Планирование экспериментального исследования трудоёмкости алгоритмов на основе бета-распределения

В.Н. Петрушин, М.В. Ульянов

Аннотация. Рассматриваются вопросы, связанные с исследованием трудоёмкости компьютерных алгоритмов. Для подтверждения результатов теоретического анализа алгоритма и прогнозирования времени выполнения его программной реализации необходимо экспериментальное исследование для получения значений функции трудоёмкости в среднем. С целью рационального, в смысле вычислительных затрат, планирования такого экспериментального исследования в статье предлагается использовать аппарат бета-распределения. Приводятся сравнительные результаты классического и предлагаемого подходов.

Введение

Создание алгоритмического обеспечения является важным этапом разработки программных систем, характеризующихся сегодня возрастающей сложностью и размерностью решаемых прикладных задач. В связи с этим требование временной эффективности алгоритмического обеспечения остаётся актуальным, несмотря на рост вычислительной мощности современных компьютеров. Временная эффективность программных реализаций, хотя отчасти и определяется выбором компьютера, операционной системы, языка реализации и тщательностью программирования, но в основном зависит от характеристик используемых алгоритмов. Наиболее употребительными характеристиками эффективности алгоритмов являются оценки временной и емкостной сложности [1], отражающие скорость роста требуемых алгоритмом ресурсов процессора и оперативной памяти. Результатом анализа алгоритма методами теории сложности вычислений [2] является его вычислительная сложность — асимптотическая оценка функции, определяющей число операций, задаваемых алгоритмом, аргументом которой является длина входа. Эта оценка коррелированна с оценкой времени выполнения программной реализации алгоритма. Однако асимптотические оценки указывают не более чем *порядок роста* функции, и результаты сравнения алгоритмов по этим оценкам будут справедливы только при достаточно больших длинах входов. Для сравнения алгоритмов в диапазоне реальных длин входов, определяемых областью применения программной системы, необходимо знание о точном числе операций, задаваемых алгоритмом, т. е. о его функции трудоёмкости. Результаты сравнительного анализа алгоритмов по функциям ресурсной эффективности могут быть использованы при проектировании программных систем на этапе разработки алгоритмического обеспечения для обоснования выбора ресурсно-оптимальных или рациональных алгоритмов в области длин входов, определяемых проблемной областью и соответствующими ограничениями технического задания.

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. Одним из наиболее значимых ресурсов является ресурс

процессора, и в аспекте этого ресурса результатом анализа является возможность прогнозирования времени выполнения [3] или трудоёмкости алгоритма, как для разных размерностей задачи, так и для различных входных данных при фиксированной размерности. Идеальным результатом для решения задач прогнозирования и сравнительного анализа можно считать получение точной функции трудоёмкости 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 [4], и, может быть, для ряда алгоритмов других классов, ввиду сложности формального описания влияния параметрической составляющей входа алгоритма на его трудоёмкость. В связи с этим реальный анализ некоторого алгоритма предполагает получение функций трудоёмкости для лучшего, среднего и худшего случая. Практически наиболее значимым результатом является функция трудоёмкости в среднем, на основе которой с достаточно хорошей точностью могут быть прогнозированы временные оценки программной реализации алгоритма [3].

Результаты теоретического анализа трудоёмкости исследуемого алгоритма должны быть подтверждены экспериментальным исследованием. Такое исследование может преследовать различные цели:

— подтверждение теоретически полученной функции трудоёмкости алгоритма в среднем, как результата его детального анализа методами теории ресурсной эффективности [5];

— получение экспериментальной зависимости трудоёмкости в среднем от длины входа для последующего определения методами регрессионного анализа оценок коэффициентов функциональной зависимости, если на этапе теоретического исследования были получены только асимптотические оценки.

Однако независимо от целей экспериментального исследования при его планировании для получения выборочного среднего значения трудоёмкости возникает задача определения необходимого числа экспериментов при фиксированной длине входа алгоритма (объёма выборки). Серия таких экспериментов для различных длин входов позволяет получить искомые точки экспериментальной зависимости трудоёмкости от длины входа. Временные затраты такого эксперимента очевидно зависят от объёма выборки для получения значимых экспериментальных данных. Классический подход к решению этой задачи, при неизвестной дисперсии, опирается на гипотезу о нормальном распределении и использует распределение Стьюдента для определения необходимого объёма выборки для получения доверительного интервала выборочного среднего при заданном уровне значимости [6]. Однако очевидно, что значения функции трудоёмкости при фиксированной длине входа ограничены как сверху (иначе нарушается свойство финитности алгоритма, и в рамках классической теории алгоритмов такая программа вообще не является алгоритмом!), так и снизу — любой алгоритм задаёт неотрицательное число базовых операций на любом допустимом входе.

Более реалистичный подход к решению этой задачи приводит к рассмотрению трудоёмкости алгоритма при фиксированной длине входа как дискретной ограниченной случайной величины, имеющей некоторое неизвестное распределение. Подход авторов состоит в аппроксимации этого неизвестного дискретного распределения непрерывным распределением, описывающим случайные величины с ограниченной вариацией и имеющим известную и хорошо изученную функцию плотности, а именно, бета-распределением. Изложению этого подхода и посвящена настоящая статья.

1. Терминология и обозначения, связанные с анализом алгоритмов

Будем рассматривать в дальнейшем, придерживаясь терминологии, введенной Э. Постом, алгоритм, как набор инструкций определенной модели вычислений, доставляющий 1-решение общей проблемы (задачи), т. е. применимый к каждой конкретной проблеме финитный 1-процесс, дающий её правильное решение [7]. В качестве модели вычислений будем рассматривать абстрактную машину, включающую процессор, адресную оперативную память и набор базовых операций, соотнесенных с про-

цедурным языком программирования высокого уровня — модели вычислений такого типа носят название «машины с произвольным доступом к памяти» [8].

Пусть D_A есть множество допустимых конкретных проблем для задачи, решаемой алгоритмом A, а его элемент $D \in D_A$ — конкретная проблема, называемая также входом алгоритма A. Под трудоёмкостью алгоритма A на входе D будем понимать число базовых операций в принятой модели вычислений, задаваемых алгоритмом на этом входе, обозначая её, как функцию от D, через $f_A(D)$. Заметим, что функция трудоёмкости для любого допустимого входа D является ограниченной целочисленной функцией целочисленного аргумента, поскольку в силу определения, алгоритм A является финитным 1-процессом, а значение $f_A(D)$ есть число базовых операций. При более детальном анализе ряда алгоритмов оказывается, что не всегда трудоёмкость алгоритма на одном входе D длины m, где $m = |D|$, совпадает с его трудоёмкостью на другом входе такой же длины. Рассмотрим допустимые входы алгоритма длины m — в общем случае существует подмножество (для большинства алгоритмов собственное) множества D_A, включающее входы, имеющие длину m, — обозначим его через D_m: $D_m = \{ D \mid |D| = m \}$. Множество D_m является конечным — обозначим через M мощность этого множества: $M = |D_m|$. Тогда алгоритм A, получая различные входы D из множества D_m, будет, возможно, задавать на каком-то из входов наибольшее, а на каком-то из входов наименьшее число операций. Такие алгоритмы образуют класс с количественно-параметрической трудоёмкостью — класс NPR, исключением являются алгоритмы класса N, для которых трудоёмкость определяется только длиной входа [4]. Будем использовать далее следующие обозначения, предложенные в [9], для числа операций, задаваемых алгоритмом A на входах длины m:

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

$f_A^{\vee}(m)$ — *лучший случай* — наименьшее число операций: $f_A^{\vee}(m) = \min\limits_{D \in D_m} \{ f_A(D) \}$.

$\overline{f_A}(m)$ — *средний случай* — среднее число операций:

$$\overline{f_A}(m) = \sum_{D \in D_m} P(D) \cdot f_A(D),$$

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

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

Поскольку изложение, связанное с планированием экспериментального исследования будет вестись для некоторой фиксированной длины входа m, то в дальнейшем у функций трудоемкости мы опускаем аргумент, т. е. $\overline{f_A} = \overline{f_A}(m)$.

Дополнительную информацию о поведении алгоритма из класса NPR можно получить, рассматривая его функцию трудоёмкости как дискретную случайную величину F_A, ограниченную минимальным и максимальным значениями. Теоретически функция распределения вероятностей для трудоемкости алгоритма при фиксированной длине входа m может быть получена на основе анализа генеральной совокупности входов — множества D_m, представляющего собой выборочное пространство Ω_m, на котором определена случайная величина F_A. При этом значение случайной величины F_A в точке выборочного пространства ω_i есть трудоёмкость алгоритма на входе D_i — $F_A(\omega_i) = f_A(D_i)$. Вводя обозначение f_A для произвольного значения трудоёмкости, как реализации случайной величины F_A, заметим, что вероятность того или иного значения f_A определяется его относительной частотой, т. е. час-

тотной встречаемостью относительно мощности множества D_m с учётом вероятности входов D. Отметим также, что мощность множества D_m, вообще говоря, значительна, например, для задачи сортировки сравнениями, эта мощность определяется числом всех возможных перестановок m различных чисел и составляет $M = m!$. Такой анализ позволяет получить гистограмму функции распределения вероятностей для значений функции трудоёмкости как дискретной ограниченной случайной величины F_A — $P(f_A)$, причём значения f_A ограничены сегментом: $f_A^{\vee} \le f_A \le f_A^{\wedge}$. Возможный вид огибающей такой гистограммы показан на рисунке.

При этом очевидно выполнено: $\sum P(f_A) = 1$, тогда среднее по генеральной совокупности значение трудоёмкости, т. е. её математическое ожидание может быть определено как

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

где обе суммы берутся по всем целочисленным значениям функции трудоёмкости f_A на сегменте $\left[f_A^{\vee}, f_A^{\wedge} \right]$ при фиксированной длине входа m.

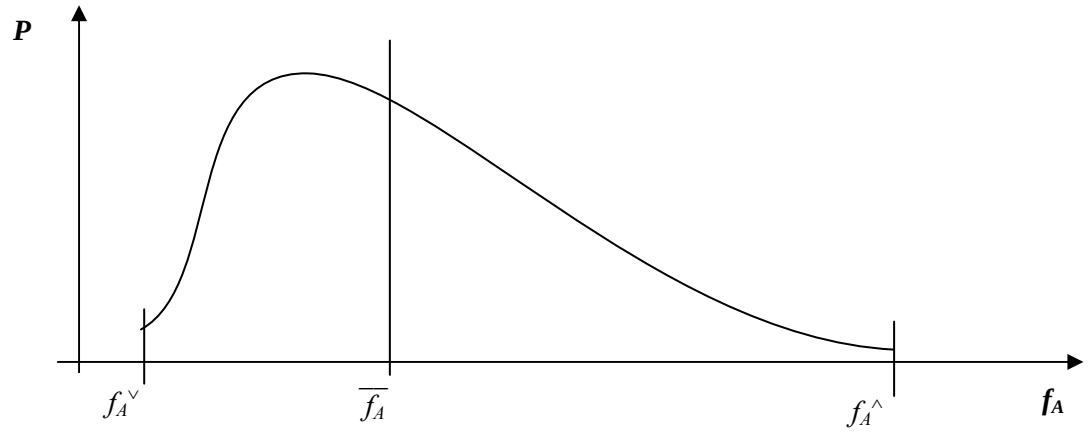
2. Этапы экспериментального исследования трудоёмкости алгоритма

Основной целью экспериментального исследования трудоёмкости алгоритма является получение значений функции трудоёмкости в среднем в зависимости от длины входа. В связи с этим организация экспериментального исследования трудоёмкости алгоритма приводит к необходимости решения следующих задач:

- модификация исходного текста программной реализации алгоритма, связанная с расстановкой счетчика, для определения значения числа выполненных базовых операций при данном входе;
- организация генерации входов алгоритма, обеспечивающая репрезентативность выборки, т. е. генерация входов, соответствующих особенностям применения данного алгоритма в исследуемой программной системе;
- планирование экспериментального исследования, состоящее в определении минимально необходимого объема выборки при фиксированной длине входа, определение сегмента и шага изменения длины.

В части генерации входов алгоритма основной задачей является обеспечение репрезентативности выборки, т. е. генерации таких входов, которые по вероятности соответствуют особенностям применения данного алгоритма в проектируемой программной системе. При этом необходимо учитывать ряд факторов, из которых отметим наиболее значимые:

- соответствие ограничениям задачи. Решаемая исследуемым алгоритмом задача может требовать выполнения ряда специфических ограничений на входные данные. Это могут быть ограничения на знаки чисел, диапазоны их изменения, алфавиты символов для входных



Огибающая гистограммы распределения вероятностей значений трудоёмкости алгоритма

строк и т. д.;

- соответствие области применения. Можно указать два различных подхода для учета данного фактора: либо используются различные реальные входы, либо предположительно известны законы распределения входных значений, отражающие специфику исходных данных проблемной области, что позволяет генерировать соответствующие входные значения. Отметим, что обеспечение соответствия генерируемых входных данных области применения в практическом плане является наиболее сложной задачей;

- особенности алгоритма и задачи, решаемой в данном экспериментальном исследовании. Как пример учета особенности алгоритма укажем на необходимость генерации различных перестановок для соответствующего подмножества входа, если исследуемый алгоритм предположительно относится к подклассу алгоритмов с функцией трудоемкости, зависящей от порядка расположения элементов в однородных массивах.

Рассмотрим более подробно этап планирования экспериментального исследования. Конкретные значения границ для области размерностей задачи определяются особенностями той проблемной области, в которой будет применяться проектируемая программная система, и, как правило, могут быть указаны пользователем. Значение шага определяется как целями исследования, так и границами области. При выборе шага так же необходимо учитывать временные затраты на полный эксперимент, особенно если область реальных размерностей значительна. Для проверки теоретической функции трудоемкости значение шага может быть выбрано в пределах от 0,05 до 0,01 длины исследуемого сегмента размерностей.

Основной задачей данного этапа, варианту решения которой и посвящена настоящая статья, является определение рационального объема выборки при фиксированной длине входа.

3. Постановка задачи

Объектом экспериментального исследования является трудоёмкость алгоритма A на случайно генерируемых входах D длины m. Закон распределения значений $f_A(D)$ неизвестен, но известно, что значения трудоёмкости являются дискретной ограниченной случайной величиной на конечном выборочном пространстве $\Omega_m = D_m$. Обозначим через n объём экспериментальной выборки — число генерируемых входов $D \in D_m$. Отметим, что поскольку эксперименты проводятся с программной реализацией алгоритма, то значение n может быть установлено пользователем. На каждом входе D_i, $i = \overline{1,n}$ регистрируется реализация случайной величины F_A — значение трудоёмкости $f_A(D_i)$, которое далее, как это принято в математической статистике, будем обозначать через f_i. На основе проведенных экспериментов по выборке рассчитываются выборочное среднее $\overline{f_э}$ и выборочная «исправленная» дисперсия S^2 для трудоёмкости алгоритма A на входах длины m. Выборочное среднее $\overline{f_э}$ есть функция объёма выборки n — $\overline{f_э} = \overline{f_э}(n)$ и является оценкой $\overline{f_A}$ — теоретической трудоёмкости алгоритма в среднем. Выборочная «исправленная» дисперсия S^2 есть оценка теоретической дисперсии трудоёмкости ${\sigma_A}^2$.

Постановка задачи: Пусть задан доверительный интервал δ и доверительная вероятность (надёжность) γ для оценки $\overline{f_A}$. Определить

$$n^* = n^*(\delta,\gamma) = \min n : P\left(\left| \overline{f_э}(n) - \overline{f_A} \right| \le \delta \right) \le \gamma . \quad (1)$$

Содержательно мы хотим определить минимальное число экспериментов (объём выборки) определения трудоёмкости алгоритма при фиксированной длине входа m, среднее значение которых $\overline{f_э}$ позволяет построить доверительный интервал длиной 2δ, который покрывает неизвестное значение $\overline{f_A}$ с надёжностью γ.

4. Решение на основе нормального распределения

Классический подход математической статистики основан на центральной предельной теореме, в силу которой при большом числе

слагаемых распределение средних суммы неограниченных случайных величин с любыми законами распределения, имеющими математическое ожидание a и дисперсию σ, и удовлетворяющих условиям Ляпунова, стремиться к нормальному распределению с математическим ожиданием a и дисперсией σ/n [11]. В силу этого для решения задачи о минимальном объеме выборки будем использовать результаты, полученные в математической статистике для доверительного интервала выборочной средней при неизвестной дисперсии для нормального закона распределения.

Рассмотрим случайную величину R, построенную на основе $\overline{f_э}(n)$, которая является случайной величиной для каждой реализации выборки объёма n:

$$r = \sqrt{n}\,\frac{\left(\overline{f_э}(n) - \overline{f_A}\right)}{S},$$

где S^2 — выборочная «исправленная» дисперсия. Доказано [11], что случайная величина R имеет распределение Стьюдента с $n-1$ степенями свободы. Обозначим через $R^{-1}(\gamma, n)$ функцию, обратную к закону распределения Стьюдента, и запишем квантиль r_γ в виде: $P\left(|r| < r_\gamma\right) = \gamma$, тогда $r_\gamma = R^{-1}(\gamma, n)$. Таким образом

$$P\left(\left|\overline{f_э}(n) - \overline{f_A}\right| \le \frac{S}{\sqrt{n}} r_\gamma \right) = \gamma,$$

и мы получаем доверительный интервал $\left(\overline{f_э}(n) - \delta, \overline{f_э}(n) + \delta\right)$ для $\overline{f_A}$, где $\delta = S \cdot r_\gamma / \sqrt{n}$, что позволяет получить значение объёма выборки для заданных δ и γ, которое определяется формулой:

$$n^* = \left\lceil \left(\frac{S}{\delta} \cdot R^{-1}(\gamma, n) \right)^2 \right\rceil. \qquad (2)$$

При $n > 50$ функция $R^{-1}(\gamma, n)$ очень слабо зависит от n, и можно считать, что

$$R^{-1}(\gamma, n) \approx R^{-1}(\gamma),$$

для значения $\gamma = 0{,}95$ и $n > 50$ значение $R^{-1}(\gamma) \approx 1{,}96$ [11]. Поскольку дисперсия и математическое ожидание функции трудоемкости являются функциями длины входа, и, следовательно, будут изменяться при различных значениях m в области исследуемого сегмента длин, то практически более удобно перейти в формуле (2) к относительным единицам. Переходя к относительной точности оценки выборочного среднего — ε, и замечая, что отношение выборочной дисперсии к выборочному среднему есть выборочный коэффициент вариации V_f, получаем формулу, приемлемую для практических расчетов (при $\gamma = 0{,}95$ и $n > 50$):

$$n^* = \frac{3{,}8416 \cdot {V_f}^2}{\varepsilon^2} \qquad (3)$$

Значение выборочного коэффициента вариации, очевидно, есть функция объема выборки. В связи с этим единственно приемлемым является метод последовательного планирования эксперимента, основанный на предварительной выборке некоторого объема, по которой вычисляется выборочное среднее, выборочная дисперсия и коэффициент вариации, с проведением последующих экспериментов для получения выборки рассчитанного объёма, после обработки которой выполняется проверка условия останова.

Таким образом, вначале значение V_f оценивается на основе предварительного эксперимента, т. е. выборки заранее определённого объёма k, например $k = 200$. Затем по формуле (3) рассчитывается значение $n^*_{(1)}$, после чего проводится $n^*_{(1)}$ экспериментов с программной реализацией алгоритма и определяется новое выборочное значение V_f и рассчитывается $n^*_{(2)}$. Если полученный объём выборки $n^*_{(2)} < n^*_{(1)}$, то эксперимент останавливается, иначе, выполняется очередная итерация до выполнения указанного условия. При выполнении условия останова $n^*_{(2)} < n^*_{(1)}$ значение ε может быть скорректировано на основе нового коэффициента вариации и предыдущего (большего) значения объёма выборки.

5. Решение на основе бета-распределения

Поскольку значения функции трудоёмкости при фиксированной длине входа ограничены сегментом $\left[f_A^{\vee}, f_A^{\wedge} \right]$, применение аппарата нормального распределения не является правомочным, и, скорее всего, приводит к завышению необходимого объёма выборки. В связи с этим авторы считают целесообразным применение аппарата, более соответствующего наблюдаемому распределению. В данном случае предлагается использовать аппарат бета-распределения, который описывает непрерывную случайную величину, имеющую ограниченный размах варьирования. Кроме того, бета-распределение обладает достаточно большой гибкостью, в смысле формы функции плотности, определяемой параметрами бета-распределения. Ещё одно важное свойство этого распределения — устойчивость [12], т. е. сумма случайных величин, подчиняющихся бета-распределению, также имеет бета-распределение. Правомочность перехода к непрерывному распределению для описания распределения значений трудоёмкости, представляющей собой дискретную ограниченную случайную величину, может быть обоснована значительным числом различных возможных значений трудоёмкости на сегменте $\left[f_A^{\vee}, f_A^{\wedge} \right]$, и возможностью проведения достаточно большого числа экспериментов, обеспечивающих репрезентативность такой выборки. Плотность распределения вероятностей бета-распределения задаётся функцией [12]:

$$B(t,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}t^{\alpha-1}(1-t)^{\beta-1},\quad t\in[0,1],\quad(4)$$

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

Для выборок с произвольными, но заведомо конечными границами диапазона изменений наблюдаемых значений случайной величины, а именно таким свойством и обладает трудоёмкость алгоритма, введём в рассмотрение нормированную случайную величину T, реализации которой t_i получаются на основе значений f_i путём следующего преобразования:

$$t_i=\frac{f_i-f^{\vee}}{f^{\wedge}-f^{\vee}},\qquad(5)$$

где $f^{\vee}$ и $f^{\wedge}$ — соответственно минимальное и максимальное значение трудоёмкости, определённое на основе теоретических функций трудоёмкости исследуемого алгоритма для лучшего и худшего случаев, а f_i — значение трудоёмкости в i-ом эксперименте $i=\overline{1,n}$. При этом очевидно, что $t_i\in[0,1]$.

Математическое ожидание и дисперсия случайной величины T, имеющей бета-распределение с параметрами α и β, соответственно равны [12]:

$$\mathbf{M}(T)=\frac{\alpha}{\alpha+\beta},\quad D(T)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.\quad(6)$$

Средние величины, наблюдаемые для выборки, извлечённой из генеральной совокупности, имеющей бета-распределение, также принадлежат сегменту $[0,1]$ и имеют бета-распределение в силу его устойчивости. Составим бета-распределение для средних значений случайной величины T — $\overline{T}$, для чего необходимо определить параметры плотности этого распределения, которые будем обозначать далее через $u(n)$ и $v(n)$, поскольку значения этих параметров определяются объёмом выборки. Выразим параметры $u(n)$ и $v(n)$ через параметры плотности распределения исходной случайной величины T — α и β. Из математической статистики известно [12], что

$$\mathbf{M}\left(\overline{T}\right)=\mathbf{M}(T),\quad D\left(\overline{T}\right)=\frac{D(T)}{n},\qquad(7)$$

где n — объем выборки. Для решения этой задачи на основании формул (6) и (7) составим следующую систему уравнений

$$\begin{cases}\dfrac{u}{u+v}=\dfrac{\alpha}{\alpha+\beta},\\[2ex]\dfrac{uv}{(u+v)^2(u+v+1)}=\dfrac{1}{n}\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.\end{cases}$$

Решая эту систему получаем

$$u(n)=\frac{\alpha}{(\alpha+\beta)}\left(n(\alpha+\beta+1)-1\right),$$

$$v(n)=\frac{\beta}{(\alpha+\beta)}\left(n(\alpha+\beta+1)-1\right). \qquad (8)$$

Подстановка $n=1$ в (8) даёт $n=1, u(1)=\alpha, v(1)=\beta$, т. е. распределение средних $\overline{T}$, совпадает с исходным распределением T, что очевидно и должно быть. Согласно центральной предельной теореме [12] при $n\to\infty$ распределение средних должно стремиться к $\mathbf{M}(T)$, а коэффициенты асимметрии и эксцесса $a(n), e(n)$, как функции объема выборки n, должны стремиться к нулю. Проверим выполнение этих условий. Согласно [13], бета-распределение имеет экстремум в точке

$$t_{\text{экстр}}(\alpha,\beta)=\frac{\alpha-1}{\alpha+\beta-2},$$

подставляя значения параметров для $\overline{T}$ из (8) и переходя к пределу, имеем

$$\lim_{n\to\infty} t_{\text{экстр}}(n)=\frac{u(n)-1}{u(n)+v(n)-2}=$$

$$=\lim_{n\to\infty}\frac{(\alpha/(\alpha+\beta))\cdot(n(\alpha+\beta+1)-1)-1}{n(\alpha+\beta+1)-3}=$$

$$=\frac{\alpha}{\alpha+\beta}=\mathbf{M}(T)$$

аналогично, используя формулы для асимметрии и эксцесса [13] получаем

$$\lim_{n\to\infty} a(n)=\lim_{n\to\infty}\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{\sqrt{\alpha\beta}\cdot\left(1/\sqrt{n}+(\alpha+\beta+1)\sqrt{n}\right)}=0,$$

$$\lim_{n\to\infty} e(n)=$$

$$=\lim_{n\to\infty}\frac{6(\alpha-\beta)^2(\alpha+\beta+1)-\alpha\beta((\alpha+\beta+1)+1/n)}{\alpha\beta((\alpha+\beta+1)+1/n)\cdot(n(\alpha+\beta+1)+2)}=0$$

Таким образом, проведённые проверки показали, что центральная предельная теорема выполняется для сумм случайных величин, имеющих бета-распределение.

Следующей задачей является определение параметров исходного бета-распределения, т. е. параметров распределения случайной величины T на основе результатов выборки. Для решения этой задачи целесообразно применить метод моментов [13]. При этом оценкой математического ожидания является выборочная средняя $\bar{t}$, а оценкой дисперсии — «исправленная» выборочная дисперсия s^2. Мы используем обозначение s^2 для «исправленной» выборочной дисперсии нормированной случайной величины T, в отличие от обозначения S^2 для «исправленной» выборочной дисперсии трудоёмкости — случайной величины F_A. Применяя метод моментов, получаем систему уравнений для определения параметров исходного 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 преобразованием по формуле (5) и имеют вид:

$$\bar{t}=\frac{1}{n}\sum_{i=1}^{n}\frac{f_i-f^{\vee}}{f^{\wedge}-f^{\vee}}=\frac{\overline{f_{\text{э}}}(n)-f^{\vee}}{f^{\wedge}-f^{\vee}}, \qquad (9)$$

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

$$=\frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(f_i-\overline{f_{\text{э}}}(n)\right)^2}{\left(f^{\wedge}-f^{\vee}\right)^2}. \qquad (10)$$

Решая полученную систему, получаем следующие формулы обращения

$$\alpha=\frac{\bar{t}}{s^2}\left(\bar{t}-\left(\bar{t}\right)^2-s^2\right),\quad \beta=\frac{1-\bar{t}}{s^2}\left(\bar{t}-\left(\bar{t}\right)^2-s^2\right). \qquad (11)$$

На основании полученных результатов предлагается следующее решение исходной задачи об вычислении минимального объёма выборки n^* для построения по выборочному среднему $\overline{f_{\text{э}}}(n)$ доверительного интервала длиной в 2δ для теоретического среднего значения трудоёмкости $\overline{f_A}$ с заданной надёжностью γ: Интервалу $\left(\overline{f_{\text{э}}}(n)-\delta,\overline{f_{\text{э}}}(n)+\delta\right)$ на ос-

новании преобразования (5) соответствует нормированный интервал $\left(\bar{t}-\delta_t, \bar{t}+\delta_t\right)$, причём

$$\bar{t} \pm \delta_t = \frac{\overline{f_э}(n) - f^{\vee} \pm \delta}{f^{\wedge} - f^{\vee}}, \qquad (12)$$

поскольку параметры бета-распределения для средних значений трудоёмкости выражаются через параметры распределения нормированной трудоёмкости по формуле (8) и являются функциями объёма выборки, то решение исходной задачи сводится к решению интегрального уравнения

$$n^* = n^*(\delta,\gamma): \quad P\left(\left|\overline{f_э}(n^*) - \overline{f_A}\right| \le \delta\right) = \gamma \Rightarrow \quad n^* : P(n^*) \ge \gamma$$

где

$$P(n) = \int_{\bar{t}-\delta_t}^{\bar{t}+\delta_t} B(x, u(n), v(n))\,dx, \qquad (13)$$

причём пределы интегрирования $\left(\bar{t}-\delta_t, \bar{t}+\delta_t\right)$ вычисляются по формуле (10) при известном значении δ, параметры бета-распределения случайной величины $\overline{T}$ — $u(n), v(n)$, как функции объёма выборки, определяются в соответствии с формулой (8), а параметры α, β распределения нормированной функции трудоёмкости — случайной величины T вычисляются на основе экспериментальных данных методом моментов по формуле (9).

6. Методика планирования экспериментального исследования трудоёмкости в среднем

На основании полученных результатов предлагается следующая методика планирования экспериментального исследования трудоёмкости алгоритма в среднем при фиксированной длине входа m, которая базируется на методе последовательного планирования эксперимента, и учитывает, что параметры исходного бета-распределения нормированной трудоёмкости α, β также являются функциями объёма выборки, а значения δ и γ являются заданными. Методика включает в себя следующие этапы:

1. *Предварительный эксперимент*: получение выборки $f_i, i = \overline{1,k}$ заранее определённого объёма k, например $k = 200$ и определение $n^*_{(1)}$.

Вычисление значения $n^*_{(1)}$ включает в себя следующие шаги:

1.1. Нормирование выборочных значений трудоёмкости по формуле (5), включая определение максимального и минимального выборочных значений или по формуле (5^*), если известны теоретические функции для лучшего и худшего случая трудоемкости — получение значений t_i.

1.2. Вычисление выборочной средней $\bar{t}$ и «исправленной» выборочной дисперсии s^2 для нормированной трудоёмкости по формулам (9) и (10).

1.3. Определение параметров α, β бета-распределения нормированной трудоёмкости методом моментов на основе $\bar{t}$ и s^2 по формуле обращения (11).

1.4. Вычисление пределов интегрирования $\left(\bar{t}-\delta_t, \bar{t}+\delta_t\right)$ для заданного значения δ по формуле (12).

1.5. Решение интегрального уравнения (13) путем последовательного увеличения предполагаемого объёма выборки n, до выполнения условия $P(n) \ge \gamma$, причём параметры $u(n), v(n)$ бета-распределения средних значений, как функции объёма выборки определяются в соответствии с формулой (8). Полученное значение $n = n^*_{(1)}$ является оценкой необходимого объёма выборки.

2. *Повторный эксперимент*: получение выборки объёма $n^*_{(1)}$: $f_i, i = \overline{1, n^*_{(1)}}$, выполнение шагов 1.1. – 1.5 для обработки новой выборки, и вычисление на этой основе значения $n^*_{(2)}$.

3. *Проверка условия останова*. Если для полученного объёма выборки выполняется условие $n^*_{(2)} < n^*_{(1)}$, то эксперимент останавливается и $n^* = n^*_{(2)}$, иначе полагаем $n^*_{(1)} = n^*_{(2)}$, и выполняется очередная итерация этапа 2 до выполнения указанного условия останова.

7. Результаты экспериментальных исследований

В качестве примера планирования экспериментального исследования алгоритма приведём экспериментальные данные для алгоритма сортировки вставками, теоретические функции трудоемкости которого получены в [5]:

$$\overline{f_A}(m) = 2{,}5m^2 + 11{,}5m - 13,$$

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

Цель эксперимента состоит в подтверждении полученного теоретического результата для трудоёмкости в среднем, с надёжностью $\gamma = 0{,}95$ и относительной ошибкой $\varepsilon = 0{,}001$ для длины входного массива $m = 100$.

Решение на основе нормального распределения. Для определения n^* было проведено предварительное исследование алгоритма с объёмом выборки равным 200, вычислен коэффициент вариации V_f, и по формуле (3) вычислено значение $n^*_{(1)}$, результаты приведены в Табл. 1.

Табл. 1.

Предварительный объём выборки	**200**
Выборочное среднее	26 081,65
Выборочная дисперсия	3 309 876,46
Выборочная сигма	1 819,31
Коэффициент вариации	0,069754
Рассчитанный объём выборки	**18 692**

Далее была извлечена выборка объёмом 18692, результаты её обработки приведены в Табл. 2. Поскольку рассчитанный объём выборки $n^*_{(2)}$ оказался меньше, чем объём выборки текущего эксперимента, то $n^* = 15821$.

Табл. 2.

Текущий объём выборки	**18 692**
Выборочное среднее	26 139,97
Выборочная дисперсия	2 814 076,73
Выборочная сигма	1 677,52
Коэффициент вариации	0,064175
Рассчитанный объём выборки	**15 821**

Решение на основе бета-распределения. Для определения n^* также было проведено предварительное исследование алгоритма с объёмом выборки равной 200. Поскольку для исследуемого алгоритма в теории получены функции трудоёмкости для худшего и лучшего случаев, нормированные значения трудоёмкости рассчитаны по формуле (5). Значение $n^*_{(1)}$ получено с применением предложенной методики путём решения интегрального уравнения (13), результаты приведены в Табл. 3.

Табл. 3.

Предварительный объём выборки	**200**
Выборочное среднее	26 081,65
Выборочная дисперсия	3 309 876,46
Теоретический минимум	1 387
Теоретический максимум	50 887
Размах	49 500
Нормированное среднее	0,498882
Нормированная дисперсия	0,001351
Альфа	91,829267
Бетта	92,240915
Дельта для eps=0,001	26,081650
Нормированное дельта	0,000527
Нижний предел интегрирования	0,498355
Верхний предел интегрирования	0,504345
Рассчитанный объём выборки	**13 165**

На следующем этапе была извлечена выборка объёмом 13165, результаты её обработки по предложенной методике приведены в Табл. 4. Поскольку рассчитанный объём выборки оказался также меньше, чем объём выборки текущего эксперимента, то $n^* = 11191$.

Табл. 4.

Текущий объём выборки	**13 165**
Выборочное среднее	26 134,57
Выборочная дисперсия	2 825 096,39
Теоретический минимум	1 387
Теоретический максимум	50 887
Размах	49 500
Нормированное среднее	0,499951
Нормированная дисперсия	0,001153
Альфа	107,903860
Бетта	107,925023
Дельта для eps=0,001	26,134573
Нормированное дельта	0,000528
Нижний предел интегрирования	0,499423
Верхний предел интегрирования	0,504345
Рассчитанный объём выборки	**11 191**

Заключение

Таким образом, в статье предложена методика планирования экспериментального иссле-

дования трудоёмкости алгоритмов в среднем, основанная на рассмотрении значений трудоёмкости алгоритма как реализаций дискретной ограниченной случайной величины, аппроксимируемой бета-распределением. Получено интегральное уравнение, решение которого даёт необходимый объём выборки для получения заданного доверительного интервала генеральной средней (теоретической трудоёмкости в среднем) с заданной надёжностью.

Предложенная методика планирования позволила для модельной задачи сократить необходимый объем выборки более чем в 1,4 раза. По результатам предварительного планирования это сокращение составило 5527, а в результате окончательно определения значения объёма выборки — 4630 экспериментов. Отметим, что поскольку предлагаемая методика использует непрерывную функцию плотности, ограниченную на сегменте, то с увеличением доверительной вероятности γ абсолютное сокращение необходимого объёма выборки будет возрастать. Например, при $\gamma = 0{,}99$ такое сокращение составляет уже 4985 экспериментов для окончательного объёма выборки.

Полученные результаты позволяют говорить о том, что применение предложенной методики сокращает временные затраты на проведение экспериментальных исследований трудоёмкости алгоритмов в среднем.

Литература

1. Кормен Т., Лейзерсон Ч., Ривест Р. Алгоритмы: построение и анализ. — М.: МЦНМО, 1999. – 960 с., 263 ил.
2. Кузюрин Н.Н., Фомин С.А. Сложность комбинаторных алгоритмов // http://www.csin.ru/courses/slozhnost-kombinatornykh-algoritmov
3. Ульянов М.В. Метод прогнозирования временных оценок программных реализаций алгоритмов на основе функции трудоёмкости // Информационные технологии. 2004. № 5. С. 54–62.
4. Ульянов М.В. Классификация алгоритмов в целях практического анализа // Информационные технологии. 2003. № 11. С. 29–36.
5. Ульянов М. В. Классификация и методы сравнительного анализа вычислительных алгоритмов. Научное издание. — М.: Издательство физико-математической литературы, 2004. — 212 с.
6. Тюрин Ю.Н.. Макаров А.А. Анализ данных на компьютере / Под ред. В.Э. Фигурнова. — М.: ИНФРА-М, 2003. — 544 с.
7. Успенский В. А. Машина Поста. — М.: Наука, 1979. — 96 с.
8. Алексеев В.Е., Таланов В.А. Графы и алгоритмы. Структуры данных. Модели вычислений — М.: Интернет университет информационных технологий: БИНОМ. Лаборатория знаний, 2006. — 320 с.
9. Ульянов М. В. Система обозначений в анализе ресурсной эффективности вычислительных алгоритмов // Вестник МГАПИ. Серия: Естественные и инженерные науки. 2004 №1(1). С.42–49.
10. Ульянов М. В. Исследование и классификация вычислительных алгоритмов на основе чувствительности функции трудоемкости // Системы управления и информационные технологии. 2004. № 4 (16). С. 97–104.
11. Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов, – 9-е изд., стер.— М.: Высш. шк., 2003.— 479 с.
12. Прохоров Ю. В., Розанов Ю. А. Теория вероятностей (Основные понятия. Предельные теоремы. Случайные процессы). — М.: Наука, 1973. — 494 с.
13. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. — М.: Наука, 1985.

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

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), профессор (2006). Автор более 70 научных работ, в том числе 3 монографий. Область научных интересов: анализ и разработка ресурсно-эффективных компьютерных алгоритмов.