011 – № 4 (86). – С. 188–194.
3. Об одном методе моделирования самоподобного стохастического процесса / Ж. В. Дейнеко, А. А. Замула, Л. О. Кириченко, Т. А. Радивилова // Вісн. Харк. нац. ун-ту ім. В. Н. Каразіна. Сер. Математичне моделювання. Інформаційні технології. Автоматизовані системи управління. – 2010. – № 890, вип. 13. – С. 53–63.
4. Шепель, В. Н. Эвристическая процедура определения подходящего распределения вероятности / В. Н. Шепель, С. С. Акимов // Компьютерная интеграция производства и ИПИ-технологии / Сборник материалов V Всероссийской научно-практической конференции. - Оренбург: Изд. ИП Осиночкин Я.В., 2011. – С. 137-139.
5. Маркович, Н. М. Методы оценивания характеристик тяжело-хвостовых случайных величин по конечным выборкам [Электронный ресурс]: Дис. ... д-ра физ.-мат. наук : 05.13.01 : М., 2004 - 206 с. - Режим доступа : http://www.lib.ua-ru.net/diss/cont/58036.html. - 10.11.2013.
6. Mark E. Crovella, Murad Taqqu and Azer Bestavros, "Heavy Tailed-Probability distributions in the World Wide Web" 5(6):835—846, December 1997.
7. Holger Drees, Laurents de Haan, and Sidney Resnick, "How to make a Hill Plot", Sept. 1998.
8. Москат, Н. А. Программно-алгоритмическое обеспечение оценки качества информационного обмена в автоматизированных системах управления железнодорожным транспортом [Электронный ресурс]:Дис. ... к-та тех. наук : 05.13.06 : Ростов-на-Дону., 2010 - 171 с. . - Режим доступа : http://www.dissercat.com/content/programmno-algoritmicheskoe-obespechenie-otsenki-kachestva-informatsionnogo-obmena-v-avtomat. - 02.11.2013.
9. Беврани, Х., Оценка параметров распределений с тяжелыми хвостами с помощью эмпирического распределения / Х. Беврани, К.Аничкин // МКО. – 2005. – ч. 2. – С. 493–495.
10. Денисов, Д. Э., Коршунов Д. А., Фосс С. Г.Нижние пределы для хвостов распределений случайно остановленных сумм / Д. Э. Денисов, Д. А. Коршунов, С. Г. Фосс // ТВП. – 52:4. –2007. – С. 794–802.
11. Denisov D. On lower limits and equivalences for distribution tails of randomly stopped sums. D. Denisov, S. Foss, D. Korshunov // Bernoulli 14 -2008 391–404.
12. Акимов, С. С. Оптимизированный алгоритм определения закона распределения вероятности по выборке из генеральной совокупности / С. С. Акимов. Известия Самарской государственной сельскохозяйственной академии [Текст] : журнал // учредители: М-во с.-х. Российской Федерации, ФГБОУ ВПО СГСХА. - Самара : СГСХА, 2013. вып. № 2. – С. 52-56.
13. Акимов С. С. Расчет вероятности дискретности для массива данных / С. С. Акимов // Научное обозрение [Текст] : журнал – Саратов. – 2013. – С. 78-82.
14. Акимов С. С. Применение коэффициентов асимметрии и эксцесса для определения закона распределения вероятностей / С. С. Акимов. Материали за 9-а международна научна практична конференция, «Новинанта за напреднали наука», - 2013. Том 53. Математика. София. «Бял ГРАД-БГ» ООД – С. 30-33.
15. Akimov S. Kvan O. Several Methods of Determining the Continuous or Discrete Distribution. Horizon Research Publishing All rights reserved, 2013, p. 185-187, DOI: 10.13189/ms.2013.010402.

Сведения об авторах:
Шепель Вячеслав Николаевич, заведующий кафедрой управления и информатики в технических системах Оренбургского государственного университета, доктор экономических наук, профессор
Акимов Сергей Сергеевич, аспирант кафедры управления и информатики в технических системах, факультета информационных технологий Оренбургского государственного университета
460018, г. Оренбург, Шарлыкское шоссе, 5, e-mail: vn_shepel@mail.ru; elite17@yandex.ru