УДК 519.216+519.224

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

МОДЕРНИЗАЦИЯ МЕТОДА ГИСТОГРАММ ДЛЯ ВЫЯВЛЕНИЯ ПРИНАДЛЕЖНОСТИ НЕИЗВЕСТНОГО МАССИВА ДАННЫХ ОПРЕДЕЛЕННОМУ ЗАКОНУ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

Рассматривается проблема совершенствования метода гистограмм для восстановления плотности распределения вероятности по выборке. Разработаны коэффициенты оценки и рассчитаны их критические значения.

Ключевые слова: гистограммный метод, коэффициенты, критические значения, закон распределения вероятности.

В задаче определения распределения вероятности по выборке возникает необходимость принятия гипотезы о его виде [1].

Для восстановления неизвестной функции плотности распределения в рамках непараметрической статистики разработан ряд методов и алгоритмов [2].

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 [3].

В то же время практика решения технических задач свидетельствует, что в подавляющем большинстве случаев для восстановления функции плотности используется метод гистограмм [4].

Гистограммному методу оценки плотности распределения вероятности посвящены труды многих исследователей [5], [6], [7], [8]. На сегодняшний день он является одним из самых популярных методов.

Однако гистограммный метод оценки является экспертным методом, что существенно ограничивает возможности его применения. В частности, необходимость экспертного оценивания исключает возможность применения ЭВМ для получения окончательных результатов. Кроме того, экспертное заключение обладает существенной долей субъективности.

Для снижения субъективности исследования, а также для получения возможности компьютерного анализа гистограммы, необходимо алгоритмизировать процесс оценки гистограммы. Для этого необходимо разработать коэффициенты, расчет которых мог бы производиться автоматически.

Кроме того, необходимо определить критические значения коэффициентов для возможности отнесения исследуемого массива данных к тому или другому закону распределения вероятности.

Цель исследования

Разработать систему коэффициентов для усовершенствования гистограммного метода определения распределения вероятностей по выборке.

Задачи исследования

– разработать коэффициенты, с учетом основных характеристик гистограммы, с целью более точного восстановления плотности распределения вероятности;

– рассчитать критические значения для полученных коэффициентов;

– произвести расчеты полученных коэффициентов на массивах данных с заранее известными законами распределения, с целью проверки адекватности модели.

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. Количество, состав, размерность и другие числовые характеристики рассматриваемых массивов данных были определены при помощи метода Монте-Карло, подробно описанном в [9]. На

описанных массивах данных производилась оценка гистограммным методом.

Результаты и обсуждение

Для наиболее точного восстановления закона вероятности разработано два коэффициента.

1. Коэффициент соотношения долей гистограммы. Рассчитывается как отношение сумм значений интервалов первой половины гистограммы и второй.

Замечание 1. Если количество интервалов в гистограмме нечетное, то интервал, находящийся в середине, необходимо разделить на два и прибавить полученный результат к правой и левой частям.

Замечание 2. Для наглядности и удобства, в случае, когда количество данных правой половины гистограммы больше левой, полученный коэффициент записывается в знаменатель дроби, числителем которой выступает -1.

Замечание 3. Для того чтобы результат коэффициента был удобен для восприятия, но не искажал бы данных, результат необходимо разделить на *n/10*.

2. Коэффициент убывания данных гистограммы. Рассчитывается как стандартное отклонение от количества значений, попавших в различные интервалы гистограммы.

Замечание 1. Для того, чтобы результат коэффициента был удобен для восприятия, но не искажал бы данных, также как и в предыдущем случае, результат необходимо разделить на *n/10*.

В результате построения гистограмм и применения метода Монте-Карло для различных распределений выявлены следующие критические значения для предложенных выше коэффициентов (таблица 1 и 2).

Как видно из таблиц, часть интервалов перекрывается. При использовании коэффициента соотношения долей гистограммы однозначно можно отделить:

– равномерное непрерывное распределение от распределения Рэлея;

– равномерное непрерывное распределение от гамма-распределения;

– равномерное непрерывное распределение от экспоненциального распределения;

– нормальное распределение от распределения Рэлея;

– нормальное распределение от экспоненциального распределения;

– логистическое распределение от распределения Рэлея;

– логистическое распределение от экспоненциального распределения.

Остальные распределения определяются данным коэффициентом неоднозначно.

При использовании коэффициента убывания данных гистограммы однозначно можно отделить:

– равномерное непрерывное распределение от всех остальных распределений;

– нормальное распределение от распределения Рэлея;

– нормальное распределение от распределения Коши;

– нормальное распределение от экспоненциального распределения;

– логистическое распределение от распределения Коши;

Таблица 1. Критические значения для коэффициента соотношения долей гистограммы

Распределение	Эмпирический интервал значений
Равномерное непрерывное	[-0,18;0,18]
Нормальное	[-0,2;0,2]
Рэлея	[0,25;1]
Логистическое	[-0,9;0,2]
Гамма	[0,2;4]
Экспоненциальное	[3;9,99]
Логнормальное	[-0,2;9,99]
Вейбулла	[-0,4;9,99]
Бета	[-0,4;9,99]
Коши	[-9,99;9,99]

Таблица 2. Критические значения для коэффициента убывания данных гистограммы

Распределение	Эмпирический интервал значений
Равномерное непрерывное	[0,15;0,45]
Нормальное	[0,5;1]
Рэлея	[1,1;1,3]
Логистическое	[0,8;1,5]
Коши	[1,8;3,6]
Экспоненциальное	[3;3,6]
Гамма	[0,7;3,2]
Логнормальное	[0,5;3,6]
Вейбулла	[0,5;3,6]
Бета	[0,5;3,6]

– логистическое распределение от экспоненциального распределения.

Все остальные случаи распределений являются неотличимыми данным методом.

Справедливость полученных значений подтверждается, также, работами авторов Ж.В. Дейнеко [10], М. А. Маталыцкого [11], Б.Ю. Лемешко [12]. В работе [10] при оценке логнормального распределения получено критическое значение 2,411. В работе [11] рассмотрено экспоненциальное распределение, критическое значение для которого оказалось равным 3,327. В работе [12] при исследовании логистического распределение было получено значение 1,42. Все полученные данные соответствуют рассчитанным интервалам, полученным в данном исследовании.

Заключение

В работе описаны коэффициенты, связанные с основными характеристиками гистограммы плотности распределения, использование которых увеличивает достоверность гистограммного метода и дает возможность алгоритмизировать сам метод.

Рассчитаны критические значения предлагаемых коэффициентов для наиболее часто встречающихся непрерывных распределений вероятности, с целью однозначного отделения внешне похожих гистограмм распределений.

Выявленные значения прошли проверку на ряде распределений, сгенерированных с помощью программы Mathcad 15. Для повышения достоверности использовались, также, данные, приведенные в статьях других исследователей.

26.06.2014

Список литературы:

1. 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.
2. Айвазян, С.А. Прикладная статистика. Основы эконометрики (в 2 -х т.) Теория вероятностей и прикладная статистика / С.А. Айвазян, В.С. Мхитарян. – М.: Юнити-Дана, 2007. – 656 с.
3. 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.
4. Сызранцев, В.Н. Адаптивные методы восстановления функции плотности распределения вероятности / В.Н. Сызранцев, Я.П. Невелев, С.Л. Голофаст // Известия ВУЗов. Машиностроение. – 2006. – №12. – С. 3–11.
5. Шепель, В.Н. Алгоритм определения эмпирической функции плотности по выборке из генеральной совокупности / В.Н. Шепель // Современные информационные технологии в науке и практике. Материалы VIII всероссийской научно-практической конференции (с международным участием). – Оренбург: ИПК ГОУ ОГУ. – 2009. – С. 224–226.
6. Маркович, Н.М. Методы оценивания характеристик тяжело-хвостовых случайных величин по конечным выборкам / Н.М. Маркович : Автореф. дис. ... д-ра физ.-мат. наук : 05.13.01 : М., 2004 – 206 с.
7. Катковник, В.Я. Непараметрическая идентификация и сглаживание данных / В.Я. Катковник. – М.: Главная редакция физико-математической литературы. – 1985. – 336 с.
8. Алейник, С.В. Метод оценки уровня клиппирования речевого сигнала / С.В. Алейник, Ю.Н. Матвеев, А.Н. Раев // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – № 3 (79). – С. 79–82.
9. Вентцель, Е.С. Исследование операций / Е.С. Вентцель. – М., «Советское радио». – 1972. – 552 с.
10. 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.
11. Маталыцкий, М.А. Метод анализа средних значений для сетей массового обслуживания, рекуррентный по моментам времени / М.А. Маталыцкий // Автомат. и телемех. – 1999. – № 11. – С. 39–45.
12. Лемешко, Б.Ю. Система статистического анализа наблюдений и исследования статистических закономерностей / Б.Ю. Лемешко, С.Н. Постовалов // Сб. «Моделирование, автоматизация и оптимизация наукоемких технологий». – Новосибирск: изд-во НГТУ, 2000. – С. 44–46.

Сведения об авторах:

Шепель Вячеслав Николаевич, профессор кафедры управления и информатики в технических системах Оренбургского государственного университета, доктор экономических наук, 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