ISSN 1814-1196
Научный вестник НГТУ
том 70, № 1, 2018, с. 153–166

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 70, No. 1, 2018, pp. 153–166

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERIN
AND CONTROL

УДК 519.224.22

DOI: 10.17212/1814-1196-2018-1-153-166

Методика расчета распределения вероятностей значений симметричных аддитивно разделяемых статистик, приближенных к их точному распределению*

А.К. МЕЛЬНИКОВ

117587, РФ, г. Москва, Варшавское шоссе, 125, стр. 17, НТЦ ЗАО «ИнформИнвест-Групп», кандидат технических наук, доцент. E-mail: ak@iigroup.ru

В работе в интересах разработки процедур обработки текстов рассматривается возможность использования точных распределений вероятностей значений статистик для построения статистических критериев согласия с равновероятным распределением. Проводится сравнение вычислительной сложности расчета точных распределений методом полного перебора и частотным методом. Показывается, что вычислительная сложность частотного метода расчета точных распределений намного меньше вычислительной сложности метода полного перебора, но и она не позволяет провести за приемлемое время вычисления точных распределений на современных высокопроизводительных вычислительных системах для практически значимых значений параметров текстов даже при кардинальной модернизации вычислительных систем путем применения новейших вычислительных элементов. За счет сужения класса используемых статистик до класса симметричных аддитивно разделяемых статистик проведен выбор направления модернизации частотного метода расчета точных распределений, заключающийся в ограничении перебираемого выборочного пространства. Показана принципиальная возможность применения модернизированного метода в областях значений параметров текстов, где высокая вычислительная сложность частотного метода не позволяет выполнить расчет точных распределений. На основе результатов по оценке вероятности значений статистики максимальной частоты проведена модернизация частотного метода расчета точных распределений, в результате которой разработана методика расчета Δ-точных распределений, которые отличаются от точных распределений не более чем на заранее заданную величину Δ. Описана пошаговая детализация методики расчета Δ-точных распределений, позволяющая применять ее для проведения практических расчетов. Приводятся конкретные результаты по применению методики расчета Δ-точных распределений для значений параметров текстов, расчет точных распределений для которых на современном этапе невозможен из-за его большой вычислительной сложности.

Ключевые слова: вероятность, статистика, критерий, точное распределение, предельное распределение, вычислительная сложность метода, производительность многопроцессорной вычислительной системы.

---

* *Статья получена 02 октября 2017 г.*

ВВЕДЕНИЕ

При построении информационных моделей задач обработки текстов [1] для выделения их из массивов, знаки в которых распределены случайным равновероятным образом, часто используются статистические критерии согласия с равновероятным распределением.

Пусть из некоторого массива, состоящего из M текстов длиной n, содержащих знаки алфавита $A_N = \{a_1, \ldots, a_N\}$ мощностью N,

$$T_{n,N}(j) = \{t_1(j), \ldots, t_n(j)\}, \quad j = \overline{1, M},$$

нужно отобрать тексты, являющиеся реализациями случайных выборок длины n из равновероятного распределения на алфавите мощностью N.

Отбор текстов с равновероятным распределением знаков производится с помощью применения критерия согласия с равновероятным распределением [2], использующим некоторую статистику S_n текста длины n, являющуюся функцией от h_i частот встречаемости знаков (исходов) текста a_i из алфавита A_N мощности N:

$$S_n = f(n, N).$$

Часть ложно отобранных как равновероятные текстов, содержащих неравновероятное распределение знаков, определяет размер применяемого критерия α [3].

Для определения размера критерия α необходимо знать вероятность распределения значений применяемой в критерии статистики S_n:

$$P\{S_n \geq c\},$$

связанного с размером критерия α соотношением [4]:

$$P\{S_n \geq c\} = \alpha.$$

В статистическом критерии согласия могут использоваться как точные распределения значений вероятности (точные распределения) статистики S_n, расчету которых посвящены работы автора [5, 6], так и предельные распределения вероятности значений (предельные распределения) статистики S_n, определяемые свойствами самой функциями f, например, как это показано Хельмертом [7] и Пирсоном [8].

Целью данной работы является выбор направления и обоснование возможности модернизации метода вычисления точных распределений вероятности значений симметричных аддитивно разделимых статистик, позволяющего вычислять распределения вероятности значений указанных статистик, сколь угодно близкие к их точным распределениям, а также разработка методики этого расчета.

1. МЕТОДЫ РАСЧЕТА ТОЧНЫХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ЗНАЧЕНИЙ СТАТИСТИК

В работах автора [5, 6] уже исследовался вопрос расчета точных распределений статистики S_n хи-квадрат χ_n, предложенной в работе [9] и исследуемой Карлом Пирсоном в [8, 10]:

$$\chi_n = \sum_{i=1}^{N} \frac{(h_i - np_i)^2}{np_i},$$

где h_i – частота встречаемости знака (исхода) a_i; n – длина текста (объем выборки); N – число исходов полиномиальной схемы (мощность алфавита A_N) и p_i – вероятность a_i-го исхода.

В работе автора [5] расчеты точных распределений статистики χ_n исследовались в общем случае, без учета свойств класса статистик, к которым она принадлежит. Показано, что вычислительная сложность такого расчета методом полного перебора для параметров текстов (n, N) сопоставима со сложностью перечисления всех текстов $T_{n,N}(j)$ длины n в алфавите мощностью N и оценивается как $O(N^n)$, что не позволяет проводить расчеты за приемлемое время для практических значений параметров.

В работе автора [6] расчеты точных распределений статистики χ_n исследовались с учетом свойств класса симметричных аддитивно разделимых статистик, к которым принадлежит статика χ_n. Показано, что вычислительная сложность такого расчета для параметров текстов (n, N) с использованием частотного метода сопоставима со сложностью перечисления всех решений уравнения

$$h_1 + \ldots + h_N = n \tag{1}$$

в целых неотрицательных числах, т. е. $0 \le h_i \le n$. Число таких решений уравнения (1) согласно [11] равно числу сочетаний с повторениями из N элементов по n:

$$\binom{N+n-1}{n}.$$

Отметим, что с каждым решением уравнения (1) связано

$$N^n \Big/ \binom{N+n-1}{n}$$

текстов $T_{n,N}(j)$ длины n в алфавите мощностью N.

Максимальные значения параметров текстов (n, N), для которых могут быть рассчитаны точные распределения статистик S_n методом полного перебора и частотным методом, рассчитаны в работах [5, 6] и приведены на рисунке.

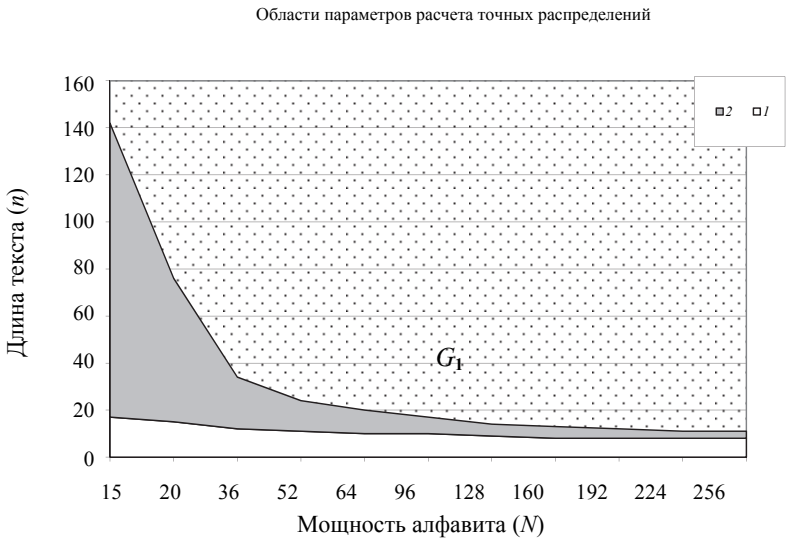

Диаграмма параметров, для которых могут быть рассчитаны точные распределения:

1 – методом полного перебора, *2* – частотным методом

Хотя вычислительная сложность частотного метода расчета точных распределений вероятностей значений статистик намного меньше вычислительной сложности метода полного перебора, но и она не позволяет за приемлемое время провести вычисления точных распределений на современных высокопроизводительных вычислительных системах [11, 12] для практически значимых значений параметров текстов $n > 50$, $N > 64$ даже при их кардинальной модернизации [13].

2. НАПРАВЛЕНИЕ МОДЕРНИЗАЦИИ ЧАСТОТНОГО МЕТОДА РАСЧЕТА ТОЧНЫХ РАСПРЕДЕЛЕНИЙ ЗНАЧЕНИЙ СТАТИСТИК

Частотный метод расчета точных распределений вероятностей значений статистик, имеющий наименьшую из рассматриваемых методов расчета точных распределений вычислительную сложность, не дает возможности провести расчет в интересующих областях изменения параметров (n, N). Это происходит из-за того, что при переборе всех возможных решений уравнения (1) каждое h_i перебирается в следующем диапазоне натуральных (целых положительных) чисел

$$\{h_i \mid i = \overline{1, N},\ h_i \in \mathrm{N},\ 0 \le h_i \le n\}\ , \qquad (2)$$

что определяет область перебора решений.

Интуитивно ясно, что тексты, частотные характеристики которых $(h_1,\ldots,\ h_N)$ мы изучаем, при равновероятном распределении знаков могут состоять только из одного какого-либо знака a_i алфавита $A_N = \{a_1,\ldots,a_N\}$ либо только из двух знаков алфавита A_N и т. д. Данное предположение ин-

терпретируется следующим образом: очень маловероятны решения уравнения (1) следующих видов:

$$\begin{bmatrix} h_1 = n \\ h_2 = 0 \\ \ldots\ldots \\ h_N = 0 \end{bmatrix}, \begin{bmatrix} h_1 = 0 \\ h_2 = n \\ h_3 = 0 \\ h_N = 0 \end{bmatrix}, \ldots., \begin{bmatrix} h_1 = 0 \\ h_2 = 0 \\ \ldots\ldots \\ h_N = n \end{bmatrix}$$

либо

$$\begin{bmatrix} h_1 = n-1 \\ h_2 = 1 \\ h_3 = 0 \\ \ldots\ldots \\ h_N = 0 \end{bmatrix}, \begin{bmatrix} h_1 = n-1 \\ h_2 = 0 \\ h_3 = 1 \\ h_4 = 0 \\ h_N = 0 \end{bmatrix}, \ldots., \begin{bmatrix} h_1 = n-1 \\ h_2 = 0 \\ \ldots\ldots \\ h_{N-1} = 0 \\ h_N = 1 \end{bmatrix}$$

и так далее.

Следовательно, для поиска возможности расчета точных распределений частотным методом [6] необходимо некоторым образом ограничить область перебора решений (2) уравнения (1) некоторым значением m:

$$\{h_i \mid i = \overline{1,N},\ h_i \in N,\ 0 \le h_i \le m,\ m < n\}. \tag{3}$$

Будем учитывать, что ограничение области перебора может привести к потере точности рассчитанного распределения вероятностей значений статистик в том смысле, что оно будет отличаться от точного распределения, которое мы посчитать не имеем возможности.

Рассмотрение вопроса ограничения области перебора (2) приводит нас к рассмотрению поведения статистики максимальной частоты M_n, связанной с параметрами области перебора $(h_1,\ldots,\ h_N)$ следующим образом:

$$M_n = \max_{i=1}^{N} h_i .$$

Опираясь на известное утверждение теории вероятности (в частности, см. [14])

$$P(A) = P(AB) + P(A\overline{B}),$$

выпишем равенство для вероятностей значений статистики с учетом возможных ограничений (3) области перебора

$$P\{S_n \ge c\} = P\{S_n \ge c, M_n > m\} + P\{S_n \ge c, M_n \le m\}. \tag{4}$$

Если удастся подобрать m так, чтобы

$$P\{S_n \ge c, M_n > m\} \le \Delta, \tag{5}$$

то

$$\left|P\{S_n \geq c\} - P\{S_n \geq c, M_n \leq m\}\right| = P\{S_n \geq c, M_n > m\}\,, \tag{6}$$

и по нашему предположению (выражение (5))

$$\left|P\{S_n \geq c\} - P\{S_n \geq c, M_n \leq m\}\right| \leq \Delta\,. \tag{7}$$

Следовательно, вместо вычисления вероятностей значений статистики S_n на всей области значений (2) $P\{S_n \geq c\}$ (точного распределения) можно будет вычислять вероятности значений статистики S_n на ограниченной области (3) $P\{S_n \geq c, M_n \leq m\}$, и вычисленные вероятности не будут отличаться от вероятностей, вычисленных на всей области значений, более чем на величину Δ. В отличие от точного значения вероятности $P\{S_n \geq c\}$ вероятность $P\{S_n \geq c, M_n \leq m\}$ будем называть Δ-точной, а соответствующее ей распределение вероятностей по аналогии с [16] – Δ-точным распределением.

Остается обсудить вопрос о том, как ограничить область перебираемых значений, т. е. как выбрать m так, чтобы $P\{S_n \geq c, M_n > m\} \leq \Delta$.

Вероятность $P\{M_n > m\}$ может быть вычислена с помощью рекуррентной формулы Б.И. Селиванова, предложенной им в 70-х годах XX века и впервые опубликованной в трудах МГУ им. М.В. Ломоносова.

$$P\{M_{n+1} < m\} = \sum_{v=0}^{n} \binom{n}{v} P\{M_{n-v} < m\}\, d_{v+1}^{(m)} \frac{1}{N^v} \tag{8}$$

с начальным условием $P\{M_0 < m\} = 1$, где $\binom{n}{\nu} = \dfrac{n!}{\nu!(n-\nu)!}$ – биномиальный коэффициент, а коэффициенты $d_{\nu+1}^{(m)}$ вычисляются по следующей рекуррентной формуле:

$$d_{v+1}^{(m)} = -\sum_{v=1}^{m-1} \binom{n}{v}\, d_{n+1-v}^{(m)} \tag{9}$$

с начальным условием

$$\begin{gathered} d_{1}^{(m)} = 1\,, \\ d_{2}^{(m)} = d_{3}^{(m)} = \ldots = d_{m-1}^{(m)} = 0\,, \\ d_{m}^{(m)} = -1\,, \\ d_{m+1}^{(m)} = m\,. \end{gathered} \tag{10}$$

Таким образом, заранее задав точность Δ, например 10^{-5}, и используя соотношения (8), (9) и (10), можем подобрать значение m, для которого $P\{M_n < m\} < 1 - \Delta$. Например, для $n = 50$, $N = 26$ и $\Delta = 10^{-5}$ $P\{M_{50} < 12\} = 0,9999992$ и соответственно $m = 12$. Следовательно, область перебора решений уравнения (2) для расчета вероятностей уменьшена с

$$\{h_i \mid i = \overline{1,26},\ h_i \in N,\ 0 \le h_i \le 50\}$$

до

$$\{h_i \mid i = \overline{1,26},\ h_i \in N,\ 0 \le h_i \le 12\}.$$

3. МЕТОДИКА РАСЧЕТА Δ-ТОЧНЫХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ ЗНАЧЕНИЙ СТАТИСТИК

Для применения модернизированного частотного метода определим μ_ν как число значений h_i из уравнения (1), равных ν. Тогда для расчета вероятностей от перечисления решений уравнения (1)

$$h_1 + \ldots + h_N = n$$

можно перейти к перебору решений системы уравнений

$$\begin{cases} \mu_0 + \mu_1 + \ldots + \mu_n = N, \\ 1\mu_1 + 2\mu_2 + \ldots + n\mu_n = n. \end{cases} \tag{11}$$

Учитывая ограничения (3) на область перебора решений уравнения (1) получаем, что при принятых ограничениях

$$\mu_{m+1} = \mu_{m+2} = \ldots = \mu_n = 0 \tag{12}$$

можно от перебора решений системы уравнений (11) перейти к перебору усеченной системы уравнений ($m < n$):

$$\begin{cases} \mu_0 + \mu_1 + \ldots + \mu_m = N, \\ 1\mu_1 + 2\mu_2 + \ldots + m\mu_m = n. \end{cases} \tag{13}$$

Выделяя независимые переменные и применяя метод их последовательного задания с определением зависимых переменных, получаем все решения системы (13):

$$\left\{\mu^{(i)} \mid \left(\mu_0^{(i)},\ \mu_1^{(i)}, \ldots,\ \mu_m^{(i)}\right),\ i = \overline{1,Z}\right\}. \tag{14}$$

Для рассматриваемого примера при $n = 50$, $N = 26$ и $\Delta = 10^{-5}$ $P\{M_{50} < 12\} = 0{,}9999992$ и соответственно $m = 12$, а вычисленное $Z = 92\,154$, что намного меньше, чем сложность просто частотного метода, равная числу сочетаний с повторениями из N элементов по n и оцениваемая как $5 \cdot 10^{19}$. Расчеты проводились с помощью программ, составленных на высокоуровневом языке Python [17] 64-битной версии 3.5.1 с использованием модуля decimal для работы с числами большой разрядности.

Заметим, что в соответствии с [18] с каждым решением системы (13) связано

$$K^{(i)} = \frac{N!}{\mu_0^{(i)}!\mu_1^{(i)}!\ldots.\mu_m^{(i)}!} \tag{15}$$

решений уравнения (1). Тогда $P^{(i)}$ – вероятность того, что решение уравнения (11)

$$\left(\mu_0{}^{(i)}, \mu_1{}^{(i)}, \ldots, \mu_n{}^{(i)}\right)$$

примет значение $\mu^{(i)}$ из (14), равна

$$P^{(i)} = K^{(i)} \frac{n!}{(2!)^{\mu_2^{(i)}} \cdot (3!)^{\mu_3^{(i)}} \cdot \ldots \cdot (m!)^{\mu_m^{(i)}} \cdot N^n} \tag{16}$$

Теперь для каждого $\left\{\mu^{(i)} \mid i = \overline{1, Z}\right\}$ мы можем вычислить $P^{(i)}$ и значение статистики $S_n^{(i)}$, например $\chi_n^{(i)}$, где

$$\chi_n^{(i)} = \chi_n\left(\mu^{(i)}\right) = \frac{N}{n} \sum_{\nu=0}^{m} \mu_\nu^{(i)} \left(\nu - \frac{n}{N}\right)^2. \tag{17}$$

Имея вычисленные $\left\{\mu^{(i)}, P^{(i)}, S_n^{(i)} \mid i = \overline{1, Z}\right\}$, можно перейти непосредственно к вычислению вероятностей $P\{S_n \geq c\}$. Для этого для всех

$$\left\{c_j \mid j = 1, 2, \ldots., \max_{i=1}^{Z} S_n^{(i)}\right\}$$

вычисляем

$$P\{S_n \geq c_j\} = \sum_{i=1}^{Z} P^{(i)} \cdot I\left(S_n^{(i)}, c_j\right),$$

где

$$I\left(S_n^{(i)}, c_j\right) = \begin{cases} 1 \text{ при } S_n^{(i)} \geq c_j, \\ 0 \text{ при } S_n^{(i)} < c_j. \end{cases}$$

Таким образом, вычисленная последовательность

$$\left\{ P\{S_n \geq c_j\} \mid, j = 1, 2, \ldots, \max_{i=1}^{Z} S_n^{(i)} \right\}$$

является дискретным распределением статистики S_n, отличающимся от точного распределения не более чем на заданную величину Δ.

После проведенных рассуждений можно перейти к формализации методики расчета Δ-точных распределений вероятностей значений симметричных аддитивно разделимых статистик.

Пусть n – длина выборки (текста) и N – мощность алфавита текста. Точное распределение вероятностей значений статистики $S_n - P_T\{S_n \geq c\}$ мы рассчитать не можем, поэтому нам надо рассчитать Δ-точное распределение $P_\Delta\{S_n \geq c\}$, отличающееся от точного не более чем на заданную величину Δ:

$$\left| P_T\{S_n \geq c\} - P_\Delta\{S_n \geq c\} \right| \leq \Delta.$$

Для этого предпринимаем следующие шаги, которые и составляют суть методики.

Шаг 1. По заданным (n, N) и выбранной точности Δ (например, 10^{-5}) для ограничения области перебора решений уравнения и нахождения m по формулам (8), (9) и (10) последовательно вычисляем

$$P\{M_1 < 2\},\ P\{M_2 < 2\},\ldots,\ P\{M_n < 2\},$$

$$P\{M_1 < 3\},\ P\{M_2 < 3\},\ldots,\ P\{M_n < 3\},$$

$$P\{M_1 < m\},\ P\{M_2 < m\},\ldots,\ P\{M_n < m\},$$

пока не выполнится условие $P\{M_n < m\} > 1 - \Delta$. Таким образом определяем m.

Шаг 2. Для перечисления всех решений системы уравнений

$$\begin{cases} \mu_0 + \mu_1 + \ldots + \mu_m = N, \\ 1\mu_1 + 2\mu_2 + \ldots + m\mu_m = n \end{cases}$$

выделяем независимые переменные и применяем метод их последовательного задания с определением зависимых переменных. Получаем все Z решений

$$\left\{\mu^{(i)} \mid \left(\mu_0^{(i)}, \mu_1^{(i)}, \ \ldots \mu_m^{(i)}\right), \ i = \overline{1, Z}\right\}.$$

Шаг 3. Для каждого решения

$$\left\{\mu^{(i)} \mid \left(\mu_0^{(i)}, \mu_1^{(i)}, \ \ldots \mu_m^{(i)}\right), \ i = \overline{1, Z}\right\}$$

вычисляем значение статистики $S_n^{(i)}$ (например, по формуле (17)), а по формулам (15) и (16) – вероятность его появления $P^{(i)}$. Таким образом, имеем вычисленные решения уравнения $\mu^{(i)}$, значения статистики от этих решений $S_n^{(i)}$ и их вероятности $P^{(i)}$ – $\left\{\mu^{(i)}, \ S_n^{(i)}, P^{(i)} \mid i = \overline{1, Z}\right\}$. Необходимо отметить, что формулы расчета значений одной и той же статистики от значений частот встречаемости знаков h_i и от значений так называемых «вторых» маркировок μ_ν отличаются друг от друга. Этот факт необходимо учитывать при проведении расчетов.

Шаг 4. Для получения распределения вероятностей значений $P\{S_n \geq c_j\}$ маркируем значения статистики $\left\{S_n^{(i)} \mid i = \overline{1, Z}\right\}$ в интервалах $\left\{c_j \mid j = 1, \ 2, \ldots, \max_{i=1}^{Z} S_n^{(i)}\right\}$ с одновременным суммированием соответствующих $P^{(i)}$.

Описание методики окончено.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

В работе рассмотрены направление и возможность модернизации частотного метода расчета точных распределений значений симметричных аддитивно разделимых статистик для его применения в области значений параметров, для которых частотный метод не может быть применим из-за его большой вычислительной сложности.

Показана возможность применения модернизированного частотного метода для расчета распределений значений симметричных аддитивно разделимых статистик, отличающихся от точных распределений не более чем на заранее заданную величину Δ.

На основе модернизации частотного метода разработана методика расчета Δ-точных распределений, отличающихся от точных распределений не более чем на заранее заданную величину.

Показана возможность применения методики расчета Δ-точных распределений для областей значений параметров статистической выборки (тек-

стов), для которых точные распределения не могут быть рассчитаны из-за своей большой вычислительной и временной сложности.

Приводятся результаты расчета Δ-точных распределений для конкретных значений параметров выборки.

Необходимость проведения статистического анализа текстов на всем практическом спектре их параметров требует исследования сложности методики расчета Δ-точных распределений для областей значений параметров текстов, в которых не могут быть применены предельные распределения, что является предметом дальнейших исследований автора.

Благодарность

Автор выражает глубокую благодарность доктору физико-математических наук, профессору А.Ф. Ронжину за постоянное внимание к работе и ее обсуждение.

СПИСОК ЛИТЕРАТУРЫ

1. *Чеповский А.М.* Информационные модели в задачах обработки текстов на естественных языках. – М.: ИНТУИТ, 2015. – 228 с. – ISBN 978-5-9556-0176-2.

2. *Крамер Г.* Математические методы статистики. – М.: Мир, 1975. – 648 с.

3. *Ивченко Г.И., Медведев Ю.И.* Введение в математическую статистику. – М.: Ленард, 2017. – 608 с. – ISBN 978-5-9710-4535-9.

4. *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. – М.: Либроком, 2014. – 352 с. – ISBN 978-5-397-04141-6.

5. *Зелюкин Н.Б., Мельников А.К.* Сложность расчета точных распределений вероятности значений статистик и область применения предельных распределений // Электронные средства и системы управления: материалы докладов XIII Международной научно-практической конференции (29 ноября – 1 декабря 2017 г.): в 2 ч. – Томск: В-Спектр, 2017. – Ч. 2. – С. 84–90.

6. *Мельников А.К.* Сложность расчета точных распределений вероятности симметричных аддитивно разделяемых статистик и область применения предельных распределений // Доклады ТУСУР. – 2017. – Т. 20, № 4. – С. 126–130.

7. *Helmert P.R.* Uber die Wahrscheinlichkeit von Potenzsummen der Beobachtungsfehler und iiber einige damit im Zusammenhange stehende Fragen // Zeitschrift für Mathematik und Physik. – 1876. – B. 21. – S. 102–219.

8. *Neyman F., Pearson E.S.* On the use and interpretation of certain test criteria for purposes of statistical inference // Biometrika. – 1928. – Vol. 20-A. – P. 175–240; 264–299.

9. *Pearson K.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling // The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. Series 5. – 1900. – Vol. 50, N 302. – P. 157–170.

10. Exact and approximate distributions of the chi-squared statistic for equiprobability / P.F. Smith, D.S. Rae, R.W. Manderscheid, S. Silbergeld // Communications in Statistics – Simulation and Computation. – 1979. – Vol. 8 (2). – P. 131–149.

11. *Корнеев В.В.* Вычислительные системы. – М.: Гелиос АРВ, 2004. – 512 с. – ISBN 5-85438-117-6.

12. *Каляев И.А., Левин И.И., Семерников Е.А.* Реконфигурируемые вычислительные системы на основе ПЛИС // Интеллект & Технологии. – 2014. – № 1 (7). – С. 40–47.

13. *Мельников А.К.* Исследование путей модернизации реконфигурируемых вычислительных систем // Известия ЮФУ. Технические науки. – 2014. – № 12 (161). – С. 83–89.

14. *Холл М.* Комбинаторика. – М.: Мир, 1970. – 424 с.

15. *Феллер В.* Введение в теорию вероятностей и ее приложения. В 2 т. Т. 1. – М.: Мир, 1984. – 528 с.

16. *Мельников А.К., Ронжин А.Ф.* Обобщенный статистический метод анализа текстов, основанный на расчете распределений вероятности значений статистик // Информатика и ее применения. – 2016. – Т. 10, вып. 4. – С. 89–95.

17. Описание языка программирования Python [Электронный ресурс]. – URL: https://www.python.org/doc/ (дата обращения: 22.03.2018).

18. *Сачков В.Н.* Комбинаторные методы дискретной математики. – М.: Наука, 1977. – 320 с.

Мельников Андрей Кимович, кандидат технических наук, главный научный сотрудник Научно-технического центра ЗАО «ИнформИнвестГрупп». Основные направления научных исследований: применение точных и предельных распределений при построении статистических критериев, организация обработки сообщений на многопроцессорных вычислительных системах, разработка методов и языков программирования. Имеет более 60 научных публикаций. E-mail: ak@iigroup.ru

DOI: 10.17212/1814-1196-2018-1-153-166

Processing complexity of exact probability distributions of symmetrical additively partitioned statistics and the application area of limit distributions[*]

A.K. MELNIKOV

STC CLSC "InformInvestGroup", 125, Varshavskoye Road, building 17, Moscow, 117587, Russian Federation, PhD (Eng.), SAC associate professor; chief research officer. E-mail: ak@iigroup.ru

In the paper, in order to develop some text processing procedures we consider a possibility of using exact distributions of statistical probabilities for creating statistical goodness-of-fit tests with an equiprobable distribution. We compare the computational complexity of exact distributions using the trivial method of full enumeration and the frequency method. It is proved that the computational complexity of the frequency method of exact distribution calculation is considerably lower than the computational complexity of the method of full enumeration. However, it also does not allow calculation of exact distributions for practically meaningful values of text parameters during an appropriate time on modern high-performance computer systems, even in case of cardinal upgrade of computer systems by means of advanced computing elements. Owing to narrowing of the class of the used statistics to the class of symmetrical additively-partitioned statistics, we have chosen the direction of improvement of the frequency method of exact distribution calculation. The aim is to restrain the enumerated sample space. Besides, we prove a principle possibility of application of the improved method in the areas of text parameters, where high computational complexity of the frequency method does not allow calculation of exact distributions. Based on the results of the estimation of maximum frequency statistic probability, we have improved the frequency method of exact distribution calculations. As a result, we have created a methodology of calculation of Δ–exact distributions, which differ from exact distributions no more than by the predetermined value Δ. A step-by-step description of the calculation methodology of Δ–exact distributions, which allows its practical application, is given. Besides, the results of the application of the methodology of the Δ–exact distribution calculation of text parameters are given. At present the calculation of exact distributions of these text parameters is impossible due to its high computational complexity.

Keywords: probability, statistics, criterion/test, exact distribution, limit distribution, computational complexity of method, performance of multiprocessor computer system

[*] *Received 02 October 2017.*

REFERENCES

1. Chepovskii A.M. *Informatsionnye modeli v zadachakh obrabotki tekstov na estestvennykh yazykakh* [Information models in tasks of processing of natural language texts]. Moscow, INTUIT Publ., 2015. 228 p.

2. Cramer H. *Mathematical methods of statistics*. Princeton, Princeton University Press, 1946 (Russ. ed.: Kramer G. *Matematicheskie metody statistiki*. Moscow, Mir Publ., 1975. 648 p.).

3. Ivchenko G.I., Medvedev Yu.I. *Vvedenie v matematicheskuyu statistiku* [Introduction to mathematical statistics]. Moscow, Lenard Publ., 2017. 608 p. ISBN 978-5-9710-4535-9.

4. Ivchenko G.I., Medvedev Yu.I. *Matematicheskaya statistika* [Mathematical statistics]. Moscow, Librokom Publ., 2014. 352 p. ISBN 978-5-397-04141-6.

5. Zelyukin N.B., Mel'nikov A.K. [Slozhnost' rascheta tochnykh raspredeleniy veroyatnosti znacheniy statistik i oblast' primeneniya predelnykh raspredeleniy]. *Elektronnye sredstva i sistemy upravleniya: materialy dokladov XIII Mezhdunarodnoi nauchno-prakticheskoi konferentsii* [Electronic facilitiesand control systems: reports of the XIIIth International scientific and practical], 29th November – 1st December, 2017. Tomsk, 2017, pt. 2, pp. 84–90. (In Russian).

6. Mel'nikov A.K. Slozhnost' rascheta tochnykh raspredelenii veroyatnosti simmetrichnykh additivno razdelyaemykh statistik i oblast' primeneniya predel'nykh raspredelenii [Processing complexity in exacting probability distributions of symmetrical additively partitioned statistics and application area of limit distributions]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki – Proceedings of Tomsk State University of Control Systems and Radioelectronics*, 2017, vol. 20, no. 4, pp. 126–130.

7. Helmert P.R. Uber die Wahrscheinlichkeit von Potenzsummen der Beobachtungsfehler und iiber einige damit im Zusammenhange stehende Fragen. *Zeitschrift für Mathematik und Physik*, 1876, B. 21, pp. 102–219.

8. Neyman F., Pearson E.S. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 1928, vol. 20-A, pp. 175–240, 264–299.

9. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. Series 5*, 1900, vol. 50, no. 302, pp. 157–170.

10. Smith P.F., Rae D.S., Manderscheid R.W., Silbergeld S. Exact and approximate distributions of the chi-squared statistic for equiprobability. *Communications in Statistics - Simulation and Computation*, 1979, vol. 8 (2), pp. 131–149.

11. Korneev V.V. *Vychislitel'nye sistemy* [The computing system]. Moscow, Gelios ARV Publ., 2004. 512 p. ISBN 5-85438-117-6.

12. Kalyaev I.A., Levin I.I., Semernikov E.A. Rekonfiguriruemye vychislitel'nye sistemy na osnove PLIS [Reconfigurable computing system on FPGA]. *Intellekt & Tekhnologii – [Intelligence & Technology*, 2014, no 1 (7), pp. 40–47.

13. Mel'nikov A.K. Issledovanie putei modernizatsii rekonfiguriruemykh vychislitel'nykh sistem [Research of possible modifications of reconfigurable computer systems]. *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki – Izvestiya Southem Federal University. Engineering sciences*, 2014, no. 12 (161), pp. 83–89.

14. Hall M. *Combinatorial theory*. Waltham, MA, Blaisdell Publ. Co., 1967 (Russ. ed.: Kholl M. *Kombinatorika*. Moscow, Mir Publ., 1970. 424 p.).

15. Feller W. *An intoroduction to probability theory and its applications*. Vol. 1. New York, John Wiley & Sons, 1970 (Russ. ed.: Feller V. *Vvedenie v teoriyu veroyatnostei i ee prilozheniya*. V 2 t. T. 1. Moscow, Mir Publ., 1984. 528 p.).

16. Melnikov A.K., Ronzhin A.F. Obobshchennyi statisticheskii metod analiza tekstov, osnovannyi na raschete raspredelenii veroyatnosti znachenii statistik [Generalized statistical method of text analysis based on calculation of probability distributions of statistical values]. *Informatika i ee primeneniya – Informatics and Applications*, 2016, vol. 10, iss. 4, pp. 89–95.

17. *Python programming language description*. Available at: https://www.python.org/doc/ (accessed 22.03.2018).

18. Sachkov V.N. *Kombinatornye metody diskretnoi matematiki* [Combinatorial methods of discrete mathematics]. Moscow, Nauka Publ., 1977. 320 p.

Для цитирования:

Мельников А.К. Методика расчета распределения вероятностей значений симметричных аддитивно разделяемых статистик, приближенных к их точному распределению // Научный вестник НГТУ. – 2018. – № 1 (70). – С. 153–166. – doi: 10.17212/1814-1196-2018-1-153-166.

For citation:

Melnikov A.K. Metodika rascheta raspredeleniya veroyatnostei znachenii simmetrichnykh additivno razdelyaemykh statistik priblizhennogo k ikh tochnomu raspredeleniyu [Processing complexity for exact probability distributions of symmetrical additively partitioned statistics and the application area of limit distributions]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Science bulletin of the Novosibirsk state technical university*, 2018, no. 1 (70), pp. 153–166. doi: 10.17212/1814-1196-2018-1-153-166.