

МЕТОДЫ РАЗЛИЧЕНИЯ СИММЕТРИЧНЫХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ (НА ПРИМЕРЕ НОРМАЛЬНОГО И ЛОГИСТИЧЕСКОГО ЗАКОНА РАСПРЕДЕЛЕНИЯ)

Одна из основных задач математической статистики – это всестороннее изучение и анализ данных. При этом широко известно, что исчерпывающей характеристикой любого массива данных является закон распределения, которому данный массив подчиняется. Существует достаточно большое количество различных методов определения закона распределения, опираясь исключительно на массив данных. Вместе с тем необходимо отметить, что все эти методы носят лишь приближенный характер. При этом стоит отметить, что некоторые законы распределения имеют достаточно сходные характеристики, что затрудняет процесс их идентификации. Одной из таких характеристик является симметрия.

Основная цель исследования – сравнение методов идентификации для симметричных законов распределения. В статье рассматриваются четыре известных метода идентификации закона распределения вероятности: критерий согласия Колмогорова, коэффициент эксцесса, оценка Хилла и метод отношений. В качестве законов распределения использовались логистическое и нормальное распределение, наиболее близкие симметричные законы. Исследования проведены в лаборатории кафедры управления и информатики в технических системах Оренбургского государственного университета. Для получения необходимых массивов данных использовался генератор случайных чисел программы Mathcad 15. Для проверки было сгенерировано 800 распределений (400 подчиняющихся логистическому закону и 400 – нормальному; $N = 100$). Для достоверности оценки параметры распределений оставались неизменными.

Было определено, что наиболее сильным методом является критерий Колмогорова. Кроме того изучались различные сочетания предлагаемых методов. При исследовании сочетаний выявлено, что наиболее результативным является применение пары критериев согласия Колмогорова – метода отношений. Все остальные пары методов дают больший процент ошибок.

Ключевые слова: логистическое распределение, нормальное распределение, методы различения, комбинация методов.

Одна из основных задач математической статистики – это всестороннее изучение и анализ данных. При этом широко известно, что исчерпывающей характеристикой любого массива данных является закон распределения, которому данный массив подчиняется.

Вопрос идентификации закона распределения по массиву данных изучается уже достаточно давно. При этом стоит отметить, что, несмотря на огромное количество методов, исследователям достаточно редко удается получить состоятельную оценку данных с точки зрения их принадлежности какому-либо конкретному закону распределения [1].

Отметим, что для восстановления неизвестной функции плотности распределения в рамках непараметрической статистики разработан ряд методов и алгоритмов [2]. Литературный анализ данной проблемы [3] показывает, что на практике достаточно хорошо разработана процедура различения симметричных и асимметричных законов распределения [4], [5], [6], [7].

При реализации большинства известных методов, основанных на асимметрии данных, в

случае доказательства принадлежности массива к симметричному закону распределения, делается вывод о нормальности данных, что ошибочно. Несостоятельность данного подхода отмечают и другие авторы [8], [9], [10].

Причина в том, что в природе существует целый ряд симметричных законов распределения, которым может подчиняться исследуемый массив. Поэтому проблема определения принадлежности массива данных к какому-либо из симметричных законов распределения была и остается весьма актуальной.

Считается, что наиболее схожими симметричными распределениями являются нормальное и логистическое [8]. Сама процедура их различения между собой весьма затруднена. При этом существует ряд методик, помогающих идентифицировать закон распределения. Потому актуальной становится задача определения наиболее подходящей методики для различения симметричных законов распределения.

Цель исследования: сравнить методы идентификации для симметричных законов распределения.

Задачи исследования:

- рассмотреть методы идентификации симметричных законов распределения;
- провести сравнение эффективности данных методов на нормально и логистически распределенных массивах данных;
- выполнить различные комбинации перечисленных методов, выбрать их наиболее оптимальное сочетание.

Материалы и методы

Исследования проведены в лаборатории кафедры управления и информатики в технических системах Оренбургского государственного университета. Для получения необходимых массивов данных использовался генератор случайных чисел программы Mathcad 15. В качестве методов идентификации были использованы следующие критерии.

1) Критерий согласия Колмогорова для проверки нормальности, рассчитанный по формуле [11]

$$\lambda = \frac{6nD_n + 1}{6\sqrt{n}}. \quad (1)$$

2) Коэффициент эксцесса в модификации Е.И. Пустыльника, исчисляемый по формуле [12]

$$Ex_{эм} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n\sigma^4} - 3, \quad (2)$$

при этом критические значения коэффициента эксцесса подчиняются формуле:

$$Ex_{кр} = 5\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}. \quad (3)$$

3) Оценка Хилла с модификацией А.Н. Гуда, определяемая по формуле [13], [14]

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \ln \frac{x_{(i)}}{x_{(k-i)}}, k < n. \quad (4)$$

4) Метод отношений, в основе которого используется формула [9]

$$K = \frac{R - Ex}{\sigma}. \quad (5)$$

Для проверки было сгенерировано 800 распределений (400 подчиняющихся логистическому закону и 400 – нормальному; $N = 100$). Для

достоверности оценки параметр распределений оставались неизменными.

Результаты и обсуждение

Сгенерированные массивы данных проверялись на соответствие нормальному закону распределения. В качестве параметра сравнения использовалось количество ошибок первого (в случае, если гипотеза о нормальности была отвергнута на нормально распределенном массиве данных) и второго рода (в случае, если гипотеза о нормальности была принята на логистически распределенном массиве данных), полученных при сравнении гипотез между собой. Полученный результат сведен в таблицу 1.

В результате проведенного сравнения различных методов идентификации можно сделать следующие выводы:

- наиболее точным среди перечисленных методов является критерий согласия Колмогорова;
- наихудшим методом среди перечисленных, является коэффициент эксцесса в модификации Е.И. Пустыльника;
- метод отношений, в среднем, дает меньше ошибок при проверке нормальных распределений, но больше при проверке логистических распределений.

Таким образом, наиболее информативным является критерий согласия Колмогорова.

Рассмотрим применение попарных комбинаций, для большей достоверности полученного результата. Для этого сопоставим пары из перечисленных методов. Все возможные комбинации составляют всего шесть пар:

- критерий согласия Колмогорова – коэффициент эксцесса;
- критерий согласия Колмогорова – оценка Хилла;
- критерий согласия Колмогорова – метод отношений;
- коэффициент эксцесса – оценка Хилла;
- коэффициент эксцесса – метод отношений;
- оценка Хилла – метод отношений.

Результат рассчитывался следующим образом: если хотя бы в одном случае основная гипотеза отвергалась, то, вне зависимости от результата второй проверки, альтернативная гипотеза была принята.

Все полученные результаты сведем в таблицу 2.

Таблица 1. Количество ошибочно принятых гипотез различными методами

Метод	Количество ошибок (%)		Сумма
	Нормальное распределение (ошибка первого рода)	Логистическое распределение (ошибка второго рода)	
Критерий согласия Колмогорова	0,25	21,25	21,50
Коэффициент эксцесса	11,00	39,75	50,75
Оценка Хилла	1,75	24,50	26,25
Метод отношений	0,75	28,75	29,50

Таблица 2. Количество ошибочно принятых гипотез при попарном применении методов

Метод	Количество ошибок (%)		Сумма
	Нормальное распределение (ошибка первого рода)	Логистическое распределение (ошибка второго рода)	
Критерий согласия Колмогорова - коэффициент эксцесса	0,25	21,25	21,50
Критерий согласия Колмогорова - оценка Хилла	0,25	19,25	19,50
Критерий согласия Колмогорова - метод отношений	0,25	16,75	17,00
Коэффициент эксцесса - оценка Хилла	1,75	24,50	26,25
Коэффициент эксцесса - метод отношений	0,75	28,75	29,50
Оценка Хилла - метод отношений	0,75	23,50	24,25

Таким образом, наиболее результативным является применение пары критерий согласия Колмогорова – метод отношений. Все остальные пары методов дают больший процент ошибок.

Отметим, что в рамках данного исследования проводились сопоставление различных вариаций трех и четырех методов. Однако ни одно из перечисленных сопоставлений не дало существенного снижения количества ошибок. Исходя из этого, можно сделать вывод о том, что применять большее количество методов с целью получения лучшего результата – нецелесообразно.

Заключение

В работе описаны некоторые методы различения симметричных законов распределения

(на примере логистического и нормального распределения). Выбраны четыре широко распространенных метода: критерий согласия Колмогорова, коэффициент эксцесса, оценка Хилла и метод отношений.

Проведенное сравнение установило, что наиболее результативным методом является критерий согласия Колмогорова.

При изучении попарных комбинаций методов выявлено, что наилучшей комбинацией является сочетание критерия согласия Колмогорова и метода отношений, в результате применения которой среднее количество ошибок при принятии гипотезы о принадлежности массива конкретному закону распределения составляет 17%.

08.06.2015

Список литературы:

1. Шепель В.Н., Акимов С.С. Определение плотности вероятности данных в прикладных исследованиях. – Современные информационные технологии в науке, образовании и практике. – Материалы XI Всероссийской научно-практической конференции. – Оренбургский государственный университет. – Оренбург. – 2014. С. 223-226.
2. Айвазян С. А., Мхитарян В. С. Прикладная статистика. Основы эконометрики (в 2-х т.) Теория вероятностей и прикладная статистика – М.: Юнити-Дана, 2007, -656 с.
3. Акимов С.С. Оценка методов восстановления законов распределения вероятности и обоснование предпочтения на основании некоторых свойств эмпирического массива данных. – Актуальные вопросы современной науки. – 2014. – № 31. С. 139-150.

4. Акимов С.С. Методы решения задачи восстановления плотности вероятности по выборке из генеральной совокупности. – Естественные и математические науки в современном мире: сб. ст. по материалам XIV междунар. науч.-практ. конф. №1 (13). Новосибирск: Изд. «СибАК», 2014. С. 29-35.
5. Айвазян С.А. Различение близких гипотез о виде плотности распределения в схеме обобщенного последовательного критерия. – Теория вероятностей и ее применения, 1965, т. X, № 4.
6. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. - М.: Наука, 1979.
7. D'Agostino R. B., Pearson E. S. A further development of test departure from normality. — *Biometrika*, 1973, 60, №3 — p. 613-622.
8. Орлов А. И. Типовые ошибки при вхождении в прикладную статистику. – [Электронный ресурс]. – Режим доступа: URL: <http://forum.orlovs.pp.ru/viewtopic.php?t=97> (заглавие с экрана).
9. Дубнер П.Н. Вычисление прямых и обратных функций распределения. // *Статистика и стохастические системы*, вып. 15, Изд-во МГУ, 1971.
10. Лемешко Б. Ю., Лемешко С. Б. Сравнительный анализ критериев проверки отклонения распределения от нормального закона // *Метрология*. 2005. № 2. С. 3–24.
11. Большев Л.Н. Смирнов Н.В. Таблицы математической статистики. М.: Наука.1983.
12. Пустыльник Е.И. Статистические методы анализа и обработки наблюдений [Текст] / Е.И. Пустыльник – «Наука», М, 1968. – 288 с.
13. 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.
14. Шепель В.Н. Использование оценки Хилла для различения законов распределения вероятности [Текст] / Шепель В. Н., Акимов С. С. // *Вестник Оренбургского государственного университета*. - 2014. - № 1, январь. - С. 75-78.
15. Акимов С.С. Проблема выбора метода восстановления закона распределения вероятности. – *Международный научно-исследовательский журнал*. – 2014. – № 1-3 (20). С. 5-8.

Сведения об авторах:

Шепель Вячеслав Николаевич, профессор кафедры управления и информатики в технических системах Оренбургского государственного университета, доктор экономических наук
460018, г. Оренбург, Шарлыкское шоссе, 5, e-mail: vn_shepel@mail.ru

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: elite17@yandex.ru