Бикритериальный метод построения и оценки качества гистограмм

В.Н. Петрушин, М.В. Ульянов, И.А. Чертихина, Е.В. Никульчев

Аннотация. В статье предлагается новый метод построения гистограмм, основанный на бикритериальной оценке их качества. Одним из критериев оценки является достоверность сегментов гистограммы, вторым – согласованность эмпирической функции распределения и функции распределения, построенной по гистограмме. Значение первого критерия растет с уменьшением числа полусегментов гистограммы, в то время как значение второго критерия возрастает с увеличением числа полусегментов. Введенный в статье, на этой основе, комплексный критерий оценки качества гистограммы, позволяет впервые оценить качество гистограммы. Максимизация значения предложенного критерия приводит к определению оптимального числа сегментов гистограммы, что позволяет повысить надежность решений, принимаемых на основе выборке, в частности, по аппроксимации наблюдаемой случайной величины некоторым известным законом распределения.

Ключевые слова: гистограмма, статистические оценки выборки, надежность гистограммы, число полусегментов гистограммы, метод оценки качества гистограммы, метод построения гистограммы по выборке.

Введение

В математической статистике одним из важнейших этапов первичного анализа экспериментальных данных, распределенных по не известному исследователю закону распределения, является построение гистограммы (этап гистограммирования), результаты которого дают представление о наблюдаемой функции плотности распределения вероятностей. Обработка экспериментальных данных известными методами математической статистики позволяет получить гистограмму для наблюдаемой случайной величины [1,2], но, в общем случае, не дает возможности оценить ее качество. Полученная гистограмма может быть использована при дальнейшем анализе экспериментальных данных в качестве:

– важного наглядного инструмента понимания и анализа поведения случайной величины в диапазоне наблюдаемого размаха варьирования и принятия решения о применимости специальных методов математической статистики для обработки и исследования полученной выборки;

– средства оценки влияния неопределенности информации, отражающейся в исходных данных, на решения практически значимых задач – нахождения таких оценок неопределенности, как оценки погрешности, оценки границ множества решений и т.п.;

– единственно возможной информационной базы для аппроксимации или идентификации функции плотности распределения вероятностей или закона распределения, наблюдаемого в экспериментально полученных данных, равно как и для проверки гипотезы о предполагаемом виде закона распределения [1, 2];

– средства выявления наиболее вероятных сегментов значений случайных величин, что позволяет более точно формулировать вероятностные критерии оценки качества исследуемых объектов (в качестве примера приведем предложенные двумя из авторов данной статьи совместно с В.А. Головешкиным понятие информацион-

[1] Работа выполнена при поддержке РФФИ (грант № 11-07-00772-а).

ной чувствительности компьютерных алгоритмов и методы ее количественной оценки, существенно опирающиеся на анализ экспериментальных данных и аппроксимацию гистограмм известными функциями плотности [3]);

– инструмента гистограммной арифметики, оперирующей вероятностным представлением входных данных, например, в виде гистограммных чисел с последующей разработкой численных операций над ними [4].

При получении аппроксимации выборок функциями плотности вероятности распределения случайных величин наличие такой информации дает возможность при расчетах учитывать и получать результаты в виде случайных величин с построенной плотностью распределения вероятностей. «В тех случаях, когда это возможно, численные операции над плотностями вероятности случайных величин позволяют существенно поднять точность расчетов при сравнительно небольшом объеме вычислений» [5].

Таким образом, гистограммы являются важным и информационно значимым механизмом исследования случайных величин. Очевидно, что качество гистограммы, построенной по экспериментальным данным, отражается на дальнейших исследовательских результатах и выводах.

1. Терминология и обозначения

Следуя в основном [1] и вводя некоторые собственные обозначения, будем использовать далее следующую терминологию и обозначения, связанные со статистическим анализом данных и гистограммированием:

X – непрерывная случайная величина, наблюдаемые значения которой составляют выборку;

n – объем выборки – число реализаций случайной величины X;

$V = \{x_1, \ldots, x_n\}$ – собственно выборка (экспериментальные данные – реализации X);

$R = (\max x_i - \min x_i), i = \overline{1,n}$ – размах варьирования выборки;

$\overline{X}$ – выборочное среднее;

S^2 – выборочная исправленная (несмещенная) оценка дисперсии;

$S = \sqrt{S^2}$ – стандартное отклонение;

$\widetilde{V} = \{\widetilde{x}_1, \ldots, \widetilde{x}_n\}$ – ранжированный вариационный ряд: элементы выборки, сортированные по не убыванию;

k – число полусегментов гистограммы (число групп);

$j = \overline{1,k}$ – номера полусегментов гистограммы;

$x_j^{(\min)}, x_j^{(\max)}$ – границы j-ого полусегмента гистограммы;

$n_j, j = \overline{1,k}$ – объем группы: число элементов выборки x_i, попавших в j-ый полусегмент $x_j^{(\min)} \le x_i < x_j^{(\max)}$, для значения $j = k$ – последнего сегмента гистограммы $x_k^{(\min)} \le x_i \le x_k^{(\max)}$;

$w_j = n_j / n, j = \overline{1,k}$ – относительные частоты в полусегментах гистограммы;

$F_V(\widetilde{x}_i) = i/n, i = \overline{1,n}$ – эмпирическая функция распределения, построенная по ранжированному вариационному ряду;

$G = \left\langle k, \left\{ \left(x_j^{(\min)}, x_j^{(\max)}, w_j \right) \right\}, j = \overline{1,k} \right\rangle$ – гистограмма, заданная числом полусегментов, с определенными для каждого полусегмента границами и соответствующими частотами;

$Gm : V \to G$ – отображение выборки в гистограмму, таким образом, Gm есть функция гистограммирования: $G = Gm(V)$;

Gm-метод – конкретная реализация функции Gm: используемый исследователем метод построения гистограмм;

$Q(V, G) = Q(V, Gm(V))$ – критерий оценки качества гистограммы G для данной выборки V.

2. История вопроса и недостатки существующих методов гистограммирования

При гистограммировании выборки возникают задачи определения числа групп и определения границ полусегментов, т.е. группировки данных в зависимости от их особенностей. Порядок решения этих задач зависит от применяемого исследователем Gm-метода.

Рассмотрим вначале Gm-методы, базирующиеся на выборе числа групп. История вопроса в теории построения гистограмм начинается с имени Г. Стержесса [6], предложившего

в 1926 году следующую формулу для определения числа групп k:

$$k = 1 + \lfloor \log_2 n \rfloor, \quad (1)$$

которая основана на рассмотрении «идеальной» гистограммы для случайной величины, подчиненной биномиальному распределению с объемом выборки, равным степени двойки. Другая формула для числа групп (полусегментов) указана в [7] и основана на сравнении стандартного отклонения в группе со стандартным отклонением средней

$$k = \left\lceil n^{1/2} \right\rceil. \quad (2)$$

Обратим внимание на то, что для больших объемов выборки рекомендуемое число групп вычисленное по формулам (1) и (2), будет значительно отличаться, так для $n = 1024$ формула (1) дает 11 полусегментов, в то время как формула (2) – 32. Заметим, что формулы (1) и (2) не позволяют определить границы полусегментов: $x_j^{(\min)}, x_j^{(\max)}$ и их длины h_j, которые в этом случае устанавливаются априори, чаще всего в виде

$$h_j = R/k \ \forall j = \overline{1,k}. \quad (3)$$

Другую группу Gm-методов образуют методы с первоначальном определением длин полусегментов. Метод, описанный И.Е. Тарасовым в [7], предполагает построение полусегментов такой длины, которая позволяет на основе известной функции плотности, интегрировать порядка $1/n$ вероятности, и требует, тем самым, знания закона распределения случайной величины.

Формулы (1) и (2) и метод из [7] не учитывают стандартного отклонения выборки. В связи с этим Скотт [8] в 1979 г. обосновывает следующую формулу для длины полусегмента

$$h = 3{,}5 \cdot S \cdot n^{-1/3}, \quad (4)$$

а Фридман и Диаконис [9] в 1981 г. предлагают метод определения длины, использующий межквартильный ранг (IQ) – разницу между верхним и нижним квартилем

$$h = 2 \cdot (IQ) \cdot n^{-1/3}. \quad (5)$$

Отметим, что формулы Скотта и Фридмана-Диакониса в отличие от [7] создают полусегменты равной длины, что приводит к неоднозначности при отсутствии кратности h в R. В отраслевой статистике группировка осуществляется по наличию или отсутствию каких-либо признаков из их устоявшегося (ставшего традиционным) набора [10]. Такой подход имеет право на существование, но довольно часто, причем вполне обосновано, нарушаются и эти правила группировки.

Таким образом, одну и туже выборку, в зависимости от выбранного Gm-метода, можно представить различными, при этом весьма непохожими, гистограммами. В связи с этим возникает возможность манипуляции данными при отсутствии критерия оценки качества гистограммы. Отметим, что в настоящее время в математической статистике нет четких критериев такой оценки [1, 2]. Существующие подходы к улучшению качества гистограммы носят эмпирический характер и выглядят, например, как рекомендации по укрупнению интервалов, полученных разбиением размаха варьирования на рекомендуемое число равных интервалов [1,2] и т.д.

3. Постановка задачи

Таким образом, для повышения точности и достоверности результатов по исследованию наблюдаемых случайных величин представляет интерес задача построения критерия оценки качества гистограммы и метода построения гистограмм, основанного на максимизации такого критерия, который определяет как число необходимых полусегментов, так и их границы на основе объема выборки и собственно регистрированных значений изучаемой случайной величины.

В соответствии с вышеизложенным в настоящей статье предлагается вариант решения следующих двух взаимосвязанных задач:

1. разработка критерия оценки качества гистограммы $Q(V,G) = Q(V, Gm(V))$, учитывающего достоверность выделенных полусегментов и качество аппроксимации полученной гистограммой эмпирической функцией распределения;

2. разработка рационального Gm-метода – Gm^* для построения гистограммы G^*, включающего определение числа полусегментов k и границ группировки данных: $x_j^{(\min)}, x_j^{(\max)}$,

который для данной выборки максимизирует значение предложенного критерия:

$$G^* = \arg\max_{G=Gm(V)} Q(V, Gm(V)). \qquad (6)$$

4. Оценка качества аппроксимации эмпирической функции распределения полученной гистограммой

Для оценки качества аппроксимации эмпирической функции распределения гистограммой, полученной каким-либо Gm-методом, авторы предлагают использовать следующий подход, основанный на применении критериев согласия.

С одной стороны, построенная на основе вариационного ряда эмпирическая функция распределения $F_V(\tilde{x}_i) = i/n$ отражает особенности поведения наблюдаемой случайной величины и использует все элементы выборки. В этом аспекте мы вправе рассматривать $F_V(\tilde{x}_i)$ как «эталонную» функцию распределения. С другой стороны, построение гистограммы, независимо от принятого Gm-метода, приводит к группировке данных. Полученная относительная частота в полусегменте w_j априорно предполагает равномерность распределения данных выборки в соответствующем полусегменте. Таким образом, полученная гистограмма может рассматриваться как аппроксимация неизвестного закона распределения кусочно-равномерными плотностями (по полусегментам). Интегрирование гистограммы приводит к получению кусочно-линейной аппроксимации эмпирической функции распределения (пример для некоторой выборки приведен на Рис. 1). Обозначим полученную интегрированием гистограммы на полном размахе варьирования кусочно-линейную аппроксимацию эмпирической функции распределения вероятностей через $F_G(x), x \in [\tilde{x}_1, \tilde{x}_n]$ и будем называть ее далее гистограммной функцией распределения. Функция $F_G(x)$ представима в следующем виде

$$F_G(x) = \{a_j x + b_j\},\ j = \overline{1,k}. \qquad (7)$$

Таким образом, возникает задача проверки гипотезы о соответствии эмпирической функции $F_V(\tilde{x}_i)$, рассматриваемой как эталонная, и гистограммной функции $F_G(x)$, вычисленной в точках вариационного ряда $F_G(\tilde{x}_i)$. Для решения этой задачи воспользуемся критерием Колмогорова [1]. В рассматриваемой ситуации статистикой критерия является величина

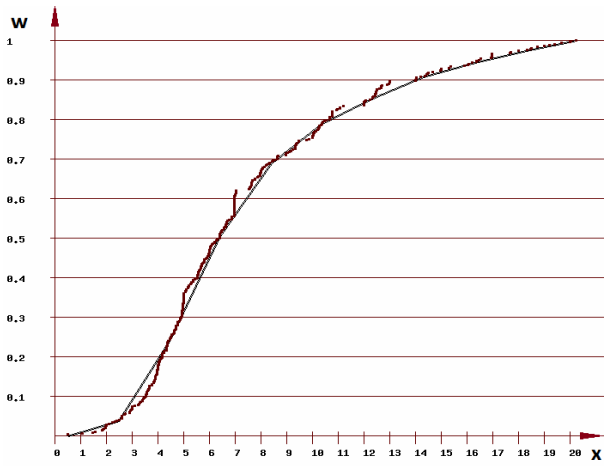


Рис. 1. Пример эмпирической функции распределения и ее кусочно-линейной гистограммной аппроксимации

$$D_n = \max_{i=\overline{1,n}} |F_V(\tilde{x}_i) - F_G(\tilde{x}_i)|. \qquad (8)$$

Теорема Колмогорова утверждает, что статистика $\sqrt{n}D_n$ зависит только от объема выборки, не зависит от закона распределения выборки (предполагая непрерывность распределения) и подчиняется следующему интегральному закону распределения вероятностей [1]

$$\lim_{n\to\infty} P(\sqrt{n}D_n \le x) = K(x) = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2x^2}. \qquad (9)$$

Сходимость по пределу достаточно быстрая, и как указано в [1], формула (9) применима при объеме выборки $n \ge 20$. Заметим, тем самым, что предлагаемый авторами подход не применим к малым ($n < 20$) выборкам.

Поскольку нашей основной задачей является оценка точности аппроксимации эмпирической функции распределения полученной гистограммой, то авторы предлагают использовать в качестве меры значение вероятности ошибки первого рода α в точке наблюдаемого значения статистики критерия Колмогорова, т.е. в точке $x = \sqrt{n}D_n$. Обозначим эту вероятность

через $\alpha(V,G)$, поскольку аппроксимация фиксированной выборки V различными гистограммами G приведет к изменению наблюдаемого значения критерия D_n по формуле (8), а следовательно, и вероятности $\alpha(V,G)$. Отметим, что значение $\alpha(V,G)$ зависит не только от предложенной гистограммы, но и от выборки, по которой строится эмпирическая функция распределения. Используя (9), приведем аналитическую формулу для вычисления $\alpha(V,G)$

$$\alpha(V,G)=\int_{\sqrt{n}D_n}^{\infty} K'(x)dx = 1-K\left(\sqrt{n}D_n\right). \qquad (10)$$

Увеличение числа полусегментов гистограммы приведет, очевидно, к лучшей аппроксимации эмпирической функции распределения, тем самым наблюдаемое значение критерия Колмогорова D_n (при фиксированной выборке $n=const$) будет уменьшаться, нижний предел интеграла (10) будет смещаться влево, что приведет к увеличению значения $\alpha(V,G)$.

Таким образом, мы вводим первый компонент комплексной оценки качества гистограммы G как качество аппроксимации этой гистограммой эмпирической функции распределения выборки $F_V(\tilde{x}_i)$: $\alpha(V,G)=1-K\left(\sqrt{n}D_n\right)$ – вероятность ошибки первого рода критерия Колмогорова в точке наблюдаемого значения критерия.

5. Оценка достоверности (надежности) полусегментов гистограммы

Для оценки достоверности разбиения выборки по полусегментам авторы предлагают использовать показатель надежности оценки среднегруппового значения. Из математической статистики известно, что интервальная оценка средней групповой формируется на основе распределения Стьюдента [2]. Пусть $\bar{x}_j$ – выборочная групповая средняя в j-ом полусегменте, а $\overline{X}_j$ – математическое ожидание групповой средней. Тогда при заданной надежности (доверительной вероятности) γ_j доверительный интервал для $\overline{X}_j$ определяется в виде [2]:

$$\overline{X}_j \in \left(\bar{x}_j-\delta_j, \bar{x}_j+\delta_j\right), \delta_j=\frac{t(\gamma_j,n_j)\cdot S_j}{\sqrt{n_j}}, \qquad (11)$$

где $t(\gamma_j,n_j)$ – значение критерия Стьюдента при выбранной доверительной вероятности γ_j и объеме группы, а $S_j=\sqrt{S_j^{\,2}}$, где $S_j^{\,2}$ – несмещенная оценка внутригрупповой дисперсии в j-ом полусегменте. Заметим, что обращением формулы (11) в случае уже имеющейся гистограммы, для которой известны значения S_j, n_j и δ_j, причем $\delta_j \le 1/2\left(x_j^{(\max)}-x_j^{(\min)}\right)$ [3], можно вычислить оценку доверительной вероятности (надежности) γ_j по следующей формуле

$$\gamma_j = t^{-1}\left(\frac{\delta_j\cdot\sqrt{n_j}}{S_j}, n_j\right), \qquad (12)$$

Очевидно, если принять гипотезу о независимости групповых средних, надежность гистограммы в целом $\gamma(G)$ будет представлять собой произведение надежности всех групповых средних γ_j. Таким образом, мы получаем второй компонент оценки качества гистограммы

$$\gamma(G)=\prod_{j=1}^{k}\gamma_j=\prod_{j=1}^{k}t^{-1}\left(\frac{\delta_j\cdot\sqrt{n_j}}{S_j}, n_j\right). \qquad (13)$$

6. Бикритериальная оценка качества гистограммы

На основании вышеизложенного авторы предлагают следующую комплексную бикритериальную оценку качества гистограммы

$$Q(V,G)=Q(\alpha(V,G),\gamma(G))=\alpha(V,G)\cdot\gamma(G), \qquad (14)$$

где значение $\alpha(V,G)$ вычисляется по формуле (10), а $\gamma(G)$ – по формуле (13).

Отметим, что критерий учитывает как достоверность (надежность) выделенных полусегментов гистограммы, так и качество аппроксимации эмпирической функции распределения полученной гистограммой. Введенная оценка качества является бикритериальной – для фиксированной выборки значительное увеличение числа полусегментов ведет к увеличению

достоверности аппроксимации $\alpha(V,G)$ и, очевидно, к уменьшению числа наблюдений в полусегментах и сокращению их длин, что влечет уменьшение доверительной вероятности в полусегментах, а, следовательно, и общего произведения $\gamma(G)$.

Приведем результаты расчета значений критерия для тестовой выборки объемом 400 (файл выборки доступен по адресу http://kaputpixel.ru/math/) для различного числа полусегментов с равной длиной. Расчетные данные приведены в Табл. 1, соответствующий график – на Рис. 2.

Заметим, что наилучшее значение критерия достигается при $k=11$ с надежностью $\gamma(G)=0{,}963$ и $\alpha(V,G)=0{,}407$, однако, если дополнительное требование состоит в наилучшей аппроксимации эмпирического распределения при понижении требованиия к надежности гистограммы до $\gamma(G)>0{,}8$, то рациональным будет значение $k=17$, при котором $\alpha(V,G)=0{,}421$.

Табл. 1. Значения критерия качества $Q(V,G)$ в зависимости от числа групп

k	$\gamma(G)$	$\alpha(V,G)$	$Q(V,G)$
10	0,981	0,149	0,146
11	0,963	0,407	0,392
12	0,962	0,259	0,249
13	0,904	0,202	0,183
14	0,903	0,315	0,284
15	0,858	0,270	0,232
16	0,805	0,304	0,245
17	0,802	0,421	0,337
18	0,752	0,292	0,219
19	0,592	0,407	0,241
20	0,378	0,528	0,200

7. Построение полусегментов гистограммы с заданной надежностью

Предложенный критерий оценки качества позволяет разработать метод построения гистограммы, максимизирующий значение $Q(V,G)$. Начнем изложение этапов этого метода с построения полусегментов, т.е. выделения групп. Идея метода формирования группы состоит в следующем. На основании (11), задавая некоторое начальное значение объема группы, мы, итерационно, можем получить такой объем группы и значения границ соответствующего полусегмента гистограммы, при котором интервальная оценка средней групповой в полусегменте j с приятой надежностью γ_j не выходит за пределы границ полусегмента. На основании (11) по заданной γ_j можно построить доверительный интервал для известного полусегмента. Теперь потребуем обязательного выполнения следующих условий, выполнение которых приводит к попаданию доверительного интервала для $\overline{X}_j$ в границы полусегмента:

$$\begin{cases} \overline{x}_j - \delta_j > x_j^{(\min)} \\ \overline{x}_j + \delta_j < x_j^{(\max)} \end{cases}, \qquad (15)$$

Полученная система неравенств позволяет построить алгоритм выделения групп в вариационном ряду выборки на основе задания доверительных вероятностей в полусегментах γ_j. Итерационное (по n_j) решение этой системы неравенств является алгоритмом выделения группы, причем в результате решения системы (15) мы получаем не только объем группы,

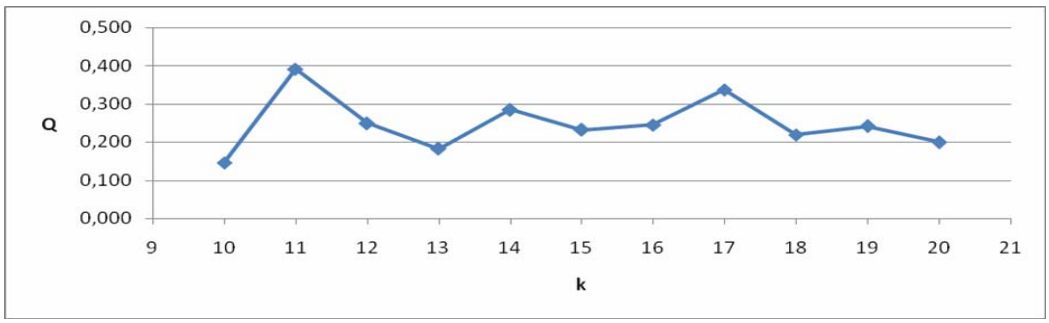


Рис. 2. Зависимость качества гистограммы от числа групп в модельной выборке

но и границы полусегмента. Предложенный подход существенно ограничивает волюнтаризм в построении гистограмм, хотя и не исключает его полностью. Если условие (15) не выполняется для последнего сегмента, то его следует присоединить к предыдущему. Заметим, что знание крайних включенных в полусегмент значений выборки оставляет возможность определенного варьирования границами полусегментов между крайним правым значением в предыдущем полусегменте и крайним левым в текущем.

Если исследователь задает надежность гистограммы $\gamma(G)$, то на основе (13) можно оценить надежности в полусегментах – $\tilde{\gamma}_j = \sqrt[k]{\gamma(G)}$. При этом предъявляемые требования по надежности влияют на длины полусегментов, но не определяют их точных границ. Возникает дополнительная задача оптимизации границ полусегментов. Для ее решения воспользуемся второй оценкой качества гистограммы по адекватности – $\alpha(V,G)$. Если максимальное отклонение, вычисляемое критерием Колмогорова (D_n), велико, то есть возможность его уменьшения за счет сужения полусегмента, которому оно принадлежит, путем перемещения его границ.

8. Учет внутригрупповой дисперсии при построение полусегментов гистограммы

Заметим, что система неравенств (15) может быть дополнена на основе оценки дисперсии в полусегменте (внутригрупповой дисперсии [2]). Дело в том, что поведение части выборки, попавшей в полусегмент гистограммы, совсем не обязательно подчиняется равномерному распределению, которым оно аппроксимируется в этом полусегменте гистограммы. В связи с этим рассмотрим требования к группе, определяемые на основе максимальной внутригрупповой дисперсии. Известно [3], что случайная величина с ограниченным носителем имеет максимальную дисперсию, если она принимает равновероятные значения на границах сегмента варьирования. Таким образом, если заменить значения вариационного ряда в полусегменте на ближайшую к этим значениям границу, то внутригрупповая дисперсия возрастет.

Пусть m_j – число значений в полусегменте, более близких к его правой границе, тогда $(n_j - m_j)$ – число значений, более близких к левой. Заменим наблюдаемые значения на соответствующие граничные, что позволяет получить оценку сверху для внутригрупповой дисперсии:

$$S_j^{\,2} \le \frac{m_j(n_j - m_j)}{n_j(n_j - 1)}\left(x_j^{(\max)} - x_j^{(\min)}\right)^2. \qquad (16)$$

Заметим, что при данной замене выборочная групповая средняя вычисляется по формуле:

$$\bar{x}_j = \frac{x_j^{(\min)}(n_j - m_j) + x_j^{(\max)} m_j}{n_j}. \qquad (17)$$

Подставим полученную оценку (16) и значение выборочной групповой средней (17) в систему неравенств (15). Решив полученную систему неравенств относительно m_j и опуская промежуточные выкладки, мы получим ограничение в виде неравенства на размещение наблюдаемых значений относительно границ полусегмента

$$\frac{n_j \cdot t^2(\gamma_j, n_j)}{n_j - 1 + t^2(\gamma_j, n_j)} < m_j \le \frac{n_j(n_j - 1)}{n_j - 1 + t^2(\gamma_j, n_j)}. \qquad (18)$$

Этот результат накладывает новое ограничение на формирование группы. При небольших значениях объема группы обязательным является близкое к симметричному распределение наблюдаемых значений относительно середины полусегмента. В случае, если двойное неравенство (18) не имеет решения, строить гистограмму нецелесообразно.

Дополнительное ужесточение требований по формированию группы и полусегмента связано с учетом случайности самой оценки внутригрупповой дисперсии. Для этого введем поправку $q(\gamma_j, n_j)$ на доверительный интервал внутригрупповой дисперсии. Метод расчета поправки приведен в [2]. В итоге объем группы n_j, при выполнении условий на размещение (18), может быть определен на основе следующей системы неравенств:

$$\begin{cases} \bar{x}_j - \dfrac{t(\gamma_j, n_j)\cdot S_j\left(1+q(\gamma_j, n_j)\right)}{\sqrt{n_j}} > x_j^{(\min)} \\ \bar{x}_j + \dfrac{t(\gamma_j, n_j)\cdot S_j\left(1+q(\gamma_j, n_j)\right)}{\sqrt{n_j}} < x_j^{(\max)} \end{cases}. \quad (19)$$

Отметим, что в соответствии с (19) рассмотрение оценки внутригрупповой дисперсии как случайной величины приводит к увеличению объема группы и обеспечивает устойчивость значения числа полусегментов гистограммы.

9. Бикритериальный метод построения гистограмм

Проведенное исследование позволяет сформулировать основные этапы предлагаемого Gm-метода, основанного на бикритериальной оценке качества гистограммы.

1. Задание оценочных значений качества гистограммы $Q(V,G)$ и его компонент $\gamma(G)$ и $\alpha(V,G)$ исходя из целей статистического исследования выборки. Поскольку предложенный критерий определен как произведение компонент, то возникает возможность интерпретации фиксированного значения качества $Q(\alpha(V,G),\gamma(G)) = \alpha\cdot\gamma = const$ как парето-оптимальной границы, на которой значения компонент критерия могут выбираться исследователем на основании дополнительных требований, заданных, например, путем введения весов для компонент комплексного критерия. Область, близкая в пространстве компонент критерия к точке пересечения гиперболы $Q(\alpha(V,G),\gamma(G)) = \alpha\cdot\gamma = \text{const}$ и прямой $b\cdot\alpha(V,G)+(b-1)\cdot\gamma(G), 0<b<1$, и определяет область выбора наилучшей гистограммы.

2. Вычисление начального значения надежности в полусегментах гистограммы $\tilde{\gamma}_j = \sqrt[k]{\gamma(G)}$ на основе выбранного значения надежности гистограммы в целом $\gamma(G)$.

3. Расчет объема групп на основе неравенства (15) или неравенства (19) в зависимости от требований исследователя при выполнении условий на размещение (18) с учетом объема выборки. Этап является итерационным – увеличение объема группы происходит до достижения заданного порога $\tilde{\gamma}_j$. Если в последнем сегменте требуемая надежность не достигается. то он объединяется с предыдущим. На этом этапе достигается выполнение требования на $\gamma(G)$.

4. Построение гистограммной функции распределения и расчет $\alpha(V,G)$ по формуле (10) с оптимизацией $\alpha(V,G)$ путем варьирования границ полусегментов. В случае, если значение по качеству аппроксимации не удовлетворяет исследователя, то возможно понижение требования по $\gamma(G)$ и переход к шагу 2 для построения новой гистограммы.

5. Расчет значения комплексного критерия качества $Q(\alpha(V,G),\gamma(G)) = \alpha\cdot\gamma$ полученной гистограммы.

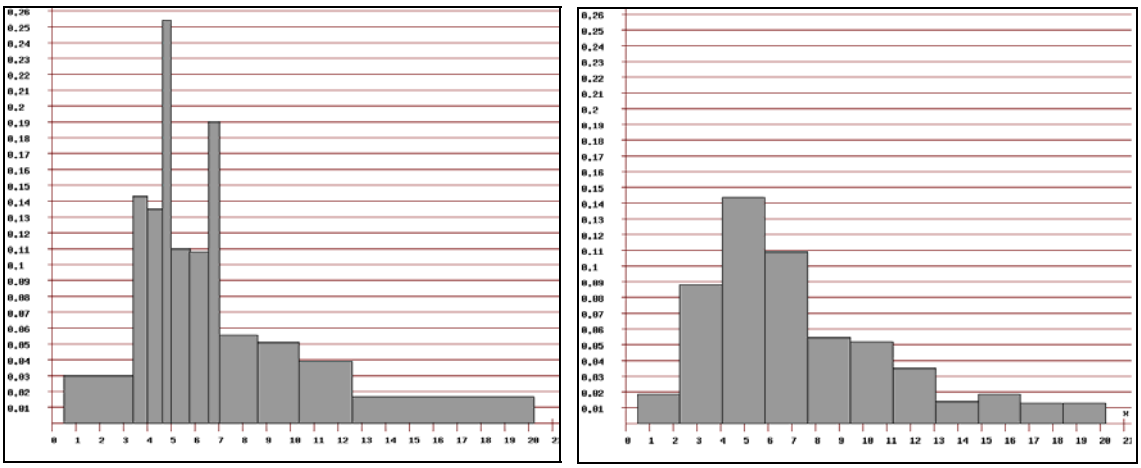
10. Модельный пример

Применение предложенного метода к тестовой выборке позволило получить следующие результаты. Проведенное гистограммирование с учетом оценки внутригрупповой дисперсии позволило улучшить значение критерия качества по сравнению с равномерным разбиением с $Q(V,G) = 0{,}392$ до $Q(V,G^*) = 0{,}491$. Значения компонент критерия приведены в Табл. 2. При этом оптимальное значение числа полусегментов осталось равным 11.

Табл. 2. Значения $Q(V,G)$ для равномерного разбиения и предложенного метода

k	$\gamma(G)$	$\alpha(V,G)$	$Q(V,G)$
11 (равномерно)	0,963	0,407	0,392
11 (предложенный метод)	0,989	0,496	0,491

Полученная гистограмма приведена на Рис. 3 а. На Рис. 3 б мы показываем для сравнения вид гистограммы при $k = 11$ с полусегментами равной длины. Заметим, что две гистограммы имеют качественные отличия, в том числе: предложенный метод позволил выявить бимодальный характер выборки, в то время как гистограмма с равномерным разбиением в окрестности моды имеет унимодальный характер. Соответствующая рациональной гистограмме аппроксимация эмпирической функции распределения показана на Рис. 4. При этом качество аппроксимации составляет $\alpha(V,G) = 0{,}496$.



а) б)

Рис. 3. Гистограмма по предложенному методу (а), по полусегментам равной длины (б)

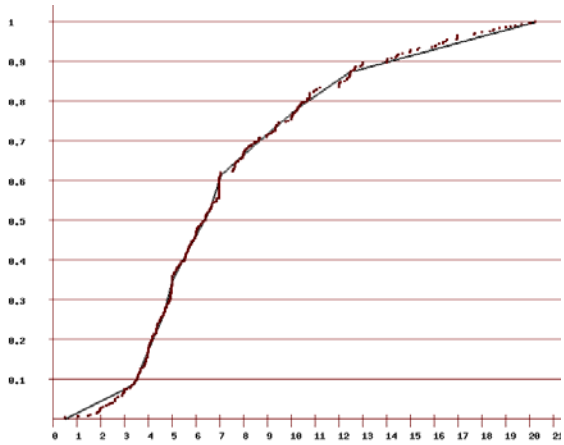


Рис. 4. Аппроксимация эмпирической функции распределения оптимальной гистограммой

Заключение

Таким образом, в статье предложен новый метод построения гистограмм, основанный на бикритериальной оценке их качества. Комплексный критерий качества гистограммы включает в себя две оценки – достоверность (надежность) сегментов гистограммы и согласованность эмпирической функции распределения с гистограммной функцией распределения. Значение первого критерия растет с уменьшением числа полусегментов гистограммы, в то время как значение второго критерия возрастает с увеличением числа полусегментов, и метод позволяет найти компромиссное (в смысле предложенного критерия) решение между этими противоречивыми критериями.

Укажем существенные отличия предложенного метода (при соблюдении условий его применимости) от существующих:

– наличие критерия оценки качества гистограммы;

– учет адекватности получаемой гистограммы исходному вариационному ряду выборки за счет рационального выбора числа сегментов;

– учет надежности показателей внутри сегментов гистограммы, обеспечивающий компромисс с показателем адекватности полученной гистограммы исследуемой выборке;

– возможность контролируемого варьирования компонентами показателя качества при его фиксированном значении, что обеспечивает гибкость учета требований исследователя.

Итерационное решение задачи максимизации значения предложенного критерия приводит к определению рационального числа сегментов гистограммы и значений их границ, что позволяет получить гистограмму, наиболее объективно отражающую исследуемую выборку. В конечном итоге предложенный метод позволяет повысить надежность решений, принимаемых на основе статистической обработки выборки, и может быть использован при построении гистограмм для различных выборок, в частности, при аппроксимации эмпирической функции распределения известными функциями плотности, при анализе информационной чувствительности компьютерных алгоритмов и при решении задач анализа загрузки компьютерных сетей.

Литература

1. Лагутин М. Б. Наглядная математическая статистика – М.: БИНОМ. Лаборатория знаний, 2007. – 472 с.
2. Гмурман В. Е. Теория вероятностей и математическая статистика – 9-е изд., стер.– М.: Высш. шк., 2003.– 479 с.
3. Петрушин В. Н.. Ульянов М. В. Информационная чувствительность компьютерных алгоритмов. – М.: ФИЗМАТЛИТ, 2010. – 224 с.
4. W.Li, J.Hym, Computer arithmetic for probability distribution variables, Reliability Engineering and System Safety, 85(2004).
5. Б. С. Добронец, О. А. Попова «Численные операции над случайными величинами и их приложения», Журн. СФУ. Сер. Матем. и физ., 4:2 (2011), С. 229–239.
6. Sturges, H. (1926) The choice of a class-interval. J. Amer. Statist. Assoc., 21, 65–66.
7. Тарасов И. Е. О выборе интервалов гистограммирования // Системы управления и информационные технологии, 2011, №2.1(44), С. 181–184.
8. Scott, D.W. (1979) On optimal and data-based histograms. Biometrika, 66, 605–610.
9. Freedman, D. and Diaconis, P. (1981) On this histogram as a density estimator: L2 theory. Zeit. Wahr. ver. Geb., 57, 453–476.
10. Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник - 4-е изд., перераб. и доп. - М.: Финансы и статистика, 2002. - 480 с: ил.

Петрушин Владимир Николаевич. Зам. зав. кафедрой Всероссийской государственной налоговой академии Минфина России. Окончил Московский государственный университет им. М.В. Ломоносова в 1974 году. Кандидат физико-математичсеких наук (1988), доцент (1991). Автор более 75 научных работ, в том числе одной монографии. Область научных интересов: теория вероятностей, математическая статистика, теория эксперимента.

Ульянов Михаил Васильевич. Профессор кафедры Национального исследовательского университета - Высшей школы экономики, профессор кафедры Московского государственного университета печати им. Ивана Федорова. Окончил Московский институт электронного машиностроения в 1979 году. Доктор технических наук (2005), профессор (2006). Автор более 70 научных работ, в том числе 5-и монографий. Область научных интересов: анализ, разработка ресурсно-эффективных компьютерных алгоритмов и оценка их качества. E-mail: muljanov@mail.ru

Никульчев Евгений Витальевич. Проректор по информатизации Всероссийской государственной налоговой академии Минфина России. Окончил Московскую государственную академию приборостроения и информатики в 1997 году. Доктор технических наук (2007), профессор (2011). Автор 124 работ, в том числе 4 монографий. Область научных интересов: системный анализ, моделирование и идентификация динамических систем, нелинейная динамика, автоматизация сложных процессов и систем. E-mail: nikulchev@mail.ru

Чертихина Ирина Александровна. Аспирантка Московского государственного университета печати им. Ивана Федорова. Окончила Московский государственный университет печати в 2009 году. E-mail: chertikhinai@gmail.com