СТАТЬИ

УДК 519.216:519.224

СРАВНЕНИЕ НЕКОТОРЫХ МЕТОДОВ ОЦЕНКИ ТЯЖЕСТИ ХВОСТОВ ПРИ ИДЕНТИФИКАЦИИ ЗАКОНА РАСПРЕДЕЛЕНИЯ

Акимов С.С., Трипкош В.А.

ФГБОУ ВО «Оренбургский государственный университет», Оренбург, e-mail: sergey_akimov_work@mail.ru

В статье рассматривается проблема идентификации закона распределения. Показано, что закон распределения является исчерпывающей характеристикой массива случайных данных. При этом, при отсутствии априорного знания закона распределения, обработка имеющегося у исследователя массива данных представляет собой определенные трудности, решение которых возможно путем идентификации закона распределения. В статье приводится краткий обзор основных методов идентификации законов распределения, при этом отмечено, что в вопросе идентификации закона распределения могут помочь характеристики, вычисляемые непосредственно по исходному массиву данных. Одной из подобных характеристик является величина тяжести хвоста распределения вероятности. Показано, что не существует единой универсальной методики оценки тяжести хвоста, в частности для целей идентификации закона распределения. В статье рассматривается метод оценки Хилла, однако показано, что сам метод достаточно трудоемок при выполнении расчетов. Исходя из этого предлагается новый метод отношений, основанный на соотношении основной и хвостовой частей распределения. При этом решается вопрос об определении границ основной и хвостовой частей распределения по исходному массиву данных. В качестве базы исследования были выбраны пять различных законов распределения, различающихся тяжестью хвоста. При этом все законы подбирались с учетом равенства остальных параметров, в частности каждый из подобранных законов распределения является симметричным относительно центра. Предлагаемый новый метод отношений сравнивается с широко известной и применяемой на практике оценкой Хилла. Доказано, что в вопросе идентификации закона распределения оценка Хилла дает, в целом, положительный результат. При этом показано, что наиболее оптимальным является применение метода отношений с уровнем 5%.

Ключевые слова: тяжесть хвоста, оценка Хилла, метод отношений, закон распределения вероятности, идентификация

COMPARISON OF SOME METHODS FOR ASSESSING THE WEIGHT OF TAILS IN IDENTIFICATION OF THE DISTRIBUTION LAW

Akimov S.S., Tripkosh V.A.

Orenburg State University, Orenburg, e-mail: sergey_akimov_work@mail.ru

The article deals with the problem of identification of the distribution law. It is shown that the distribution law is an exhaustive characteristic of a random data array. In this case, in the absence of a priori knowledge of the distribution law, the processing of the data array available to the researcher presents certain difficulties, the solution of which is possible by identifying the distribution law. The article provides a brief overview of the main methods for identifying distribution laws, while it is noted that characteristics calculated directly from the original data array can help in identifying the distribution law. One of these characteristics is the severity of the tail of the probability distribution. It is shown that there is no single universal method for assessing the severity of the tail, and, in particular, for the purpose of identifying the distribution law. The article discusses the Hill assessment method, but it is shown that the method itself is quite laborious when performing calculations. Based on this, a new relationship method is proposed based on the ratio of the main and tail parts of the distribution. At the same time, the question of determining the boundaries of the main and tail parts of the distribution according to the original data array is solved. Five different distribution laws were selected as the basis for the study, differing in the weight of the tail. At the same time, all the laws were selected taking into account the equality of the remaining parameters, in particular, each of the selected distribution laws is symmetric about the center. The proposed new method of relations is compared with the widely known and applied in practice Hill's estimate. It is proved that in the question of identifying the distribution law, Hill's estimate gives, in general, a positive result. At the same time, it was shown that the most optimal is the application of the method of relations with a level of 5%.

Keywords: tail severity, Hill's estimate, relationship method, probability distribution, identification

В настоящее время проблемам обработки данных в массивах отводится значительное место во многих прикладных науках. Данное обстоятельство объясняется тем, что от правильной обработки набора данных напрямую зависит конечный результат, и, в случае если изначально данные были интерпретированы и обработаны неверно, исследователь может получить в итоге значительно искаженные данные, что сведет на нет все его усилия.

Как известно, исчерпывающей характеристикой массива случайных данных в различных исследованиях является закон распределения, которому подчиняется данный массив [1]. Однако в прикладных исследованиях закон распределения, как правило, бывает неизвестен или же не определен однозначно [2].

Проблеме идентификации закона распределения с опорой исключительно на массив данных посвящено достаточно

большое количество трудов отечественных и зарубежных ученых и отведена большая роль в области теории вероятностей и математической статистики [3].

В рамках поставленной задачи разработано огромное количество различных методов и процедур, среди которых можно выделить следующие: метод гистограмм, приближение сплайнами, корневая оценка плотности, рекуррентная ядерная оценка, интегральная оценка, стохастическая регуляризация, метод проекций, структурная минимизация риска, оценка максимума правдоподобия и т.д. При этом, несмотря на все многообразие подходов к решению задачи идентификации закона распределения, абсолютное большинство методов идентификации закона распределения базируется на его основных особенностях: асимметрии, эксцессе и т.д. [4].

Немаловажным в данном списке является метод оценки тяжести хвостов распределения. При этом широко известно, что значительное количество процедур идентификации законов распределения могут быть достоверно применимы только для распределений, имеющих легкие хвосты [5]. Аналитика массивов данных, которые подчиняются распределениям с тяжелыми хвостами, должна проводиться неклассическими статистическими методами, поскольку имеет место нарушение основного условия Крамера, описывающего производящую функцию моментов [1].

В настоящий момент не существует единого стандарта оценки тяжести хвоста. При этом чаще других используется оценка Хилла [5]. Однако никаких обоснований предпочтений данной оценки, как правило, не приводится. Потому применение данной оценки нуждается в достоверной проверке на различных законах распределения.

Цель исследования: проверить адекватность оценки Хилла для определения тяжести хвоста в различных законах распределения.

Задачи исследования:

– подобрать законы распределения с разной степенью тяжести хвостов;

– оценить тяжесть хвоста при помощи различных методов оценок;

– провести сравнение полученных результатов.

Материалы и методы исследования

Исследования выполнены в лаборатории кафедры управления и информатики в технических системах Оренбургского государственного университета.

В качестве распределений были взяты следующие: равномерное непрерывное; нормальное; Стьюдента; логистическое; Коши. Данные законы являются очень распространенными, а кроме того, обладают сильно различающимися кривыми распределения, поведение которых, безусловно, влияет на конечный результат.

Подбор распределений осуществлялся исходя из тяжести их хвостов; при этом для чистоты эксперимента предпочтения отдавались распределениям, не имеющим асимметрии.

Для определения количества каждого из видов законов распределений в данном исследовании использовалась распространенная формула, взятая из работы В.Н. Дианова «Перспективные направления повышения надежности вычислительной техники и систем управления» [6]:

$$k_p = \frac{4}{p(1-p)\varepsilon^2}, \quad (1)$$

где P – вероятность события, ε – допустимая погрешность вычислений.

Предпочтение данному способу отдано в силу заведомой ненормальности большинства распределений.

Вероятность, как правило, задается на уровне 0,95. Погрешность для данной проверки берется на уровне 0,5, поэтому общее число каждого исследуемого распределения составляло 337.

Для проведения эксперимента на генераторе случайных чисел программы Mathcad 15 были сгенерированы три массива данных (n = 1000), подчиняющихся равномерному (*U*), нормальному (*N*), логистическому (*log*) закону распределения, а также распределению Стьюдента (*St*) и Коши (*C*), с такими параметрами, чтобы размах вариации для каждого был примерно одинаков.

Оценка Хилла для исследования тяжести хвоста определялась по следующей формуле:

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \ln \frac{x_{(i)}}{x_{(k-1)}}, k < n, \quad (2)$$

Поскольку данная формула, как приведено в некоторых исследованиях [7], не дает численного результата, то используем уже известную систему преобразований, описанную в работе «Модели оценки параметров телекоммуникационного трафика в автоматизированных информационно-управляющих системах» [8].

Искомая оценка находилась для каждой из квантилей, по результатам чего были построены квантильно-квантильные графики [9]. Затем строилась касательная к полученному графику и вычислялись уравнения регрессии с основой в виде полинома второй

степени. Далее через тангенсы угла касательной и оси абсцисс вычислялся искомый параметр, являющейся оценкой тяжести [10].

Данный метод является наиболее универсальным, он описан в большинстве литературных источников и применяется на практике чаще других [11]. Вместе с тем в отечественных и зарубежных трудах нет никаких исследований, отображающих точность и достоверность данного подхода [12]. Однако необходимо отметить, что указанный метод Хилла далеко не единственный, который может быть применен для сравнения хвостовой части изучаемых распределений [13].

В качестве альтернативной оценки использовался метод отношений. Он основывается на том, что с увеличением тяжести хвоста его значения начинают значительно отличаться от среднего. Для реализации этого метода из ранжированного массива данных получают интервальный ряд, который тоже ранжируют. Далее выбирают хвостовую часть, которая составляет, как правило, от 1% до 20%.

Результаты исследования и их обсуждение

Для реализации метода отношений хвостовая часть бралась на уровнях 1%, 3%, 5%, 10%, 15% и 20%.

Результаты проверки приведены в табл. 1.

Как видно из представленной таблицы, оценки, полученные различными способами, различаются между собой. Для того чтобы однозначно отделить один закон распределения от другого, необходимо, чтобы интервалы оценок не пересекались, иначе говоря, чтобы максимум оценки одного распределения был меньше минимума другого распределения. Для выявления различий в оценках построим таблицы разностей для каждого из рассматриваемых законов распределений (табл. 2–5).

Как видно из таблицы, ни один из методов не помогает отличить равномерное непрерывное распределение от нормального распределения. Равномерное распределение от распределения Стьюдента можно отличить при помощи оценки Хилла или метода отношений с уровнем 5%. Равномерное распределение от логистического распределения позволяют отличить методы отношений с уровнями 3%, 5%, 10%, 15%. Наконец, равномерное непрерывное распределение от распределения Коши можно отличить при помощи любого из перечисленных методов.

Согласно данным таблицы, нормальное распределение можно отличить только от распределения Коши при помощи метода отношений с уровнями 1%, 3%, 5%, а также при помощи оценки Хилла. Остальные описанные случаи не дают возможности различения перечисленных распределений.

Таблица 1

Min и *max* значения оценок тяжести хостов разными методами

Метод	Законы распределения									
	Равномерный непрерывный		Нормальный		Стьюдента		Логистический		Коши	
	Min	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*
Отношений, 1%	4,57	7,46	7,28	25,34	7,16	24,91	7,39	25,26	107,8	198568
Отношений, 3%	4,15	6,77	6,11	21,25	6,59	21,07	7,04	22,17	69,31	63427
Отношений, 5%	3,77	6,15	5,98	20,8	6,28	18,37	6,43	19,8	21,84	4505
Отношений, 10%	3,47	5,66	5,64	22,66	5,59	16,49	6,01	19,78	19,37	1157
Отношений, 15%	3,40	5,55	5,51	23,03	5,23	19,35	5,59	22,63	16,44	118,21
Отношений, 20%	3,24	5,28	5,02	23,09	5,14	22,43	5,13	22,79	14,26	73,45
Оценка Хилла	0,12	0,33	0,3	0,37	0,35	0,593	0,32	0,47	1,05	3,12

Таблица 2

Разности максимума оценки тяжести хвоста равномерного непрерывного распределения с минимумами других распределений

Метод	Законы распределения			
	Нормальный	Стьюдента	Логистический	Коши
Отношений, 1%	0,17	0,30	0,07	**–100,34**
Отношений, 3%	0,66	0,18	**–0,27**	**–62,54**
Отношений, 5%	0,17	**–0,13**	**–0,28**	**–15,69**
Отношений, 10%	0,02	0,07	**–0,35**	**–13,71**
Отношений, 15%	0,04	0,32	**–0,04**	**–10,89**
Отношений, 20%	0,26	0,14	0,15	**–8,98**
Оценка Хилла	0,03	**–0,02**	0,01	**–0,72**

Таблица 3

Разности максимума оценки тяжести хвоста нормального распределения с минимумами других распределений

Метод	Законы распределения		
	Стьюдента	Логистический	Коши
Отношений, 1%	18,18	17,95	**–82,46**
Отношений, 3%	14,66	14,21	**–48,06**
Отношений, 5%	14,52	14,37	**–1,04**
Отношений, 10%	17,07	16,65	3,29
Отношений, 15%	17,80	17,44	6,59
Отношений, 20%	17,95	17,96	8,83
Оценка Хилла	0,02	0,05	**–0,68**

Таблица 4

Разности максимума оценки тяжести хвоста распределения Стьюдента с минимумами других распределений

Метод	Законы распределения	
	Логистический	Коши
Отношений, 1%	17,52	**–82,89**
Отношений, 3%	14,03	–48,24
Отношений, 5%	11,94	**–3,47**
Отношений, 10%	10,48	**–2,88**
Отношений, 15%	13,76	2,91
Отношений, 20%	17,30	8,17
Оценка Хилла	0,273	**–0,457**

Как показывают данные таблицы, распределение Стьюдента также отличается только от распределения Коши при помощи метода отношений с уровнями 1%, 3%, 5%, 10% или оценки Хилла; все перечисленные виды оценки не помогают отличить распределение Стьюдента от логистического распределения.

Таблица 5

Разности максимума оценки тяжести хвоста логистического распределения с минимумами других распределений

Метод	Распределение Коши
Отношений, 1%	**–82,54**
Отношений, 3%	**–47,14**
Отношений, 5%	**–2,04**
Отношений, 10%	0,41
Отношений, 15%	6,19
Отношений, 20%	8,53
Оценка Хилла	**–0,58**

Согласно представленным в таблице данным, логистическое распределение можно отличить от распределения Коши при помощи методов отношений с уровнями 1%, 3%, 5% или оценки Хилла.

Из приведенных выше данных очевидно, что наиболее приемлемым является метод отношений с уровнем 5%, который позволяет различать наибольшее количество законов распределения из представленного перечня. Данный метод более оптимален, чем оценка Хилла, поскольку с большей точностью позволяет отличать логистическое распределение от равномерного непрерывного распределения.

Заключение

В работе представлен способ идентификации некоторых законов распределения на основе оценки тяжести их хвостов. Выбраны пять законов распределений с различной степенью тяжести хвоста, также выбран ряд оценок, включающий в себя как широко применяемую оценку (оценку Хилла), так и более оригинальные оценки (метод отношений с различным уровнем).

Проведены попарные сравнения законов распределения в зависимости от тяжести хвоста каждым из перечисленных методов, все полученные оценки приведены в соответствующих таблицах.

Полученные разными способами оценки тяжести хвоста сравнивались при помощи метода разниц. Данным способом определено, что наилучший результат в проблеме идентификации закона распределения показывает метод отношений с уровнем 5%.

Необходимо также отметить, что ни один из методов не выдал гарантированного результата отделения одного закона распределения от другого. Соответственно, для объективизации оценки необходима комбинация указанного метода с каким-либо другим с целью получения более точной и объективной картины распределения исследуемых массивов данных.

Список литературы

1. Акимов С.С. Оптимизированный алгоритм определения закона распределения вероятности по выборке из генеральной совокупности // Известия Самарской государственной сельскохозяйственной академии. 2013. № 2. С. 52–56.

2. Сызранцев В.Н., Невелев Я.П., Голофаст С.Л. Адаптивные методы восстановления функции плотности распределения вероятности // Известия ВУЗов. Машиностроение. 2006. № 12. С. 3–11.

3. Беврани Х., Аничкин К. Оценка параметров распределений с тяжелыми хвостами с помощью эмпирического распределения // МКО. 2005. Ч. 2. С. 493–495.

4. Акимов С.С. Использование коэффициентов асимметрии и эксцесса при гистограммном методе определения закона распределения вероятности // Известия Оренбургского государственного аграрного университета. 2014. № 1 (45). С. 225–227.

5. Шепель В.Н., Акимов С.С. Использование оценки Хилла для различения законов распределения вероятности // Вестник Оренбургского Государственного университета. 2014. № 1 (162). С. 75–78.

6. Дианов В.Н. Перспективные направления повышения надежности вычислительной техники и систем управления // Надежность. 2004. № 3(10). С. 33–47.

7. Глебов В.И., Криволапов С.Я. О принадлежности к области притяжения устойчивых законов распределений, обобщающих распределение Коши // Успехи современной науки. 2016. Т. 6. № 11. С. 32–35.

8. Гуда А.Н., Бутакова М.А., Москат Н.А. Модели оценки параметров телекоммуникационного трафика в автоматизированных информационно-управляющих системах // Вопросы современной науки и практики. Ун-т им. В.И. Вернадского. 2010. № 4–6(29). С. 71–87.

9. Волынская А.В. О двух методах приближенного построения оператора преобразования законов распределения случайных процессов // Фундаментальные исследования. 2015. № 2–7. С. 1383–1386.

10. Капля Е.В. Параметрическое обобщение логистического закона распределения в статистическом анализе динамики направления ветра // Альтернативная энергетика и экология. 2015. № 17–18. С. 42–47.

11. Буторина О.В., Осипова М.Ю. Особенности статистического анализа современного производственного цикла // Вектор науки Тольяттинского государственного университета. Серия: Экономика и управление. 2018. № 1(32). С. 5–12.

12. Галкин В.М., Ерофеева Л.Н., Лещева С.В. Оценки параметра распределения Коши // Труды НГТУ им. Р.Е. Алексеева. 2014. № 2. С. 314–319.

13. Сиротин В.П., Архипова М.Ю. Декомпозиция распределений в моделировании социально-экономических процессов: монография. М.: МЭСИ, 2011. 146 с.