Доверительная трудоемкость – новая оценка качества алгоритмов

М. В. Ульянов, В. Н. Петрушин, А. С. Кривенцов

Аннотация. Рассматриваются вопросы,связанные с оценкой качества компьютерных алгоритмов по критерию трудоемкости. Классически применяемая оценка трудоемкости в среднем позволяет получить значимые результаты только в статистическом смысле, т. е. оценить алгоритм на большом числе входов с фиксированной длиной. В статье вводится интервальная оценка – доверительная трудоемкость, построенная по аналогии с доверительными интервалами математической статистики. Предлагается использовать бета-распределение для аппроксимации распределения значений трудоемкости как ограниченной дискретной случайной величины, приводится методика определения доверительной трудоемкости как функции длины входа алгоритма.

Ключевые слова: бета-распределение, доверительная трудоемкость, критерий согласия Пирсона, метод моментов, трудоемкость алгоритма.

Введение

Вопросы построения комплексных критериев оценки качества алгоритмов являются важными при решении задачи выбора рациональных алгоритмов в рамках разработки алгоритмического обеспечения программных систем. Одним из важных критериев оценки алгоритма до сих пор остается временная эффективность, что связано с практической необходимостью решения задач большой размерности и возрастающими требованиями к системам реального времени. При этом в обоих случаях рост размерности задач не компенсируется ростом вычислительной мощности современных компьютеров, что приводит к необходимости построения комбинированных эффективных алгоритмов и прогнозирования времени их выполнения.

Для теоретического решения этих задач при оценке временной эффективности в диапазоне реальных длин входов, определяемых областью применения программной системы, необходимо знание точного числа операций, задаваемых алгоритмом, т.е. его функции трудоемкости. Отметим в связи с этим, что достаточно часто алгоритмы, имеющие асимптотически оптимальную оценку вычислительной сложности, не всегда могут быть эффективны на реальном диапазоне длин входов, что объясняется большими коэффициентами у компонент функции трудоемкости.

Практически значимыми результатами анализа некоторого алгоритма является получение таких сведений, которые могли бы дать возможность прогнозирования ресурсных затрат, требуемых этим алгоритмом при решении задач из данной проблемной области. Идеальным результатом для решения задач прогнозирования и сравнительного анализа можно считать получение точной функции трудоемкости алгоритма. Эта функция должна учитывать не только длину входа, но и влияние значений элементов входа на число задаваемых алгоритмом базовых операций в принятой модели вычислений. К сожалению, такая функция может быть получена только для количественно-зависимых алгоритмов, образующих класс N [1], и, может быть, для ряда алгоритмов других классов на его трудоемкость. В связи с этим реальный анализ некоторого алгоритма предполагает по-

лучение функций трудоемкости для лучшего, среднего и худшего случая как функций длины входа. В настоящее время практически наиболее используемой оценкой является функция трудоемкости в среднем, на основе которой с достаточно хорошей точностью могут быть прогнозированы (в статистическом смысле) временные оценки программной реализации алгоритма [2].

Однако проблема состоит в том, что оценка в среднем, являясь статистической и точечной, не позволяет получить какие-либо сведения о поведении алгоритма на конкретных входах, что важно как для задач большой размерности, так и для систем реального времени. Использование оценки в худшем случае приводит к существенному завышению временного прогноза из-за малой вероятности входов, обеспечивающих максимум трудоемкости при фиксированной размерности задачи. Таким образом, интерес представляет задача построения практически значимой интервальной оценки трудоемкости алгоритма, что и составляет предмет исследования настоящей статьи.

Одно из возможных решений этой задачи, предлагаемое авторами, связано с рассмотрением трудоемкости алгоритма при фиксированной длине входа как дискретной ограниченной случайной величины, имеющей некоторое неизвестное распределение. Подход авторов состоит в построении доверительного интервала трудоемкости на основе аппроксимации неизвестного дискретного распределения значений трудоемкости непрерывным распределением с ограниченной вариацией, в качестве которого предлагается использовать бета-распределение. Получаемое решение позволяет с заданной доверительной вероятностью указать более реальную правую границу трудоемкости алгоритма при фиксированной длине входа.

1. Функция трудоемкости алгоритма

Комплексная оценка качества алгоритма предполагает получение количественных оценок компонент, входящих в комплексный критерий. Важной составляющей таких комплексных критериев являются ресурсные характеристики алгоритма – временная и емкостная эффективности. Поскольку объектом исследования является алгоритм, а не его программная реализация, то в качестве ресурсной оценки времени должна выступать оценка, отражающая операционные затраты алгоритма при решении конкретной задачи. Этот подход приводит к необходимости определения элементарных или базовых операций, в которых будут оцениваться эти операционные затраты. Таким образом, теоретическое исследование алгоритма должно опираться на фиксацию абстрактной модели вычислений, обладающую априорными базовыми операциями. При этом подходе исследуется запись алгоритма как последовательность базовых операций принятой модели вычислений, а количественная мера его временной эффективности определяется как число базовых операций, задаваемых алгоритмом на конкретном входе.

Далее будем использовать понятия и обозначения, связанные с оценкой временной эффективности алгоритма в фиксированной модели вычислений. Такой моделью может быть, например, «машина с произвольным доступом к памяти» [3].

Пусть D_A есть множество допустимых конкретных проблем для задачи, решаемой алгоритмом A, элемент этого множества $D \in D_A$ – конкретная проблема, называемая также входом алгоритма.

Под *трудоемкостью алгоритма* A на входе D будем понимать число базовых операций в принятой модели вычислений, задаваемых алгоритмом на этом входе, обозначая ее как функцию от D через $f_A(D)$. Заметим, что функция трудоемкости для любого допустимого входа D является ограниченной целочисленной функцией, отображающей D_A на N, поскольку в силу классического определения по Э. Посту [4], алгоритм A является финитным 1-процессом.

При более детальном анализе ряда алгоритмов оказывается, что не всегда трудоемкость алгоритма на одном входе D длины n, где $n = |D|$, совпадает с его трудоемкостью на другом входе такой же длины. Рассмотрим допустимые входы алгоритма длины n – в общем случае существует подмножество (для боль-

шинства алгоритмов собственное) множества D_A, включающее все входы, имеющие длину n, – обозначим его через D_n: $D_n = \{ D \mid |D| = n \}$. Множество D_n является конечным – это очевидно, если, например, рассматривать n как число бит входа или как число элементов на входе, имеющих фиксированную битовую длину. Обозначим через M мощность этого множества: $M = |D_n|$.

Поскольку конечное множество целых чисел имеет наибольший и наименьший элемент, то алгоритм A, получая различные входы D из множества Dn, будет задавать на каком-то из входов наибольшее, а на каком-то из входов наименьшее число операций. Такие алгоритмы образуют класс с количественно-параметрической трудоемкостью – класс NPR, исключением являются алгоритмы класса N, для которых трудоемкость определяется только длиной входа [1]. Будем использовать далее следующие обозначения, предложенные в [5], для числа операций, задаваемых алгоритмом A на входах длины n как функций длины входа:

$f_A^{\wedge}(n)$ – *худший случай* – наибольшее число операций: $f_A^{\wedge}(n) = \max\limits_{D \in D_n} \{ f_A(D) \}$. Отметим, что в теории сложности вычислений под сложностью алгоритма понимается асимптотическая оценка функции $f_A^{\wedge}(n)$ в оценках O или Θ.

$f_A^{\vee}(n)$ – *лучший случай* – наименьшее число операций: $f_A^{\vee}(n) = \min\limits_{D \in D_n} \{ f_A(D) \}$.

$\overline{f_A}(n)$ – *средний случай* – среднее число операций, определяемое как математическое ожидание трудоемкости на вероятностной мере входов алгоритма:

$$\overline{f_A}(n) = \sum_{D \in D_n} P(D) \cdot f_A(D),$$

где $P(D)$ есть вероятность входа D для анализируемой области применения алгоритма. В случае, если все входы $D \in D_n$ считаются равновероятными, то:

$$\overline{f_A}(n) = \frac{1}{M} \sum_{D \in D_n} f_A(D).$$

2. Трудоемкость как дискретная ограниченная случайная величина

Дополнительную информацию о поведении алгоритма из класса *NPR* можно получить, рассматривая его функцию трудоемкости как дискретную случайную величину F_A, ограниченную минимальным и максимальным значениями. Теоретически функция распределения вероятностей для трудоемкости алгоритма при фиксированной длине входа n может быть получена на основе анализа генеральной совокупности входов – множества D_n, представляющего собой выборочное пространство Ω_n, на котором определена случайная величина F_A. При этом значение случайной величины F_A в точке выборочного пространства ω_i есть трудоемкость алгоритма на входе D_i – $F_A(\omega_i) = f_A(D_i) = f_i$. Вводя обозначение f_A для произвольного значения трудоемкости как реализации случайной величины F_A заметим, что вероятность того или иного значения f_A определяется стандартным образом на выборочном пространстве Ω_n на основе вероятности входов D:

$$P(f_A) = \sum P(D), D : f_A(D) = f_A, D \in D_n.$$

Такой анализ позволяет получить на основе полного эксперимента по всем $D \in D_n$ (гипотетически, поскольку число элементов в D_n астрономически велико) распределение вероятностей для значений функции трудоемкости как дискретной ограниченной случайной величины, причем значения f_A ограничены сегментом: $f_A^{\vee} \le f_A \le f_A^{\wedge}$. При этом, очевидно, выполнено: $\sum P(f_A) = 1$, тогда математическое ожидание трудоемкости задается формулой

$$\mathbf{M}(F_A) = \sum f_A \cdot P(f_A),$$

где обе суммы берутся по всем значениям функции трудоемкости f_A на теоретическом сегменте варьирования $f_A^{\vee} \le f_A \le f_A^{\wedge}$ при фиксированной длине входа n.

Экспериментальное исследование алгоритма с целью определения частотной встречаемости значений трудоемкости состоит в проведении ряда экспериментов с его программной реализацией при фиксированной длине входа. Для каждого эксперимента генерируется случайный допустимый алгоритмом вход и фиксируется число базовых операций. В предположении о том, что для исследуемого алгоритма в теории получены функции трудоемкости для лучшего и худшего случаев, теоретический размах варьирования $f_A^{\wedge} - f_A^{\vee}$ разбивается на выбранное число полусегментов. Далее определяется частотная встречаемость значений трудоемкости в полученных полусегментах и строится экспериментальная гистограмма относительных частот. В качестве примера на Рис. 1 и Рис. 2 показаны ненулевые части гистограммы относительных частот трудоемкости для алгоритма сортировки вставками и алгоритма сортировки поиском минимума при длине входа n =100,

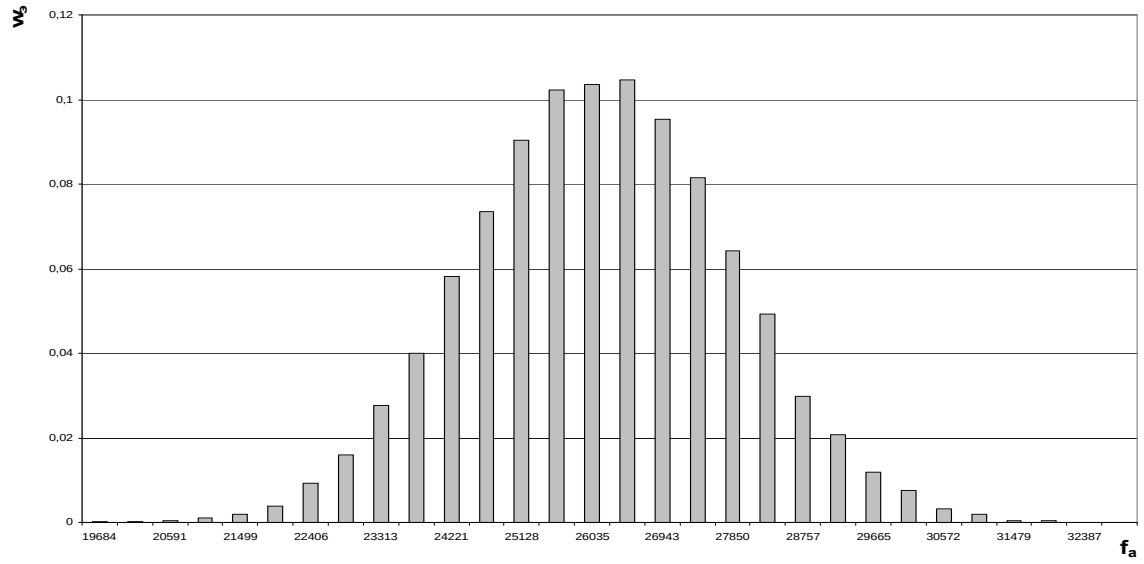

Рис. 1. Ненулевая часть гистограмма относительных частот трудоёмкости для алгоритма сортировки вставками при n =100

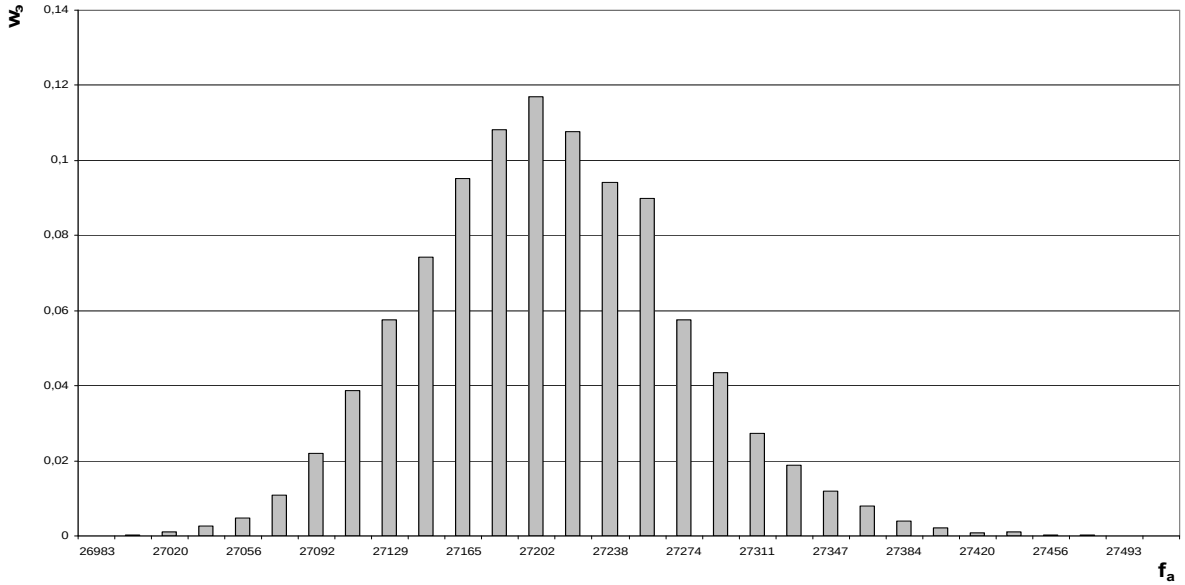

Рис. 2. Ненулевая часть гистограммы относительных частот трудоёмкости для алгоритма сортировки поиском минимума при n =100

полученные по результатам обработки 20000 экспериментов (все экспериментальные исследования и расчеты выполнены А. С. Кривенцовым).

Отметим еще раз, что часто используемая в анализе алгоритмов оценка трудоемкости в среднем (математическое ожидание) не корректна для оценки единичных входов, т. к. возможно наблюдение любого значения трудоемкости в теоретическом диапазоне с определенными, не равными нулю, вероятностями. Вариант оценки по моде также некорректен – знание, что именно это значение трудоемкости встречается наиболее часто, не есть гарантия того, что это значение мы будем наблюдать в конкретном эксперименте. Более того, для несимметричных распределений мода, медиана и математическое ожидание в общем случае не совпадают и могут значительно различаться. Таким образом, точечные оценки трудоемкости как дискретной ограниченной случайной величины – мода, медиана и математическое ожидание не - могут быть использованы как гарантирующие оценки, а очевидно гарантирующая оценка по максимуму – теоретическая трудоемкость в худшем случае – дает слишком завышенные временные прогнозы.

В силу вышесказанного актуальной является задача построения оценки трудоемкости, опирающейся на рассмотрение функции трудоемкости как дискретной ограниченной случайной величины, которая дает практически значимую интервальную оценку для конкретных входов алгоритма.

3. Аппроксимация гистограммы относительных частот трудоемкости бета-распределением

Основной проблемой вероятностного подхода к описанию трудоемкости алгоритма является сложность теоретического доказательства того факта, что значения трудоемкости данного алгоритма имеют определенный закон распределения. Это возможно в некоторых частных случаях, результаты по точечным оценкам математического ожидания и дисперсии для некоторых алгоритмов получены Д. Кнутом в [6]. В общем случае задача может быть поставлена как задача нахождения такого распределения, которое обладало бы определенной гибкостью формы и давало бы приемлемую аппроксимацию наблюдаемой гистограммы относительных частот трудоемкости.

Поскольку значения функции трудоемкости при фиксированной длине входа ограничены: $f_A^{\vee} \le f_A \le f_A^{\wedge}$, то необходимо рассматривать функции распределения, ограниченные на сегменте. С другой стороны, число различных значений трудоемкости достаточно велико, что позволяет перейти к рассмотрению непрерывных распределений, описываемых функциями плотности. Основываясь на результатах, опубликованных в [7], авторы предлагают использовать аппарат бета-распределения, который описывает непрерывную случайную величину, имеющую ограниченный размах варьирования. Кроме того, функция плотности бета-распределения, являясь двухпараметрической, обладает достаточно большой гибкостью формы. Еще одно важное свойство этого распределения – устойчивость [8], т. е. сумма случайных величин, подчиняющихся бета-распределению, также имеет бета-распределение.

Правомочность перехода к непрерывному распределению для описания распределения значений трудоемкости, представляющей собой дискретную ограниченную случайную величину, может быть обоснована значительным числом различных возможных значений трудоемкости на сегменте $\left[f_A^{\vee}, f_A^{\wedge} \right]$ и возможностью проведения достаточно большого числа экспериментов, обеспечивающих репрезентативность такой выборки.

Плотность распределения вероятностей для бета-распределения задается функцией [8]:

$$b(x,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in [0,1], \tag{1}$$

где $\Gamma(\cdot)$ – гамма функция Эйлера, а α и β – параметры функции плотности бета-распределения.

На Рис. 3 и Рис. 4 показаны функции плотности бета-распределения при различных параметрах. Отметим, что если $\alpha = \beta$, то функция плотности симметрична, при $\alpha > 1, \beta > 1$ и $\alpha < \beta$ мода бета-распределения смещена влево,

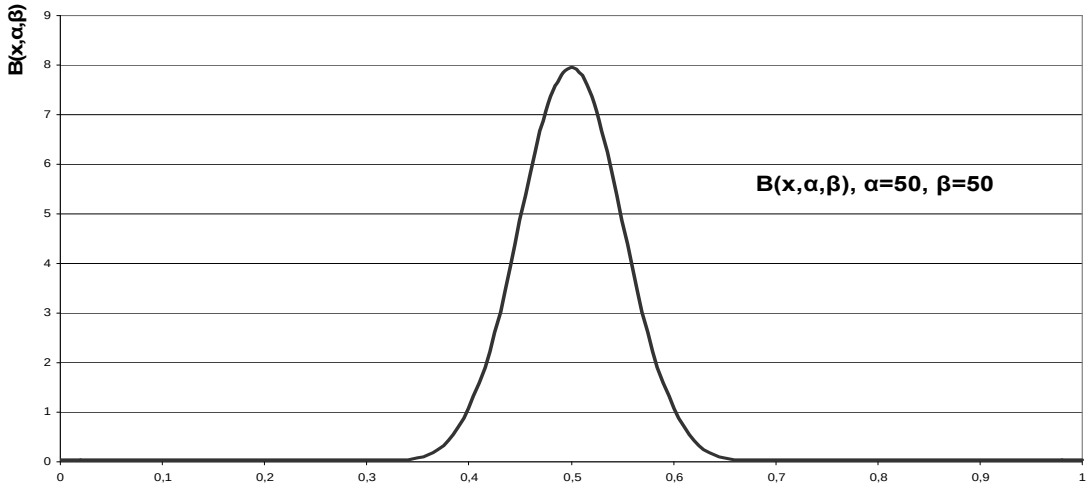

Рис. 3. Симметричная функция плотности бета-распределения ($\alpha = \beta = 50$)

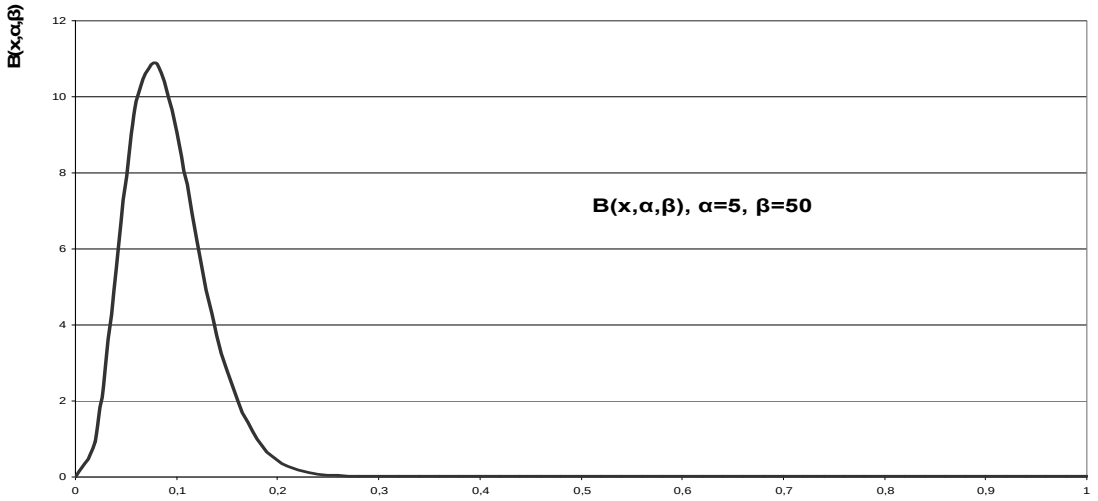

Рис. 4. Функция плотности бета-распределения с модой, смещенной влево ($\alpha < \beta$)

а при $\alpha > 1, \beta > 1$ и $\alpha > \beta$ – смещена вправо, если $\alpha = \beta = 1$, то бета-распределение вырождается в равномерное – $b(x,1,1) \equiv 1$.

Для выборок с произвольными, но заведомо конечными границами диапазона изменений наблюдаемых значений случайной величины, а именно таким свойством и обладает трудоемкость алгоритма при фиксированной длине входа, введем в рассмотрение нормированную случайную величину T, реализации которой t_i получаются на основе значений f_i путем следующего преобразования:

$$t_i = \frac{f_i - f^{\vee}}{f^{\wedge} - f^{\vee}}, \qquad (2)$$

где $f^{\vee}$ и $f^{\wedge}$ – соответственно минимальное и максимальное значение трудоемкости, определенное на основе теоретических функций трудоемкости исследуемого алгоритма для лучшего и худшего случаев при данной длине входа, а f_i – значение трудоемкости в i-ом эксперименте для случайного допустимого входа: $f_i = f_A(D_i)$, $i = \overline{1,m}$, при этом очевидно, что $t_i \in [0,1]$. После нормированмя экспериментальных данных возможно построение гистограммы относительных частот для случайной величины T.

Если параметры аппроксимирующего бета-распределения известны, то путем интегрирования функции плотности бета-распределения

по полусегментам, полученным при построении экспериментальной гистограммы относительных частот для случайной величины T, могут быть получены теоретические относительные частоты бета-распределения и проверена гипотеза о правомерности такой аппроксимации.

4. Восстановление параметров бета-распределения методом моментов

Возможность использования бета-рапределения для аппроксимации гистограммы относительных частот трудоемкости при фиксированной длине входа порождает задачу определения параметров этого аппроксимирующего бета-распределения, т. е. определения параметров бета-распределения распределения случайной величины T на основе результатов выборки.

Дополнительно возникает вопрос об определении необходимого числа экспериментов – m, обеспечивающего заданную точность выборочных значений с заданной доверительной вероятностью. Вариант решения этого вопроса, также опирающийся на аппроксимацию распределения относительных частот трудоемкости бета-распределением, подробно изложен в [7].

Для решения задачи определения параметров аппроксимирующего бета-распределения целесообразно применить метод моментов [9], при этом оценкой математического ожидания является выборочная средняя $\bar{t}$, а оценкой дисперсии – «исправленная» выборочная дисперсия s^2.

Математическое ожидание и дисперсия случайной величины T, имеющей бета-распределение с параметрами α и β, соответственно равны [8]:

$$\mathbf{M}(T)=\frac{\alpha}{\alpha+\beta},\quad D(T)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \quad (3)$$

Применяя метод моментов, получаем систему уравнений для определения параметров аппроксимирующего бета-распределения:

$$\begin{cases}\dfrac{\alpha}{\alpha+\beta}=\bar{t},\\ \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}=s^2,\end{cases}$$

где значения $\bar{t}$ и s^2 определяются на основе преобразований f_i по формуле (2) и имеют вид:

$$\bar{t}=\frac{1}{m}\sum_{i=1}^{m}\frac{f_i-f^{\vee}}{f^{\wedge}-f^{\vee}}=\frac{\overline{f_{\text{э}}}(n)-f^{\vee}}{f^{\wedge}-f^{\vee}},$$

$$s^2=\frac{1}{m-1}\sum_{i=1}^{m}\frac{\left(f_i-f^{\vee}-\overline{f_{\text{э}}}(n)+f^{\vee}\right)^2}{\left(f^{\wedge}-f^{\vee}\right)^2}=$$

$$\cdot=\frac{1}{m-1}\sum_{i=1}^{m}\frac{\left(f_i-\overline{f_{\text{э}}}(n)\right)^2}{\left(f^{\wedge}-f^{\vee}\right)^2}. \quad (4)$$

Решением данной системы, являются следующие формулы обращения

$$\alpha=\frac{\bar{t}}{s^2}\cdot\left(\bar{t}-\left(\bar{t}\right)^2-s^2\right),\quad \beta=\frac{1-\bar{t}}{s^2}\cdot\left(\bar{t}-\left(\bar{t}\right)^2-s^2\right). \quad (5)$$

Таким образом, на основе m экспериментов с программной реализацией алгоритма на входах фиксированной длины n, результатами которых являются значения трудоемкости $f_i, i=\overline{1,m}$, по формулам (4) определяются нормированная выборочная средняя и нормированная исправленная выборочная дисперсия. На их основе по формуле (5) рассчитываются параметры бета-распределения, аппроксимирующего гистограмму относительных частот нормированных значений функции трудоемкости как дискретной ограниченной случайной величины.

5. Проверка гипотезы о бета-распределении

Правомерность использования бета-распределения для аппроксимации гистограммы относительных частот трудоемкости требует подтверждения средствами математической статистики. Выдвигаемая стандартно нулевая гипотеза состоит в том, что бета-распределение не противоречит наблюдаемому в эксперименте распределению относительных частот трудоемкости как случайной величины. Проверка такого рода гипотез осуществляется с использованием критериев согласия.

В данном случае авторы применяют наиболее распространенный критерий, инвариантный к виду закона распределения, – критерий согла-

сия Пирсона [10]. В рассматриваемой задаче выдвигается нулевая гипотеза о том, что нормированная трудоемкость (случайная величина T) имеет бета-распределение.

Пусть сегмент $[0,1]$ варьирования T разбит на s не обязательно равных полусегментов: $[x_i, x_i + \Delta x_i], i = \overline{1,s}$, в которых по результатам m экспериментов определены эмпирические частоты

$$m_i, i = \overline{1,s}, \sum_{i=1}^{s} m_i = m,$$

на основе которых рассчитаны относительные частоты

$$w_i = \frac{m_i}{m}, i = \overline{1,s}, \sum_{i=1}^{s} w_i = 1,$$

а для предполагаемого закона распределения с известными параметрами, в данном случае – бета-распределения, путем интегрирования функции плотности в полусегментах $[x_i, x_i + \Delta x_i], i = \overline{1,s}$ получены теоретические частоты p_i:

$$p_i = \int_{x_i}^{x_i + \Delta x_i} b(x, \alpha, \beta)\, dx. \qquad (6)$$

Тогда в качестве критерия проверки гипотезы принимается случайная величина:

$$\chi^2 = m \cdot \sum_{i=1}^{s} \frac{(w_i - p_i)^2}{p_i}, \qquad (7)$$

имеющая закон распределения χ^2 с k степенями свободы. Число степеней свободы определяется равенством $k = s - 1 - r$, где r – число параметров распределения, в данном случае $r = 2$. Область принятия нулевой гипотезы при правосторонней критической области определяется неравенством $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}(\alpha', k)$, где α' – принятый уровень значимости, значение $\chi^2_{\text{набл}}$ вычисляется по формуле (7), а значение $\chi^2_{\text{кр}}(\alpha', k)$ определяется по теоретическому распределению χ^2. Обычно выбираемое значение уровня значимости составляет $\alpha' = 0{,}05$ [10].

Проверим гипотезу о бета-распределении для алгоритма сортировки вставками при $n = 100$. Функции трудоемкости этого алгоритма для лучшего и худшего случаев, необходимые для нормирования значений имеют вид [1]:

$$f_A^{\wedge}(n) = 5n^2 + 9n - 13,\ f_A^{\vee}(n) = 14n - 13.$$

Сегмент $[0,1]$ был разбит на 75 равных полусегментов, по результатам 20000 экспериментов при $n = 100$ по формуле (2) были рассчитаны нормированные значения t_i, получены эмпирические частоты, выборочная средняя и дисперсия. По формуле (5) рассчитаны параметры аппроксимирующего бета-распределения: $\alpha = \beta = 110{,}3$, на основе которых по формуле (7) вычислены теоретические частоты. Полученные эмпирические и теоретические частоты показаны на Рис. 5. Для большей наглядности вместо наложения гистограмм было выбрано представление в виде кусочно-линейных функций, маркеры кусочно-линейной функции проставлены по серединам интервалов.

Наблюдаемое значение критерия χ^2 рассчитано по формуле (7), критическое значение – стандартной функцией MS Excel. В результате проверки гипотезы получены следующие результаты:

$$\chi^2_{\text{набл}} = 56{,}69,\ \chi^2_{\text{кр}}(0{,}05, 72) = 92{,}80,$$

$$\chi^2_{\text{набл}} < \chi^2_{\text{кр}}(\alpha', k).$$

Таким образом, нет оснований отвергать нулевую гипотезу, и предположение о возможности аппроксимации гистограммы экспериментальных относительных частот нормированной трудоемкости для алгоритма сортировки вставками при $n = 100$ бета-распределением является оправданным.

Аналогичная проверка гипотезы была проведена для алгоритма сортировки поиском минимума при длине входа $n = 50$. Теоретические функции трудоемкости этого алгоритма, используемые для нормирования сегмента варьирования, имеют вид [11]:

$$f_A^{\wedge}(n) = 4n^2 + 11n - 14,$$
$$f_A^{\vee}(n) = 2{,}5n^2 + 12{,}5n - 14.$$

В результате рассчитанные на основе экспериментальных данных по формуле (5) параметры аппроксимирующего бета-распределения

составили: $\alpha = 106{,}76$; $\beta = 902{,}21$, на основе которых по формуле (7) вычислены теоретические частоты. Полученные эмпирические и теоретические частоты показаны на Рис. 6.

В результате проверки гипотезы получены следующие результаты:

$$\chi^2_{\text{набл}} = 38{,}35,\ \chi^2_{\text{кр}}(0{,}05, 72) = 92{,}80,$$

$$\chi^2_{\text{набл}} < \chi^2_{\text{кр}}(\alpha', k).$$

Таким образом, нет оснований отвергать нулевую гипотезу, и предположение о возможности аппроксимации гистограммы экспериментальных относительных частот нормированной трудоемкости алгоритма сортировки поиском минимума при $n = 50$ бета-распределением является оправданным.

6. Понятие доверительной трудоемкости

Основной целью статьи является построение такой интервальной оценки трудоемкости алгоритма, которая, будучи содержательной для

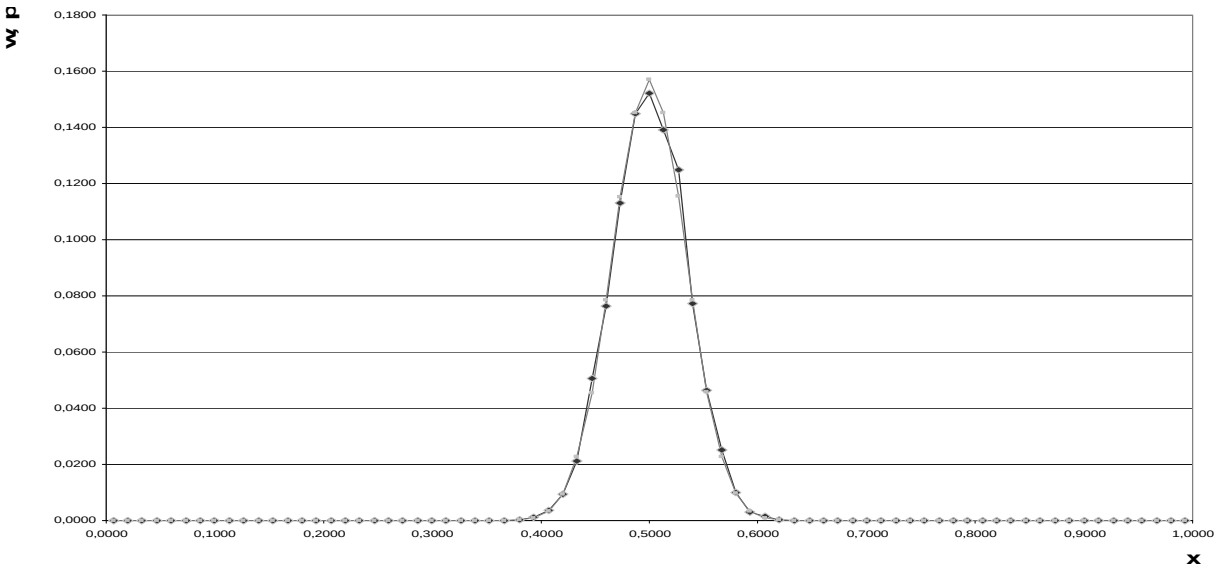

Рис. 5. Теоретические и эмпирические частоты для алгоритма сортировки вставками при *n*=100 с разбиением нормированного сегмента [0,1] на 75 полусегментов

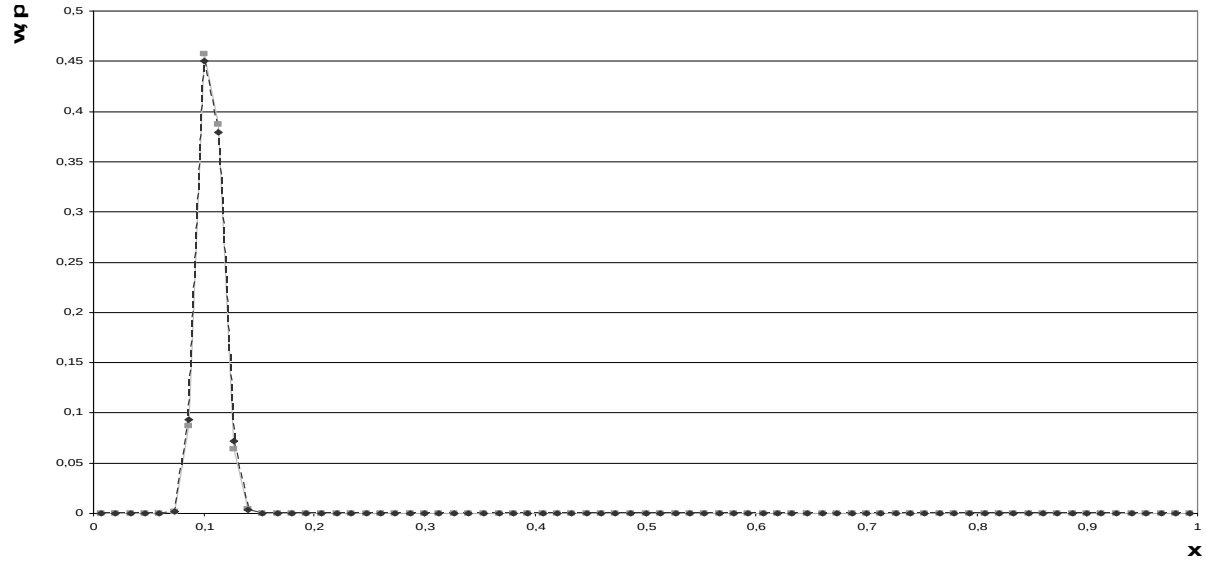

Рис. 6. Теоретические и эмпирические частоты для алгоритма сортировки поиском минимума при n=50 с разбиением нормированного сегмента [0,1] на 75 полусегментов

оценки единичных входов, была бы практически более приемлемой, чем теоретический определенный сегмент варьирования трудоемкости между лучшим и худшим случаями при фиксированной длине входа. Проблема заключается в том, что для большинства алгоритмов значения трудоемкости, достаточно близкие к худшему случаю, имеют незначительную частотную встречаемость (Рис. 1, 2, 5, 6). С другой стороны, практически более важно ограничить значения трудоемкости сверху, чем снизу.

Предлагаемое решение основано на классическом подходе математической статистики, и связано с построением доверительных интервалов оцениваемых величин с заданной доверительной вероятностью [10]. В данном случае оцениваемой величиной является трудоемкость алгоритма при фиксированной длине входа как случайная величина, аппроксимируемая бета-распределением. Если задана доверительная вероятность γ, то путем решения интегрального уравнения возможно определение таких значений пределов интегрирования x_0, x_1, при которых интеграл от функции плотности бета-распределения равен γ. Поскольку нас очевидно устраивают значения трудоемкости, равные или близкие к лучшему случаю, хотя они и могут иметь малые вероятности, то из бесконечного множества решений интегрального уравнения мы выбираем одно, задаваемое пределами $x_0 = 0$, $x_1 = x_\gamma$, т.е. левым γ-квантилем бета-распределения. Переходя от нормированных значений к реальному диапазону варьирования, мы получаем значение трудоемкости, которое не будет превышено для единичного входа с доверительной вероятностью γ, – будем называть это значение *доверительной трудоемкостью* (на уровне γ) и обозначать его через f_γ. Иными словами, для некоторого единичного входа алгоритма трудоемкость будет заключена между лучшим случаем и доверительной трудоемкостью, т.е. в сегменте $\left[f^\vee, f_\gamma\right]$ с вероятностью γ. Например, для приведенной на Рис.4 функции плотности бета-распределения при $\gamma = 0{,}95$ значение $x_\gamma = 0{,}161545$, а для аппроксимирующего бета распределения с параметрами $\alpha = 106{,}76$; $\beta = 902{,}21$, полученными для алгоритма сортировки поиском минимума при $n = 50$ (Рис. 6), при $\gamma = 0{,}95$ значение $x_\gamma = 0{,}122164$. Отметим, что длина сегмента доверительной трудоемкости в первом случае более чем в 6, а во втором – более чем в 8 раз меньше теоретического сегмента варьирования. Именно сокращение длины сегмента для оценки трудоемкости алгоритма на некотором входе фиксированной длины и составляет цель авторов статьи. Эта ситуация проиллюстрирована на Рис. 7.

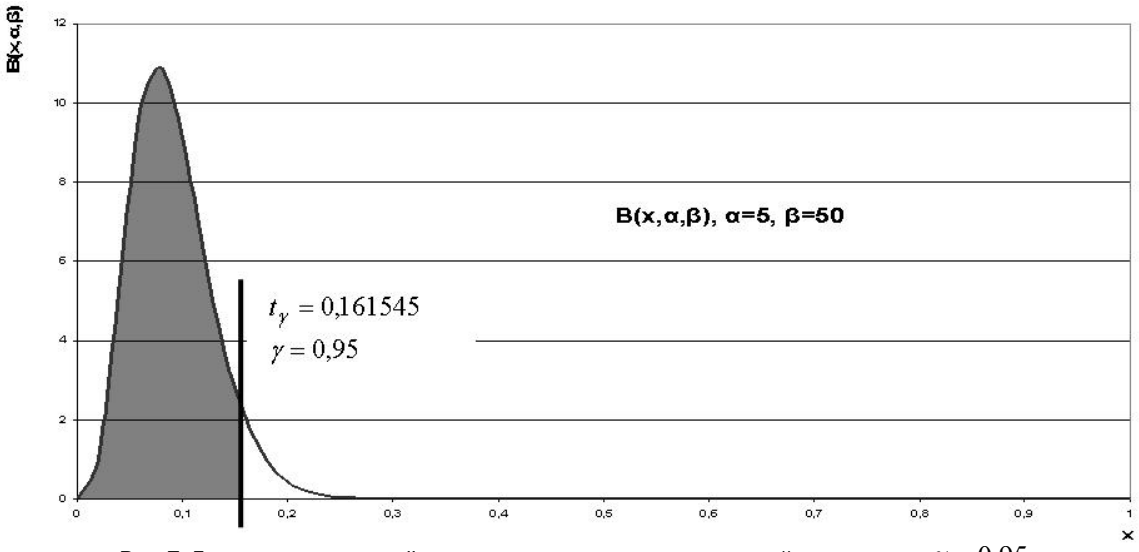

Рис. 7. Доверительная трудоёмкость алгоритма для доверительной вероятности $\gamma = 0{,}95$

Таким образом, если функция $B^{-1}(\gamma,\alpha,\beta)$ есть функция, обратная к интегрированной плотности бета-распределения

$$B^{-1}(\gamma,\alpha,\beta)=s:\ B(s,\alpha,\beta)=\int_0^s b(x,\alpha,\beta)dx=\gamma,$$

то при заданной доверительной вероятности γ значение $x_\gamma=B^{-1}(\gamma,\alpha,\beta)$ и доверительная трудоемкость определяется преобразованием в реальный сегмент варьирования по формуле

$$f_\gamma=f^\vee+x_\gamma\cdot\left(f^\wedge-f^\vee\right). \qquad (8)$$

Сегмент $\left[f^\vee,f_\gamma\right]$ представляет собой интервальную оценку трудоемкости с доверительной вероятностью γ, и прогноз времени выполнения на основе доверительной трудоемкости состоятелен на уровне γ. Для симметричных функций плотности, т.е. при $\alpha=\beta$, значение $x_\gamma>1/2$, однако для больших значений α и β оно ненамного превышает $1/2$. Можно говорить о том, что в этом случае мы имеем почти двукратное сокращение длины оценивающего сегмента. Наибольший эффект от предлагаемого решения будет достигаться для сильно левоассиметричных функций плотности, т. е. при $\alpha<<\beta$. Отметим, что понятие доверительной трудоемкости введено для фиксированной длины входа алгоритма. В целях практического сравнительного анализа алгоритмов необходимо ввести функцию доверительной трудоемкости, аргументом которой является длина входа $f_\gamma(n)$.

7. Прогнозирование дисперсии трудоемкости по экспериментальным данным

Получение значений функции $f_\gamma(n)$ на интересующем разработчика алгоритмического обеспечения сегменте размерностей требует значительных вычислительных затрат, т.к. параметры аппроксимирующего бета-распределения определяются методом моментов на основе экспериментальных данных. Сокращение этих временных затрат может быть достигнуто за счет прогнозирования выборочной дисперсии и выборочной средней функциями регрессии, построенными на основе анализа экспериментальных данных для некоторых длин входа. В качестве примера приведем данные по регрессии значений нормированной выборочной дисперсии (формула (4)), полученные на основе обработки 27 значений длины входа от 100 до 360 с шагом 10 для алгоритма сортировки вставками. Наилучшей в смысле максимума значения R^2 в данном случае является функция вида ae^{-bn}, а уравнение регрессии имеет вид (расчеты выполнены в MS Excel):

$$s^2=0{,}0022\cdot e^{-0{,}0077n}. \qquad (9)$$

Полученное уравнение регрессии приводит к нулевой дисперсии при $n\to\infty$, однако ненулевая асимптота дисперсии при $n\to\infty$ существует, но столь мала, что не определяется статистическими методами.

Полученные результаты показаны на Рис. 8.

Полученное уравнение позволяет прогнозировать значение выборочной дисперсии для больших значений размерности с приемлемой погрешностью, что позволяет при наличии теоретической функции трудоемкости в среднем или уравнения регрессии для выборочного среднего получить параметры аппроксимирующего бета-распределения без проведения экспериментальных исследований путем экстраполяции по длине входа алгоритма.

8. Методика определения функции доверительной трудоемкости

Для определения значений функции доверительной трудоемкости алгоритма $f_\gamma(n)$, аргументом которой является длина входа, с целью последующего прогнозирования его временной эффективности предлагается следующая методика, которая иллюстрируется простым примером определения функции доверительной трудоемкости для алгоритма сортировки вставками.

Методика включает в себя два этапа – этап предварительного исследования, целью которого является проверка гипотезы о законе распределения значений трудоемкости алгоритма как дискретной ограниченной случайной величины, и этап основного исследования, на котором оп-

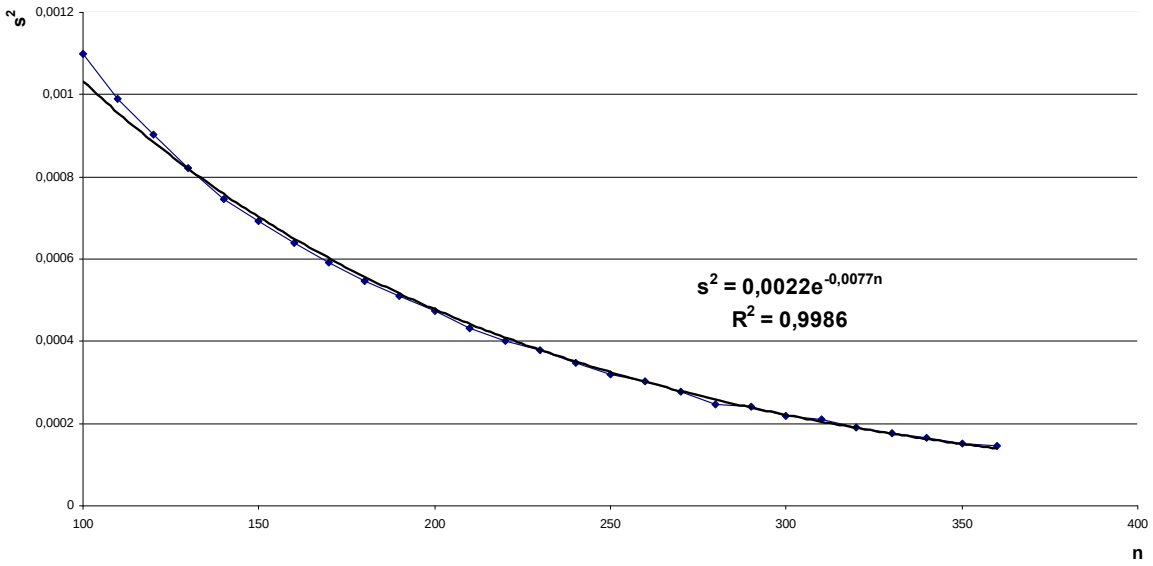

Рис. 8. Экспериментальные данные и уравнение регрессии для выборочной дисперсии значений трудоемкости алгоритма сортировки вставками

ределяются значения доверительной трудоемкости как функции длины входа алгоритма.

Этап предварительного исследования (проверка гипотезы о законе распределения):

1. Фиксация некоторого значения длины входа n из реального сегмента длин в области применения данного алгоритма. В рассматриваемом примере $n = 100$.

2. Определение необходимого числа экспериментов m с программной реализацией для получения гистограммы относительных частот значений трудоемкости. В данном случае $m = 20000$.

3. Проведение экспериментального исследования и получение значений $f_i = f_A(D_i)$, $i = \overline{1,m}$.

4. Получение теоретических функций трудоемкости алгоритма для лучшего и худшего случаев, как функций длины входа. Для алгоритма сортировки вставками эти функции имеют вид: $f_A^{\wedge}(n) = 5n^2 + 9n - 13$, $f_A^{\vee}(n) = 14n - 13$.

5. Выбор числа полусегментов для гистограммы частот значений трудоемкости. В рассматриваемом примере гистограмма строилась на 75 полусегментах.

6. Нормирование значений экспериментальной трудоемкости и построение на основе полученных данных, гистограммы относительных частот в полусегментах (Рис. 5).

7. Вычисление выборочной средней и выборочной дисперсии по формулам (4).

8. Формулировка гипотезы и расчет параметров аппроксимирующего закона распределения. В данном случае выдвигается гипотеза о бета-распределении. Параметры бета-распределения рассчитываются по формулам (5). В данном случае: $\alpha = \beta = 110{,}3$.

9. Расчет теоретических частот по функции плотности, для бета-распределения расчет выполняется по формуле (6) (результаты приведены на Рис. 5).

9. Расчет наблюдаемого значения критерия Пирсона по формуле (7),– $\chi^2_{\text{набл}} = 56{,}69$.

10. Проверка гипотезы о законе распределения. Если нет оснований отвергнуть нулевую гипотезу, то переход к основному этапу исследования. В противном случае – выбор другого закона распределения и повторная проверка гипотезы. В данном примере $\chi^2_{\text{кр}}(0{,}05,72)$=92,80, и нет оснований отвергнуть нулевую гипотезу.

Этап основного исследования:

1. Определение сегмента значений длин входа, соответствующего особенностям применения данного алгоритма в разрабатываемой программной системе. Например, алгоритм сортировки вставками будет применяться для массивов длиной от 100 до 800.

2. Определение сегмента значений длин входа, для которого будут проводиться экспериментальные исследования. В данном примере таким сегментом является сегмент от 100 до 360.

3. Выбор шага изменения длины входа в экспериментальном исследовании. В данном случае значение шага равно 10.

4. Выбор необходимого числа экспериментов с программной реализацией алгоритма для фиксированной длины входа, например, по методике, изложенной в [7], для определения выборочной средней и дисперсии. В данном случае $m = 20000$.

5. Расчет на основе экспериментальных данных значений выборочной средней и дисперсии для каждого значения n. В данном случае n изменяется от 100 до360 с шагом 10.

6. Анализ экспериментальных данных – построение уравнения регрессии для выборочной дисперсии. Результаты показаны на Рис. 8, а уравнение регрессии задается формулой (9).

7. Расчет на основе полученных результатов параметров аппроксимирующего бета-распределения по формулам (5) как функций длины входа – $\alpha(n), \beta(n)$. Для рассматриваемого примера график функции $\alpha(n)$ показан на Рис. 9.

8. Выбор значения доверительной вероятности и вычисление значений левого γ-квантиля бета –распределения: $x_\gamma(n) = B^{-1}(\gamma, \alpha(n), \beta(n))$. В рассматриваемом примере $\gamma = 0{,}95$, $\alpha = \beta$, график значений $x_\gamma(n)$ показан на Рис. 10.

9. Вычисление значений функции доверительной трудоемкости по формуле

$$f_\gamma(n) = f^\vee(n) + x_\gamma(n) \cdot \left(f^\wedge(n) - f^\vee(n)\right)$$

для исследуемого сегмента длин входа. На Рис. 11 показан график значений доверительной трудоемкости и трудоемкости в худшем случае для алгоритма сортировки вставками на сегменте [100,800]. Особо отметим, что доверительная трудоемкость получена для значения доверительной вероятности $\gamma = 0{,}95$, т. е. в 95% случаев по вероятности наблюдаемая в единичном эксперименте трудоемкость алгоритма не будет превышать значение доверительной трудоемкости – для рассматриваемого примера эти значения почти в два раза меньше трудоемкости в худшем случае на всем исследуемом сегменте длин входа.

Заключение

Таким образом, в статье предложен новый подход к оценке качества компьютерных алгоритмов по критерию трудоемкости, основанный на рассмотрении значений трудоемкости алгоритма при фиксированной длине входа как ограниченной случайной величины. Введенная интервальная оценка – довери-

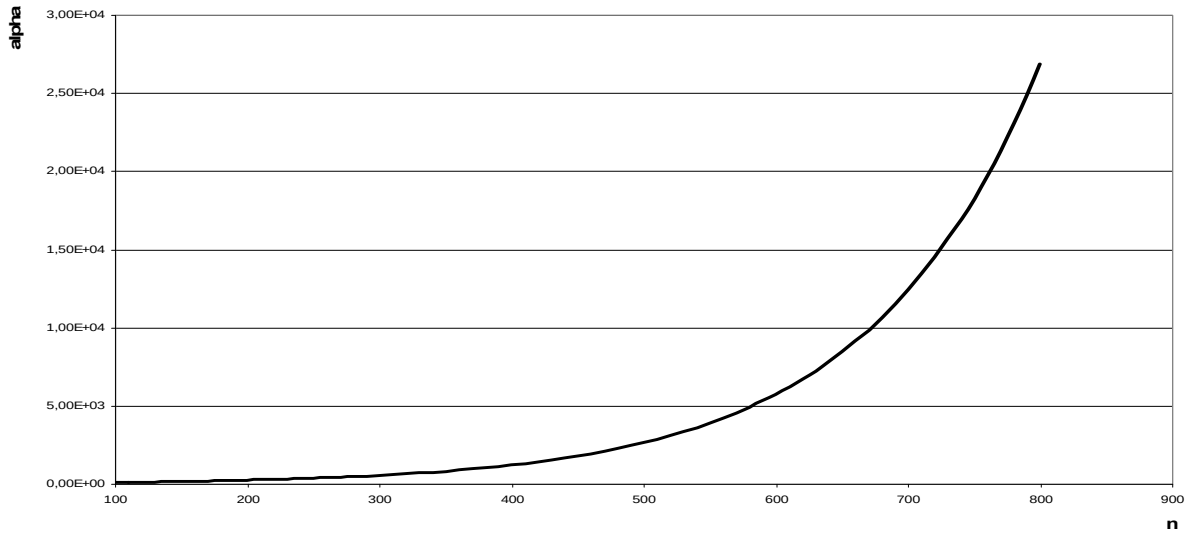

Рис. 9. График функции $\alpha(n)$ — параметра α аппроксимирующего бета-распределения для алгоритма сортировки вставками

тельная трудоемкость – позволяет значительно уменьшить оцениваемый сегмент варьирования трудоемкости при приемлемых значениях доверительной вероятности. С использованием критерия согласия Пирсона показано, что для модельных примеров – двух алгоритмов сортировки – правомерно использовать бета-распределение для аппроксимации распределения частот значений трудоемкости как ограниченной дискретной случайной величины. Сформулирована методика определения доверительной трудоемкости как функции длины входа алгоритма.

Полученные результаты могут быть использованы для повышения достоверности прогнозирования временной эффективности компьютерных алгоритмов и более качественного решения задачи выбора рациональных алгоритмов на основе сравнительного анализа функций доверительной трудоемкости вместо применяемого традиционно сравнения по трудоемкости в среднем случае.

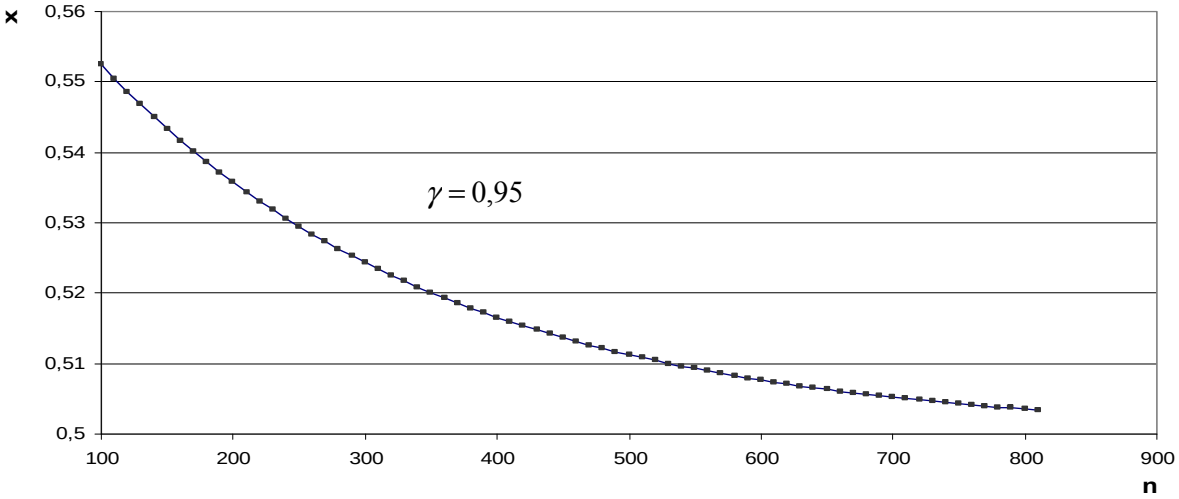

Рис. 10. График зависимости левого γ -квантиля бета –распределения $x_\gamma(n)$ от длины входа для алгоритма сортировки вставками

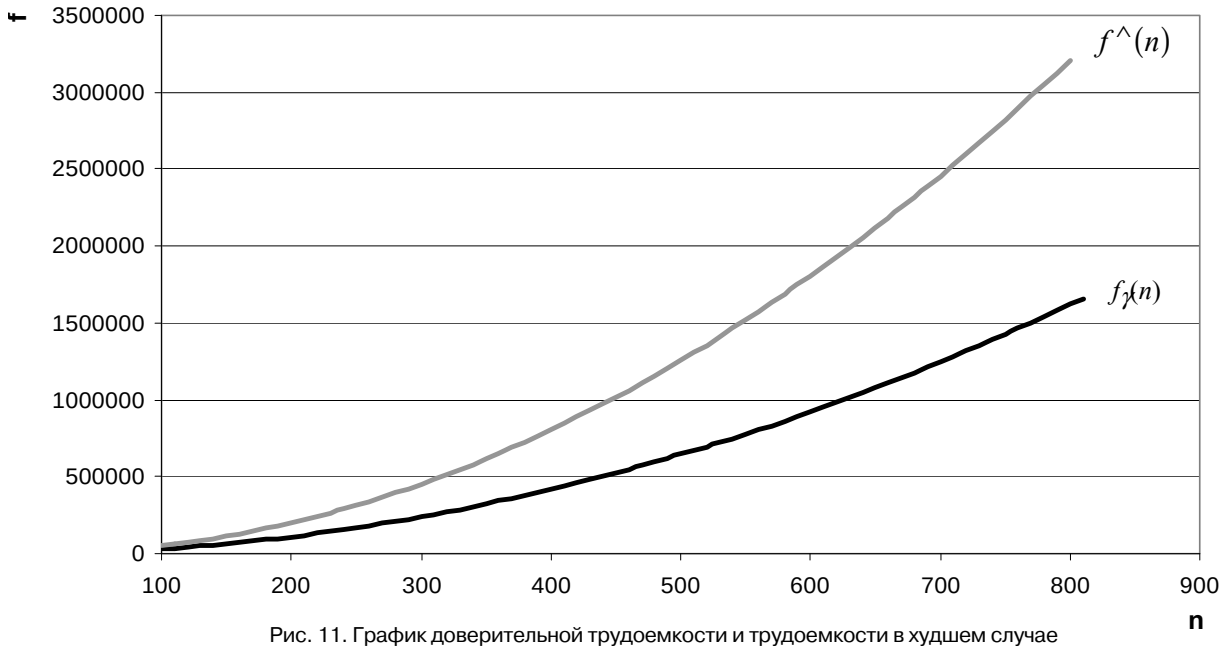

Рис. 11. График доверительной трудоемкости и трудоемкости в худшем случае для алгоритма сортировки вставками

Литература

1. Ульянов М. В. Ресурсно-эффективные компьютерные алгоритмы. Разработка и анализ. – М.: ФИЗМАТЛИТ, 2008. – 304 с.
2. Ульянов М. В. Метод прогнозирования временных оценок программных реализаций алгоритмов на основе функции трудоемкости // Информационные технологии. 2004. № 5. С. 54–62.
3. Алексеев В. Е., Таланов В. А. Графы и алгоритмы. Структуры данных. Модели вычислений – М.: Интернет университет информационных технологий: БИНОМ. Лаборатория знаний, 2006. – 320 с.
4. Успенский В. А. Машина Поста. – М.: Наука, 1979. – 96 с.
5. Ульянов М. В. Система обозначений в анализе ресурсной эффективности вычислительных алгоритмов // Вестник МГАПИ. Серия: Естественные и инженерные науки. 2004 №1(1). С.42–49.
6. Кнут Д. Э. Искусство программирования, том 1. Основные алгоритмы, 3-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2002. – 720 с.
7. Петрушин В. Н., Ульянов М. В. Планирование экспериментального исследования трудоемкости алгоритмов на основе бета-распределения // Информационные технологии и вычислительные системы, 2008. № 2. С. 81–91.
8. Прохоров Ю. В., Розанов Ю. А. Теория вероятностей (Основные понятия. Предельные теоремы. Случайные процессы). – М.: Наука, 1973. – 494 с.
9. Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985.
10. Гмурман В. Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов, – 9-е изд., стер.– М.: Высш. шк., 2003.– 479 с.
11. Ульянов М. В. Классификация и методы сравнительного анализа вычислительных алгоритмов. Научное издание. – М.: Издательство физико-математической литературы, 2004. – 212 с.

Ульянов Михаил Васильевич. Профессор кафедры «Прикладная математика и моделирование систем» Московского государственного университета печати. Окончил Московский институт электронного машиностроения в 1979 году. Доктор технических наук (2005 г.), профессор (2006 г.). Автор более 75 научных работ, в том числе 5 монографий. Область научных интересов: анализ и разработка ресурсно-эффективных компьютерных алгоритмов.

Петрушин Владимир Николаевич. Доцент кафедры «Прикладная математика и моделирование систем» Московского государственного университета печати. Окончил физический факультет Московского университета в 1974 году. Кандидат физико-математичсеких наук (1988 г.), доцент (1991 г.). Автор более 75 научных работ. Область научных интересов: теория вероятностей, математическая статистика, теория эксперимента.

Кривенцов Александр Сергеевич. Студент 5-го курса Московского государственного университета приборостроения и информатики, кафедра «Управление и моделирование систем». Область научных интересов: исследование, анализ и разработка компьютерных алгоритмов.