УДК 004.056:519.1

**А.К. Мельников**

# Сложность расчета точных распределений вероятности симметричных аддитивно разделяемых статистик и область применения предельных распределений

Рассматривается применение частотного метода для расчета точных распределений вероятности симметричных аддитивно разделяемых статистик, анализируется его вычислительная и временная сложность. Исследуются значения граничных параметров, для которых на современном этапе частотным методом возможен расчет точных распределений. Сравниваются возможности по расчету точных распределений с использованием частотного метода и тривиального метода полного перебора. Исследуются значения граничных параметров текстов, для которых возможен расчет точных распределений на современном этапе. Рассматриваются возможности применения точных и предельных распределений вероятностей значений статистик для построения критериев согласия в рамках анализа текстовой информации.

**Ключевые слова:** вероятность, статистика, критерий, точное распределение, предельное распределение, вычислительная сложность метода, производительность многопроцессорной вычислительной системы.
**doi:** 10.21293/1818-0442-2017-20-4-126-130

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

Пусть из некоторого массива текстов длины $n$, состоящих из знаков алфавита $A_N=\{a_1, ..., a_N\}$ мощности $N$,

$$T_{n,N}(v) = \{t_1(v), \dots, t_n(v)\},\ v = \overline{1,M}$$

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

Выбор текстов с равновероятным распределением знаков производится с помощью применения критерия согласия с равновероятным распределением, использующим некоторую статистику $S_n$ текста длины $n$, являющейся 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 $\alpha$.

Для определения размера критерия согласия $\alpha$ необходимо знать вероятность распределения значений, применяемых в критерии статистик $S_n$

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

связанных с размером критерия $\alpha$ соотношением

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

В критерии согласия могут использоваться как точные значения вероятности распределения значений статистики $S_n$, расчету которых без ограничения на вид функции $f$ посвящена работа автора [2], так и предельные значения вероятности, определяемые свойствами самой функции $f$, например, как это показано Хельмертом и Пирсоном [3].

В данной работе не исследуется зависимость размера применяемого критерия от вида функции $f$ статистки $S_n$, как это делалось в [5], а рассматривается эта зависимость от вида применяемого распределения. Вводится лишь одно ограничение на вид статистики критерия согласия, требуется её симметричность относительно используемых в ней частот встречаемости знаков (исходов) алфавита $a_i$ – $h_i$, рассматривается класс симметричных аддитивно разделимых статистик [6].

Целью данной работы является описание границ областей возможности вычисления точных распределений вероятностей значений симметричных аддитивно разделимых статистик на современном этапе развития вычислительной техники и её сравнение с границей области применения предельных распределений.

**Расчет точных распределений значений симметричных аддитивно разделимых статистик**

В работе [2] автором уже исследовался вопрос расчета точных распределений статистики $S_n$ хи-квадрат – $\chi_n$, предложенной Карлом Пирсоном в [6, 7]:

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

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. Но расчеты точных распределений статистики $\chi_n$ исследовались в общем случае, без учета свойств класса статистик, к которым она принадлежит.

Одним из замечательных свойств класса симметричных аддитивно разделимых статистик, к которым принадлежит статистика $\chi_n$, является то, что при равновероятном распределении знаков текста

$$\{P(t_i(v) = a_j) = 1/N \mid i = 1, ..., n;\ \ j = 1, ..., N\}$$

они симметричны относительно входящих в них частот встречаемости знаков текста – $h_i$.

Тогда для расчета распределения вероятности значений статистики (распределение вероятности) $S_n$

$$P\{S_n \ge c\}$$

для равновероятной полиномиальной схемы, когда

$$\{p_i = 1/N \mid i = 1, ..., N\},$$

можно перейти от перечисления всех текстов длины $n$ в алфавите $A_N = \{a_1, ... , a_N\}$ мощности $N$

$$T_{n,N}(v) = \{t_1(v), \dots , t_n(v)\}, \ v = 1, ..., N^n$$

к перечислению всех решений уравнения

$$h_1 + \dots + h_N = n \qquad (1)$$

в неотрицательных целых числах, т.е. $0 \le h_i \le n$. Число таких решений (1) равно [8] числу сочетаний с повторениями из $N$ элементов по $n$

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

С каждым решением уравнения (1) связано

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

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

Тогда вычислительная сложность (число машинных операций) расчета точных распределений значений статистик $P_T\{S_n \ge c\}$ методом, использующим частоты входящих в тексты знаков алфавита, назовем его частотным методом – $C_{\text{чм}}(P_T\{S_n \ge c\})$, определяется количеством решений уравнения (1) в неотрицательных целых числах и имеет вид

$$C_{\text{чм}}(P_T\{S_n \ge c\}) = C(S_n(h_1^{(i)},...,h_N^{(i)})) \times \binom{N+n-1}{n}, \quad (2)$$

где $C(S_n(h_1^{(i)},...,h_N^{(i)}))$ есть вычислительная сложность расчета значений статистики $S_n$ для одного решения $(h_1^{(i)},...,h_N^{(i)})$ уравнения (1) с учетом вклада этого решения в $P_T\{S_n \ge c\}$.

По аналогии с [9] определим вычислительную сложность расчета одного значения статистики $S_n$ от $(h_1^{(i)},...,h_N^{(i)})$ в $(3N + 100)$ машинных операций:

$$C(S_n(h_1^{(i)},...,h_N^{(i)})) = 3N + 100, \qquad (3)$$

где $3N$ – число операций для вычисления статистики $S_n$ на равновероятном полиномиальном распределении с $N$ исходами, 100 – число вспомогательных операций.

Аналогично [2] для проведения расчета точных распределений их вычислительная сложность $C_{\text{чм}}(P_T\{S_n \ge c\})$ должна обеспечиваться производительностью используемых вычислительных средств $\Pi_{\text{вс}}$ и временем $T$ проведения расчета.

$$C_{\text{чм}}(P_T\{S_n \ge c\}) \le \Pi_{\text{вс}} \times T. \qquad (4)$$

Следовательно, исходя из (2)–(4) основным соотношением, связывающим значения параметров текста ($n$, $N$), для которых рассчитывается точное распределение, производительность используемых для расчета вычислительных средств $\Pi_{\text{вс}}$ и время проведения расчета $T$ является

$$C(S_n(h_1^{(i)},...,h_N^{(i)})) \times \binom{N+n-1}{n} \le \Pi_{\text{вс}} \times T. \qquad (5)$$

Для условий, принятых в [2] к производительности вычислительных средств: $\Pi_{\text{вс}} = 10^{16}$ операций в секунду, и времени расчета: $T$ = 30 дней или 2 592 000 с, поведем оценку параметров ($n$, $N$), для которых на современном этапе точные распределения могут быть рассчитаны частотным методом.

Принимая во внимание ограничения к $\Pi_{\text{вс}}$ и $T$ и (5), параметры $n$ и $N$ должны удовлетворять следующему соотношению:

$$\binom{N+n-1}{n} \le \frac{2{,}59 \times 10^{22}}{3N + 100}. \qquad (6)$$

Используя принятые выше предположения, для всех натуральных целых $N$ от 2 до 256 были рассчитаны максимальные значения параметра $n$, для которых на вычислительном ресурсе производительностью $10^{16}$ операций в секунду (оп./с) за «приемлемое» время 1 месяц могут быть рассчитаны точные распределения статистик $S_n$ рассматриваемым частотным методом. Значения параметров приведены в табл. 1 вместе со значениями, полученными при тех же условиях расчета методом полного перебора в [2].

Таблица 1

**Максимальные значения параметров текстов, для которых могут быть рассчитаны точные распределения статистик $S_n$ методом полного перебора и частотным методом**

| Число исходов полиномиальной схемы (мощность алфавита) $N$ | Объем выборки (длина текста) $n$ | |
|---|---|---|
| | Метод полного перебора | Частотный метод |
| 5 | 28 | 41584 |
| 7 | 23 | 3233 |
| 8 | 22 | 1498 |
| 9 | 21 | 834 |
| 10 | 20 | 527 |
| 15 | 17 | 142 |
| 20 | 15 | 76 |
| 26 | 14 | 50 |
| 36 | 12 | 34 |
| 52 | 11 | 24 |
| 64 | 11 | 21 |
| 96 | 10 | 17 |
| 128 | 9 | 14 |
| 192 | 8 | 12 |
| 256 | 8 | 11 |

Максимальные значения параметров текстов ($n$, $N$), соответствующие табл. 1, приведены на рис. 1.

Анализ значений табл. 1 показывает, что применение частотного метода расчета точных распределений для алфавитов мощности $N$ от 2 до 64 дает возможность рассчитывать точные распределения симметричных аддитивно разделимых статистик $S_n$ для длин сообщений $n$, в 2 раза и более превосходящих длины сообщений, для которых возможен расчет точных распределений методом полного перебора. Для алфавитов мощности $N$ более 64 разницу в длинах сообщений $n$, для которых могут быть рассчитаны точные распределения сравниваемыми методами, можно считать незначительной.

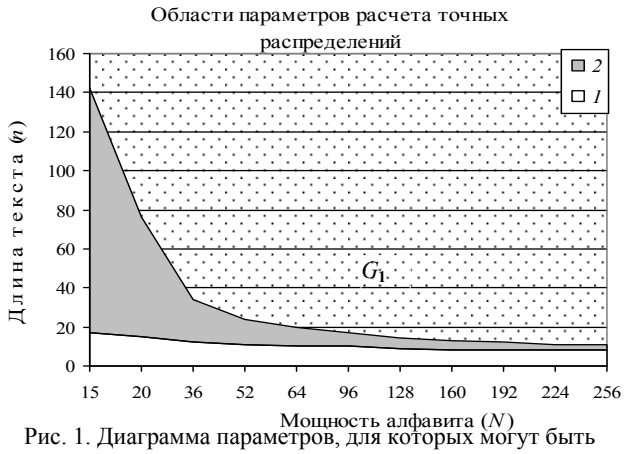


Рис. 1. Диаграмма параметров, для которых могут быть рассчитаны точные распределения: *1* – методом полного перебора, *2* – частотным методом

В [2] автором было показано, что ни разумное увеличение время расчета точных распределений до одного года, ни гипотетическая возможность использования вычислительных средств на два порядка более производительных до $10^{18}$ оп./с, не дадут сколько-нибудь практически значимого увеличения длин сообщений, для которых можно рассчитать точные распределения статистик.

Для полноты исследования возможностей частотного метода расчета точных распределений и сравнения его с методом полного перебора приведем значения вычислительной сложности расчета точных распределений для различных длин текстов $n$ и мощностей алфавита $N$. Результаты расчетов приведены в табл. 2.

Таблица 2

**Вычислительная сложность расчета точного распределения симметричных аддитивно разделимых статистик $S_n$ методом полного перебора и частотным методом**

| Параметры полиномиальной схемы | | Вычислительная сложность (оп./с) | |
|---|---|---|---|
| Мощность алфавита $N$ | Длина текста $n$ | Метод полного перебора | Частотный метод |
| 26 | 50 | $2{,}12\times10^{73}$ | $9{,}36\times10^{21}$ |
| 26 | 100 | $1{,}97\times10^{144}$ | $2{,}31\times10^{28}$ |
| 26 | 150 | $1{,}55\times10^{215}$ | $2{,}26\times10^{32}$ |
| 64 | 30 | $6{,}01\times10^{56}$ | $6{,}42\times10^{26}$ |
| 64 | 50 | $1{,}00\times10^{93}$ | $1{,}08\times10^{35}$ |
| 128 | 30 | $9{,}61\times10^{65}$ | $7{,}10\times10^{34}$ |
| 256 | 30 | $1{,}71\times10^{75}$ | $2{,}97\times10^{43}$ |
| 256 | 50 | $2{,}76\times10^{123}$ | $6{,}66\times10^{60}$ |

Расчеты значений табл. 1 и 2 проводились с помощью программ, составленных на высокоуровневом языке Python [10] 64-битной версии 3.5.1 с использованием модуля decimal для работы с числами большой разрядности.

Сравнение вычислительной сложности расчета точных распределений статистик показывает, что частотный метод имеет гораздо меньшую вычислительную сложность, отличающуюся от вычислительной сложности метода полного перебора на много порядков, но и его сложность не позволяет этим методом проводить расчеты для практически значимых значений параметров на современном этапе развития вычислительной техники [11].

По аналогии с [2] область параметров, для которых частотным методом могут быть рассчитаны точные распределения, назовем областью $O_1$. Верхнюю границу области $O_1$ обозначим через $G_1$. Напомним, что по построению $O_1$ аналогично [2] содержит только пары натуральных целых чисел $(n, N)$.

**Области применения предельных и точных распределения значений симметричных аддитивно разделимых статистик**

Автором в работах [2, 9] уже отмечалось важное свойство статистики $\chi_n$, из которого следует, что при

$$m = \min_{i=1}^{N} np_i \to \infty$$

её предельное распределение не зависит от вероятностей исходов полиномиальной схемы $p_i$, ..., $p_N$ и совпадает с $\chi^2$-распределением с $(N - 1)$ степенью свободы [3, 7].

В рассматриваемом в статье случае при равновероятном полиномиальном распределении

$$\{p_i = 1/N \mid i = 1, ..., N\}$$

должно выполняться условие

$$m = n/N \to \infty.$$

Результаты исследования вопроса, начиная с какого $m$ можно пользоваться предельным распределением, рассматривались автором в [10]. В [12] предлагается ограничение $m \geq 5$, в [13] – $m \geq 20$, в [14] – $m \geq 30$.

В [2] для $m = 5$ автором было проведено построение области предельных распределений $O_3$, области значений параметров $(n, N)$ для которых согласно предположению Фишера [13] могут применяться предельные распределения – табл. 2 в [2]. Обозначим аналогично [2] нижнюю границу области $O_3$ через $G_2$. Значения параметров области предельных распределений показаны на рис. 2.

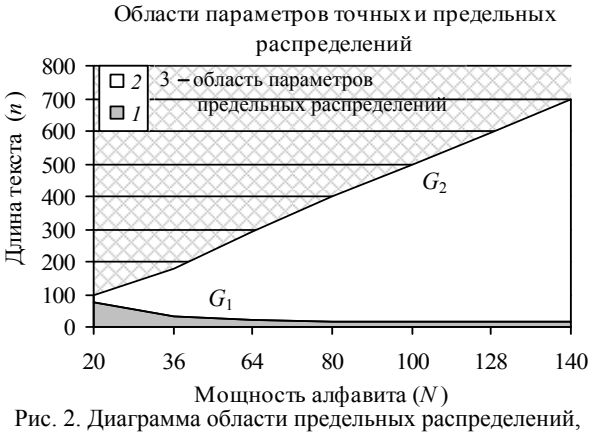


Рис. 2. Диаграмма области предельных распределений, ограниченная снизу границей $G_2$: *1* – область точных распределений, *2* – область неопределенности, *3* – область предельных распределений

**Ограничения на применение точных и предельных распределений симметричных аддитивно разделимых статистик**

Анализ диаграммы областей точных и предельных распределений симметричных аддитивно разделимых статистик на рис. 2 показывает, что между областью точных распределений $O_1$ и областью пре-

дельных распределений $O_3$ продолжает располагаться, как и в случае общего вида статистик [2], область неопределенности, для которой на современном этапе не могут быть рассчитаны точные распределения, а предельные распределения не могут быть использованы. Аналогично [2] обозначаем эту область через $O_2$. Напомним, что по построению нижней границей области $O_2$ является граница $G_1$, определяющая верхнюю границу области возможного расчета точных распределений $O_1$ частотным методом. Одновременно верхней границей области $O_2$ является граница $G_2$, определяющая нижнюю границу области применения предельных распределений.

Расчеты значений минимального уровня производительности вычислительных средств, требуемого для расчета частотным методом точных распределений для всех параметров из области неопределенности, могут быть получены из условия (4), принимающего при предположении о времени расчета $T$, равного 1 месяцу, и стремлении $m$ к бесконечности

$$n = m\,N$$

следующий вид:

$$\Pi_{\text{вс}} = \binom{N+n-1}{n} \times (3N+100)/2592000\,. \qquad (7)$$

С помощью выражения (7) проводим расчеты значений минимальной производительности вычислительных средств $\Pi_{\text{вс}}$, которые могут позволить для разных алфавитов частотным методом рассчитать за один месяц точные распределения для области неопределенности при $m$, равном 5, 20 и 30 соответственно. Результаты расчетов частично приведены в табл. 3.

Таблица 3

**Минимальные значения производительности вычислительных средств, необходимые для расчета точных распределений симметричных аддитивно разделимых статистик $S_n$ во всей области неопределенности частотным методом**

| Число исходов полиномиальной схемы (мощность алфавита) $N$ | Значение производительности $\Pi_{\text{вс}}$ (оп./с) | | |
|---|---|---|---|
| | $m = 5$ | $m = 20$ | $m = 30$ |
| 17 | $9{,}1\ 10^{13}$ | $1{,}3\ 10^{23}$ | $7{,}6\ 10^{25}$ |
| 18 | $1{,}4\ 10^{15}$ | $7{,}3\ 10^{24}$ | $6{,}2\ 10^{27}$ |
| 19 | $2{,}0\ 10^{16}$ | $4{,}0\ 10^{26}$ | $5{,}1\ 10^{29}$ |
| 20 | $3{,}0\ 10^{17}$ | $2{,}2\ 10^{28}$ | $4{,}2\ 10^{31}$ |
| 21 | $4{,}5\ 10^{18}$ | $1{,}2\ 10^{30}$ | $3{,}5\ 10^{33}$ |

Анализ результатов расчета значений производительности вычислительных средств, необходимых для получения частотным методом точных распределений симметричных аддитивно разделимых статистик $S_n$ для конкретных значений мощности алфавита $N$ во всей области неопределенности $O_2$, при условии доступности вычислительного ресурса производительностью не более $10^{16}$ оп./с, показывает:

– точные распределения симметричных аддитивно разделимых статистик $S_n$ могут быть рассчитаны для мощности алфавита $N$, меньшей 19;

– для мощностей алфавита $N$ больше 20 точные распределения рассматриваемых статистик $S_n$ не могут быть посчитаны за «приемлемое» время расчета в один месяц;

– для мощностей алфавита $N$ больше 20 точные распределения рассматриваемых статистик $S_n$ не могут быть посчитаны даже при увеличении «приемлемого» времени расчета с одного месяца до одного года;

– опираясь на оценку показателей роста производительности современных и перспективных вычислительных средств [15], можно утверждать, что производительность, требуемая для расчета точных распределений рассматриваемых статистик $S_n$ в области неопределенности, не достижима в ближайшие десятилетия.

Результаты анализа возможностей расчетов точных распределений симметричных аддитивно разделимых статистик $S_n$ частотным методом и взаимного расположения областей применений точных и предельных распределений статистик для построения критериев согласия показывают нерешенность проблемы выбора распределения вероятности значений статистик для построения критерия согласия для текстов, параметры которых $(n,N)$ находятся в области неопределенности. Следовательно, применение частотного метода даже для более узкого класса симметричных аддитивно разделимых статистик $S_n$ не решает проблемы неопределенности в выборе вида распределения при построении критериев согласия и не снимает актуальности разработки методов расчета точных распределений статистик.

**Заключение и выводы**

В работе рассмотрен частотный метод расчета точных распределений значений симметричных аддитивно разделимых статистик, применяемых при анализе текстов для построения статистических критериев согласия. Проведен расчет параметров границ применения частотного метода.

Показана возможность применения частотного метода для расчета точных распределений значений симметричных аддитивно разделимых статистик для широкого спектра параметров текстов.

Проведен расчет границы области параметров текстов – области точных распределений, для которых на современном этапе за «приемлемое» время частотным методом могут быть рассчитаны точные распределения.

Проведено сравнение частотного метода расчета точных распределений с методом полного перебора, показано, что для класса симметричных аддитивно разделимых статистик частотный метод позволяет значительно расширить спектр параметров текстов, для которых могут быть рассчитаны точные распределения.

Показано, что в случае применения частотного метода, как и в случае применения метода полного перебора, ни увеличение производительности вычислительных средств, ни разумное увеличение времени расчета не приводит к практически значимому увеличению значений параметров текстов, для которых могут быть рассчитаны точные распределения.

Сравнение границы области предельных распределений и границы области точных распределений, полученных частотным методом, показало наличие между ними, как и в случае применения для расчета точных распределений метода полного перебора, непрерывной области неопределенности, содержащей параметры, для которых не могут быть рассчитаны точные распределения и неприменимы предельные распределения, так как их применение приводит к потере размерности статистического критерия и ставит под сомнение правильность решения о проверке гипотезы.

Рассчитаны и проанализированы значения производительности вычислительных средств, необходимой для расчета частотным методом точных распределений в области неопределенности. Сделан вывод о том, что требуемая производительность практически не достижима в ближайшее время.

Необходимость проведения статистического анализа текстов на всём практическом спектре их параметров требует ликвидации области неопределенности. Как показано в статье, применение частотного метода расчета точных распределений как альтернативы тривиальному методу полного перебора не решает данной задачи. Решение задачи по расчету точных распределений в области неопределенности может лежать в направлении разработки новых подходов к понятию точных распределений, что является предметом дальнейших исследований автора.

Автор выражает глубокую благодарность д.ф.-м.н., профессору А.Ф. Ронжину за постоянное внимание к работе и её обсуждение.

*Литература*

1. Чеповский А.М. Информационные модели в задачах обработки текстов на естественных языках. − М.: Национальный открытый университет «ИНТУИТ», 2015. − 228 с.

2. Зелюкин Н.Б. Сложность расчета точных распределений вероятности значений статистик и область применения предельных распределений / Н.Б. Зелюкин, А.К. Мельников // Электронные средства и системы управления: матер. докладов XIII Междунар. науч.-практ. конф. (29 ноября – 1 декабря 2017 г.): в 2 ч. – Томск: В-Спектр, 2017. – Ч. 2. – 258 с. – С. 84–90.

3. Helmert P.R. Uber die Wahrscheinlichkeit von Potenzsummen der Beobachtungsfehler etc. // Z.f. Math. u. Phys. – 1876. – B. 21. – P. 102–219.

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

5. Колодзей А.В. Критерии согласия для схемы выбора без возвращения, основанные на заполнении ячеек в обобщенной схеме размещения // Тезисы докл. VI Междунар. Петрозаводской конф. «Вероятностные методы в дискретной математике», Петрозаводск, 10–16 июня 2004 г. – Обозрение прикладной и промышленной математики. – 2004. – Т. 11, № 2. – С. 239–240.

6. Ивченко Г.И. Математическая статистика / Г.И. Ивченко, Ю.И. Медведев. − М.: Книж. дом «ЛИБРОКОМ», 2014. – 352 с.

7. 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 // Philos. Mag. Ser. 5. – 1900. – Vol. 50, No. 302. – P. 157−170.

8. Smith P.F. Exact and approximate distributions of the chi-squared statistic for equiprobability / P.F. Smith, D.S. Rae, R.W. Manderscheid, D.S. Silbergel // Commun. Stat., – 1979. – Vol. 8, No. 2. – P. 131−149.

9. Холл М. Комбинаторика. − М.: Мир, 1970. – 424 с.

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

11. Описание языка программирования Python [Электронный ресурс]. – Режим доступа: www.python.or, www.python.org/docs, свободный (дата обращения: 20.08.2017).

12. Мельников А.К. Исследование путей модернизации реконфигурируемых вычислительных систем в интересах решения вычислительно трудоемких задач // Вестник компьютерных и информационных технологий. – М.: Изд. дом «Спектр», 2016. − № 2 (140). – С. 52−59.

13. Фишер Р.А. Статистические методы для исследователей. − М.: Госстатиздат, 1958. − 73 с.

14. Кендалл М.Г. Теория распределений / М.Г. Кендалл, А. Стьюарт. − М.: Наука, 1966. – 302 с.

15. Крамер Г. Математические методы статистики. − М.: Мир, 1975. − 648 с.

16. Реконфигурируемые мультиконвейерные вычислительные структуры / В.А. Каляев, И.И. Левин, Е.А. Семерников, В.И. Шмойлов. – Ростов-н/Д: Изд-во ЮНЦ РАН, 2008. – 397 с.

---

**Мельников Андрей Кимович**
Канд. техн. наук, доцент ВАК по специальности,
гл. науч. сотр. НТЦ ЗАО «ИнформИнвестГрупп»
Тел.: +7 (495) 287-00-35
Эл. почта: ak@iigroup.ru

Melnikov A.K.

**Processing complexity in exacting probability distributions of symmetrical additively partitioned statistics and application area of limit distributions**

In the paper the authors consider application of the frequency method to calculate the exact probability distributions of symmetrical additively partitioned statistics, and analyze its computational and time complexity. The values of boundary parameters are investigated, for which, at present, could be calculated the exact distributions using the frequency method. The possibilities of exact distributions calculation with the help of the frequency method and the trivial method of full enumeration are compared. Besides, the analyses is done for the values of boundary parameters of texts, for which it is now possible to calculate exact distributions. The application possibilities for exact and limit probability distributions of statistics are considered, and could be used further for fitting criterion creation within symbolic data analysis.

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