УДК 519.245

СГЛАЖИВАНИЕ ГИСТОГРАММ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ МЕТОДОМ МОНТЕ-КАРЛО

А.Н. ЛЫСЮК, канд. техн. наук, доц. С.С. ДЕРЕЧЕННИК
(Брестский государственный технический университет)

Рассмотрена задача формирования эмпирического распределения для соотношения двух дискретных величин с переменным знаменателем. Перечислены основные проблемы, возникающие при использовании классических методик. Показан фрактальный эффект, присутствующий на итоговой гистограмме, указаны причины его возникновения. Сформулирован теоретический алгоритм решения этой проблемы, предложена версия его реализации для вычислительной машины. Основная идея предлагаемого подхода заключается в генерировании методом Монте-Карло множества значений, по которым строится итоговая гистограмма, устраняющая все перечисленные недостатки классических методик. Важным фактором является то, что генерирование значений осуществляется в рамках некоторого диапазона, границы которого рассчитываются на первом шаге алгоритма. В результате экспериментального сравнения классической и предлагаемой методики показано преимущество последней. В качестве дополнения предложен способ оценивания параметров получаемых эмпирических распределений, основанный на статистике интервальных данных.

Введение. В настоящее время в Брестском государственном техническом университете проводится ряд исследований, направленных на совершенствование методики расчета учебной нагрузки кафедр и эффективное перераспределение трудового (преподавательского) ресурса [1]. Так как главным фактором, влияющим на объем учебной нагрузки, является количество обучающихся студентов, возникает частная задача прогнозирования данного показателя, который из-за наличия событий отчисления/восстановления, является случайной величиной.

При исследовании процессов отчисления/восстановления в качестве анализируемой величины было принято соотношение y_i количества отчисленных/восстановленных студентов на академическом потоке к общему количеству студентов:

$$y_i = \frac{m_i}{n_i}, \quad i = 1,2,...,L, \tag{1}$$

где m_i – количество отчисленных/восстановленных студентов; n_i – общее количество студентов; L – общее количество анализируемых академических потоков.

Однако в процессе исследования возникла следующая проблема: в связи с тем, что академические потоки на различных специальностях и курсах неоднородны по численному составу, то для анализа данных неприменимы классические методы математической статистики [2]. На рисунке 1 показан результат применения общепринятой методологии, основанной на построении гистограмм.

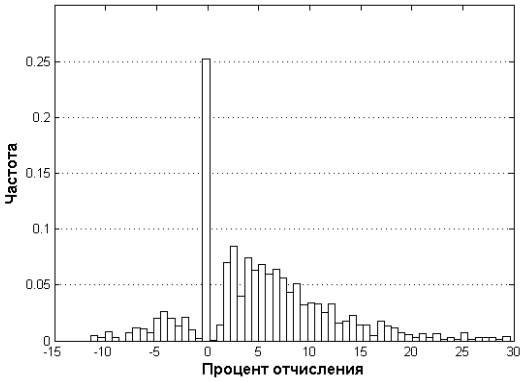


Рис. 1. Гистограмма процента отчислений

По данной гистограмме сложно делать предположения относительно истинной природы искомого распределения. Так, видно, что для нулевого значения соответствующая частота максимальна, тогда как для значений, находящихся рядом с нулевым, частота имеет локальные минимумы. Это обусловлено наличием фрактального эффекта у подобных распределений, как показано далее.

Использование методов непараметрической статистики, в частности ядерного сглаживания [3], также не позволяет избавиться от указанного эффекта, поскольку данный подход не учитывает конкретных значений n_i. Следует отметить, что данная проблема может возникнуть не только в социальных сферах, но и в производственных задачах, что свидетельствует о ее актуальности. Для решения данной задачи необходим принципиально новый подход, который и представлен в данной работе.

Постановка задачи

Объект – некоторая неделимая сущность, которая может находиться в определенном состоянии. В приведенном выше примере понятию объект соответствует отдельный студент, который может быть отчисленным, восстановленным либо ни тем, ни другим (нейтральным).

Группа – набор из числа n_i объектов, для которого известно количество m_i «особых» (не нейтральных) объектов в рамках рассматриваемой группы.

Композиция групп – совокупность из нескольких групп для которой: $n_{гр} = \sum n_i$, $m_{гр} = \sum m_i$, где индексы i соответствуют группам, входящим в рассматриваемую композицию.

Исходная выборка – выборка, состоящая из значений y_i, рассчитанных по формуле (1).

Итоговая выборка – выборка, на основании которой производится анализ: строятся гистограммы, делаются предположения и т.д. Для классической методологии исходная и итоговая выборки тождественны.

Сформулируем основное свойство, которому должна удовлетворять итоговая выборка.

Свойство масштабирования – эмпирическое распределение, полученное из итоговой выборки, которое должно быть применимо не только для отдельных групп, но и для их композиции. В частности, это означает, что выборочное среднее итоговой выборки должно быть равным

$$\overline{y} = \sum_{i=1}^{L} m_i \Big/ \sum_{i=1}^{L} n_i . \qquad (2)$$

Таким образом, исходная задача заключается в нахождении итоговой выборки для имеющегося набора групп, представленных целыми по своей природе значениями m_i и n_i. Причем в общем случае $n_i \neq n_j$ при $i \neq j$, что отличает данную задачу от задач, для которых может быть найдено дискретное распределение, а именно:

1) в исходной выборке отсутствует информация о размерах групп n_i, для которых рассчитывается соотношение y_i, тем самым теряется существенная для анализа информация;

2) исходная выборка не удовлетворяет свойству масштабирования. Поясним этот момент на примере двух групп, для которых $n_1 = 6$, $m_1 = 2$, $n_2 = 4$, $m_2 = 2$ соответственно. Из (1) получаем $y_1 = 1/3$ и $y_2 = 1/2$, откуда выборочное среднее значение будет равным $(1/3 + 1/2)/2 = 5/12$, что не будет соответствовать ожидаемому значению $(m_1 + m_2)/(n_1 + n_2) = 2/5$;

3) данное распределение не может быть сведено к дискретному по двум причинам. Первая причина показана на рисунке 2, на котором изображен типичный фрактал дискретного распределения соотношения двух дискретных величин (ДВ), где значения n_i и m_i берутся равномерно на интервале от 0 до 100. В частности, как и в случае решаемой выше задачи с анализом отчисления, видны пики в значениях 0 и 1, так как им соответствует максимальное количество возможных рациональных дробей. Данный эффект будем называть *фрактальным*.

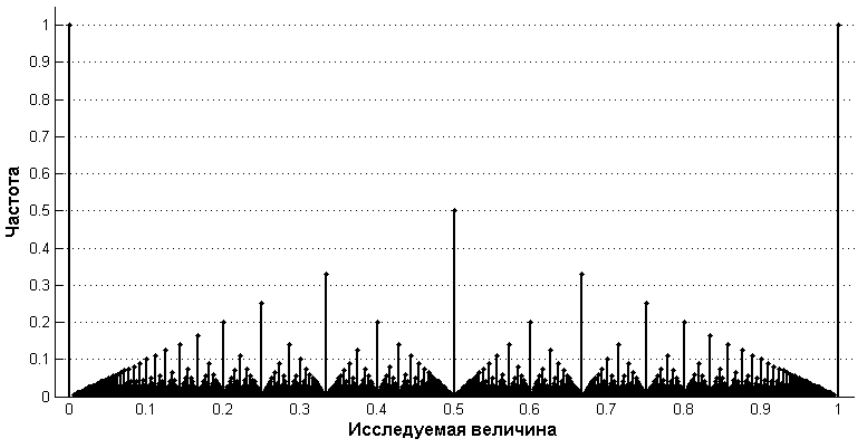


Рис. 2. Фрактальное дискретное
распределение соотношения двух дискретных величин

Вторая причина заключается в том, что при попытке использования полученных дискретных значений y_i для вычисления m_i мы неизбежно будем сталкиваться с дробными величинами, которые необходимо округлять;

4) при построении гистограммы будут наблюдаться локальные пики, связанные, как уже упоминалось выше, с фрактальным эффектом (см. рис. 1 и 2).

Таким образом, итоговая выборка должна:

- удовлетворять свойству масштабирования;
- отражать информацию о величине n_i;
- устранять фрактальный эффект;
- сводиться к исходной выборке в случае одинаковых значений n_i.

Последнее требование фактически сводит классический вариант решения задачи в частный случай, что, на наш взгляд, является еще одним достоинством предлагаемой методики.

Математическая модель и методика формирования итоговой выборки

Предлагаемая методика основывается на предположении о непрерывности исходного закона распределения $g(y)$ исследуемой величины y. В этом случае значениям n_i и m_i соответствует не отдельное дискретное значение y_i, а некоторый интервал, границы которого определяются по следующей формуле:

$$y_i \in \left(\frac{m_i - 0,5}{n_i}; \ \frac{m_i + 0,5}{n_i} \right). \tag{3}$$

В основе этой формулы лежит простое выражение $m_i = [n_i \cdot y_i]$, где квадратные скобки соответствуют операции округления до ближайшего целого. Очевидно, что данному выражению удовлетворяет не одно, а множество значений y_i из интервала, определяемого по формуле (3).

Для построения $g(y)$, воспользуемся его представлением в виде композиции нескольких распределений:

$$g\ \ y\ = \sum_{i=1}^{L} C_i g_i\ \ y\ \ , \tag{4}$$

где C_i – вещественные коэффициенты.

В качестве $g_i(y)$ используем равномерные распределения с границами, определяемыми из (3). На выбор данного типа распределения повлияли следующие теоретические предпосылки:

- равномерное распределение будет приводить к устранению фрактального эффекта, показанного на рисунке 2;
- при неизвестном распределении нельзя заранее утверждать, какое из значений y является более вероятным, а какое – менее вероятным, тем самым равносильны все допущения;
- математическое ожидание y_i должно быть равным m_i/n_i (см. свойство масштабируемости), что справедливо в случае равномерного распределения;
- как известно, ошибка округления подчиняется равномерному распределению [4].

Для определения значений C_i запишем свойство масштабируемости в следующем виде:

$$E\ \ g\ \ y\ \ = \frac{\sum_{i=1}^{L} m_i}{\sum_{i=1}^{L} n_i} = \frac{\sum_{i=1}^{L} E\ \ g_i\ \ y\ \ \cdot n_i}{\sum_{i=1}^{L} n_i}, \tag{5}$$

где E □ – математическое ожидание случайной величины.

Из данного уравнения на основании свойства суммирования математических ожиданий [4] получаем следующие значения коэффициентов C_i:

$$C_i = n_i \Big/ \sum_{i=1}^{L} n_i . \tag{6}$$

Сформулируем общий алгоритм нахождения итогового распределения:

шаг 1: из исходного набора групп, представленных своими значениями n_i и m_i, по (3) вычисляются интервалы равномерного распределения $g_i(y)$;

шаг 2: по (6) рассчитываются значения коэффициентов C_i;

шаг 3: на основании (4) производится объединение полученных на шаге 1 функций равномерного распределения.

Рассмотрим далее, каким образом на основании непрерывного распределения $g(y)$ построить итоговую выборку. Для этого необходимо разыграть случайную величину y на соответствующем распределении по методу Монте-Карло. Как известно, для композитного распределения метод разыгрывания может быть применен к каждой составляющей с последующим объединением полученных результатов [5]. Таким образом, каждый интервал с равномерным распределением будет представлен своим набором ДВ. Согласно (4), количество ДВ для соответствующего интервала должно быть пропорционально значению C_i. Как видно из (6), данному условию удовлетворяют следующие значения: n_i, $2n_i$, $3n_i$ и т.д.

Отметим, что итоговую выборку можно получить из исходного набора групп без предварительного построения непрерывного распределения, а соответствующий алгоритм примет следующий вид:

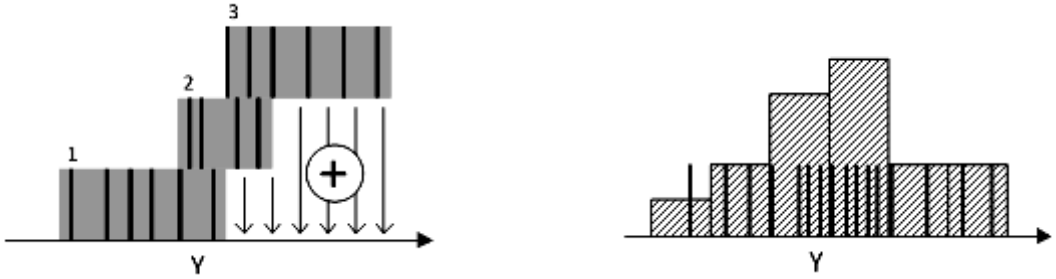
шаг 1′: из исходных наборов групп по (3) вычисляются интервалы равномерного распределения величины y;

шаг 2′: на каждом из полученных интервалов $k \cdot n_i$ раз ($k \in N$) осуществляется разыгрывание случайной величины y;

шаг 3′: полученные на шаге 2′ результаты розыгрыша для всех i объединяются в одну общую выборку, которая и является итоговой.

Сформированная таким образом выборка может быть использована в дальнейшем для построения гистограммы, проверки гипотез, нахождения среднего значения, дисперсии и т.д.

Процесс формирования выборки и построения ее гистограммы представлен на рисунках 3, *а* и 3, *б*, где рассматриваются три группы, для которых интервалы равномерного распределения отмечены на рисунке 3, *а* темным фоном. В каждом из интервалов проводится разыгрывание случайной величины (результаты розыгрышей показаны черными штрихами). Полученные значения объединяются в одну выборку (рис. 3, *б*), для которой показана возможная гистограмма.



а) *б*)

Рис. 3. Процесс формирования выборки и построения ее гистограммы:
а – иллюстрация методики формирования выборки; *б* – итоговая гистограмма

Объем итоговой выборки определяется следующим образом:

$$S = k\sum_{i=1}^{L} n_i, \quad k = 1, 2\ldots, \tag{7}$$

где значение константы k определяется на втором шаге алгоритма.

Очевидно, что объем итоговой выборки будет гораздо больше объема исходной выборки. В связи с этим может возникнуть ошибочное мнение о том, что в данной задаче используются известные методы «размножения выборки», такие как метод «складного ножа» или «бутстреп-метод» [6], которые на основании одной выборки формируют целый класс похожих выборок. К сожалению, данные методы на практике оказываются бессмысленными, так как не дают исследователю никакой дополнительной информации [7].

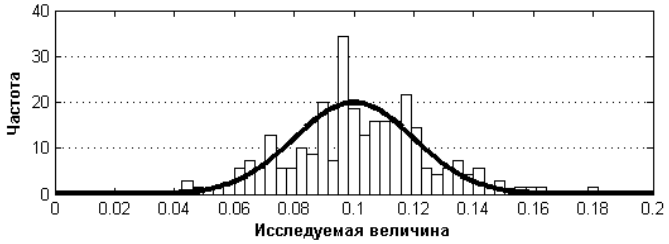
В предложенной методике увеличение размера выборки вызвано попыткой смоделировать теоретически построенную модель, в которой учитывается дополнительная информация об интервалах и размерах групп. Поэтому ее сложно отнести к классу методов «размножения выборки». Как будет показано далее, с помощью данного подхода удается достичь лучших по сравнению с классическими методиками результатов.

Сравнение методик

Проведем сравнение классической методики построения эмпирического распределения с предложенной методикой. В качестве исследуемого набора групп сгенерируем согласно таблице, представленной ниже, две последовательности, в первую из которых войдут значения n_i, а во вторую – m_i. После генерации соответствующих выборок построим их гистограммы с увеличенным количеством интервалов. Гистограммы приведены на рисунке 4, где также показана линия известного теоретического распределе-

ния. Даже визуально заметно преимущество предлагаемого подхода: полученное эмпирическое распределение во втором случае лучше согласуется с теоретическим распределением, чем в первом.

Наименование параметра	Значение параметра
Количество групп (L)	200
Минимальный размер группы ($\min\{n_i\}$)	40
Максимальный размер группы ($\max\{n_i\}$)	100
Распределение размера групп	Экспоненциальное с математическим ожиданием, равным 10
Теоретическое распределение y	Нормальное с математическим ожиданием, равным 0,1, и среднеквадратическим отклонением, равным 0,02



а)

б)

Рис. 4. Сравнение двух методик:
классической (*а*) и предлагаемой (*б*)

Дадим точечную оценку математического ожидания: для первого случая она составит 0,1025; для второго – 0,1024. Полученные значения весьма далеки от значения 0,1, заявленного нами при генерации выборок (см. таблицу). Интервальное оценивание также не дает удовлетворительного результата: 0,1025 ± 0,0001 при уровне доверия 0,95. Следовательно, можно сделать вывод, что операция округления приводит к искажению исходного теоретического распределения, поэтому очевидно необходимы другие подходы к оцениванию параметров распределения.

В качестве основы для оценивания воспользуемся сравнительно новым направлением в области математической статистики – статистикой интервальных данных.

Оценивание параметров распределения

Статистика интервальных данных отличается от классической статистики тем, что оперирует не числами, а интервалами [8]. Такая парадигма основана на представлении о неточности измерений, погрешностей в округлениях и др. Так, очевидно, что в результате округления получаются не истинные значения, а приближенные (формула (2)), которые и формируют результирующую выборку.

$$y_i = x_i + \varepsilon_i, \qquad i = 1, 2, ..., L, \tag{8}$$

где x_i – истинное значение некоторой величины; y_i – соответствующее приближенное значение; ε_i – погрешность оценки.

Статистика $R_n \to R_1$:$f(y)$, рассчитанная для приближенных значений, в общем случае отличается от статистики $R_n \to R_1$:$f(x)$, рассчитанной для истинных значений. Тем самым ставится задача оценивания $f(x)$ по $f(y)$, для чего в статистике интервальных данных вводится базовое понятие – *нотна*.

Нотной называют максимальную разницу между $f(y)$ и $f(x)$:

$$N_f\ y\ = \sup\left| f\ y\ - f\ x\ \right| = \sup\left| \sum_{1 \le i \le n} \frac{\partial f\ y}{\partial x_i} \varepsilon_i + O\ \Delta^2 \right|. \tag{9}$$

Средний квадрат ошибки для некоторой статистики равен

$$\max M \quad f(y) - a \quad ^2 = \frac{\sigma^2}{n} + N_f^2 \quad y \quad + o\left(\Delta^2 + \frac{1}{n} \right). \tag{10}$$

В отличие от классических формул, при возрастании объема выборки средний квадрат ошибки не будет стремиться к нулю, откуда следует два важных вывода:

1) статистика является смещенной оценкой и для нахождения доверительного интервала необходимо учитывать значение нотны;

2) нет необходимости в безграничном увеличении размеров выборки с целью нахождения лучшей оценки.

Рассмотрим пример оценивания математического ожидания на основе выборочного среднего.

Оценкой математического ожидания является выборочная средняя, которая записывается в виде

$$f \quad y \quad = \frac{\sum_{i=1}^{S} y_i}{S}, \tag{11}$$

где S – общий объем итоговой выборки, вычисляемый по (7).

Из (9) находим значение нотны для выборочного среднего:

$$N_f \quad y \quad = \frac{\sum_{i=1}^{L} n_i \cdot \left(\frac{0,5}{n_i} \right)}{S} = \frac{0,5L}{S}. \tag{12}$$

Выражение в скобках представляет собой математическое ожидание ошибки округления ε_i для соответствующего интервала. Таким образом, значение нотны определятся размерами исходной и итоговой выборок. Интервальная оценка математического ожидания в данном случае будет следующей:

$$\left(f \quad y \quad - A \quad \gamma \quad - \frac{0,5L}{S}; f \quad y \quad + A \quad \gamma \quad + \frac{0,5L}{S} \right), \tag{13}$$

где $A(\gamma)$ – границы с доверительной вероятностью γ, определяемые из распределения Стьюдента или нормального распределения.

Как правило, $0,5L/S >> A(\gamma)$, поэтому в практических задачах для оценки математического ожидания можно пользоваться следующей формулой:

$$\left(f \quad y \quad - \frac{0,5L}{S}; \quad f \quad y \quad + \frac{0,5L}{S} \right). \tag{14}$$

Для приводимого выше примера значение нотны будет равным 0,0098. Следовательно, интервальная оценка математического ожидания будет равна (0,1024 ± 0,0098), что согласуется с теоретическим распределением. Как показывают результаты 1000-кратного численного моделирования, выборочное среднее изменяется в рамках интервала (0,0957; 0,1047), значит, реальное значение нотны в два раза меньше рассчитанного по (14). Причиной этого явилась специфика распределения размеров отдельных групп, что требует проведения дополнительных исследований.

Отметим, что данный подход может быть с успехом использован для оценивания других статистик: дисперсии, моды, коэффициента вариации и т.д.

Заключение

В данной работе была предложена новая методика решения задачи анализа соотношения двух ДВ. Данная задача является достаточно распространенной на практике и, к сожалению, способы ее решения зачастую приводят к негативным результатам, к которым можно отнести: отсутствие масштабируемости (невозможности обобщить результат для композиции групп) и наличие фрактального эффекта, который приводит гистограммы к виду «гребенки».

В качестве решения предлагается математическая модель в виде непрерывного распределения величины соотношения. Данное распределение представляет собой композицию равномерных распределений с определяемыми границами. Такая композиция полностью устраняет фрактальный эффект. Коэффициенты, участвующие при формировании итогового распределения, позволяют использовать его не только для групп, но и для композиций групп, что соответствует свойству масштабируемости.

Предложен алгоритм формирования итоговой выборки, основанный на использовании метода Монте-Карло. Для работы данного алгоритма требуются только наборы исходных ДВ, а также генератор случайных чисел. Дополнительно предложен способ оценивания статистик по итоговой выборке, теоретической основой для которого служит статистика интервальных данных.

В ходе сравнительного анализа классической и предложенной методик была продемонстрирована несостоятельность первой из них. В результате использования данной методики для анализа отчисления студентов была получена гистограмма, представленная на рисунке 5.

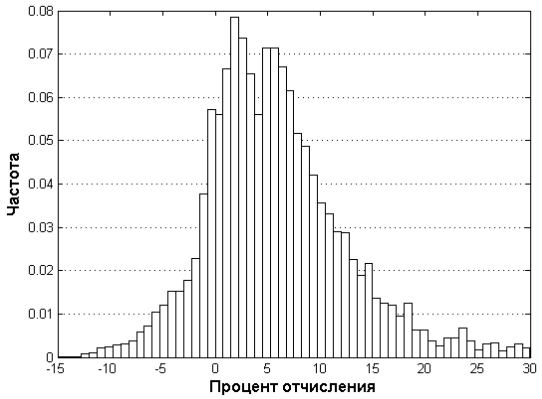


Рис. 5. Бимодальное распределение отчисления студентов

На данной гистограмме отчетливо видно бимодальное распределение, которое соответствует теоретическим предпосылкам. Кроме того, выборочное среднее, равное 0,0577, соответствует средней отчисляемости по вузу. Среди недостатков предлагаемой методики стоит отметить значительное увеличение размера выборки и, как следствие, вычислительной сложности. Самый простой способ преодоления данного недостатка заключается в случайном отборе нескольких значений из итоговой выборки. Другой недостаток заключается в необходимости реализовывать соответствующее программное обеспечение для работы с такими данными, так как ни в одном из стандартных математических пакетов таких функций не предусмотрено.

ЛИТЕРАТУРА

1. Автоматизация расчета объема учебной работы кафедр / С.С. Дереченник [и др.] // Проблемы проектирования и производства радиоэлектронных средств: сб. материалов V междунар. науч.-техн. конф., Новополоцк, 29 – 30 мая 2008 г.: в 3-х т. – Новополоцк: ПГУ, 2008. – Т. III. – С. 287 – 289.
2. Управление качеством продукции. Инструменты и методы менеджмента качества / С.В. Пономарев [и др.]; под общ. ред. С.В. Пономарева – М.: РИА «Стандарты и качество», 2005. – 248 с.
3. Bowman, A.W. Applied Smoothing Techniques for Data Analysis / A.W. Bowman, A. Azzalini. – New York: Oxford University Press, 1997. – 193 p.
4. Гмурман, В.Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов / В.Е. Гмурман. – 9-е изд., стер. – М.: Высш. шк., 2003. – 479 с.
5. Ермаков, С.М. Метод Монте-Карло и смежные вопросы / С.М. Ермаков. – 2-е изд. – М.: Наука, 1975. – 471 с.
6. Эфрон, Б. Нетрадиционные методы многомерного статистического анализа / Б. Эфрон. – М.: Финансы и статистика, 1988. – 263 с.
7. Орлов, А.И. О реальных возможностях бутстреп как статистического метода / А.И. Орлов // Заводская лаборатория. – 1987. – Т. 53, № 10. – С. 82 – 85.
8. Орлов, А.И. Прикладная статистика: учебник / А.И. Орлов – М.: Изд-во «Экзамен», 2004. – 656 с.

Поступила 16.07.2012

SMOOTHING HISTOGRAM OF DISCRETE DISTRIBUTIONS WITH MONTE CARLO METHOD

A. LYSIUK, S. DERECHENNIK

We have considered the task of forming the empirical distribution for the two discrete variables ratio with a varying denominator. The main problems that arise when using classical methods were listed. Fractal effect, which is present at the final histogram, has been shown, and we explained the reasons for its occurrence. We have formulated a theoretical algorithm of this problem, and we have also offered its version for computers. The main idea of the approach is generate a set of values using the Monte Carlo method. These values are used to construct the final histogram. The resulting histogram eliminates all the classical methods' disadvantages. Experimental comparison has showed that the proposed method is better than the classical method. In addition, we have proposed a method to estimate the parameters of the final empirical distributions. This method is based on the statistics for data having interval.