УДК 519.213.42+519.712

Алгоритм точного вычисления распределений разделимых статистик и его применения

А.М. Зубков[*], **М.В. Филина**[*]

[*] *Математический институт им. В.А. Стеклова Российской академии наук, 119991, г. Москва, ул. Губкина, д. 8*

Аннотация. Описан вычислительно эффективный алгоритм точного вычисления распределений разделимых статистик. Он был применен для вычисления распределений статистики Пирсона и некоторых статистик в полиномиальных схемах размещений частиц по ячейкам. Проведено сравнение полученных точных распределений для конечных выборок с соответствующими предельными распределениями.

Ключевые слова: точные распределения статистик, точные вычисления распределений.

1. Введение

Для проверки статистических гипотез используются различные критерии, распределения статистик которых обычно аппроксимируются стандартными распределениями из соответствующих предельных теорем. Но вопрос о точности таких аппроксимаций в случаях небольших объемов выборок изучен недостаточно. Кроме того, в некоторых прикладных задачах необходимо обеспечить малое значение вероятности ошибки первого рода, и в этом случае требуются точные оценки хвостов распределений. Точное вычисление распределений во многих случаях сопряжено с высокой вычислительной сложностью или большим объемом необходимой оперативной памяти.

Данная задача исследовалась различными авторами. Например, для вычисления функции распределения статистики Пирсона в [1] использовались производящие функции, а в [2] — быстрое преобразование Фурье. Методы вычисления распределений разделимых статистик (также использующие производящие функции) были предложены в [3]. В данной работе описан вычислительно эффективный алгоритм расчета точных распределений разделимых статистик в полиномиальной схеме наблюдений. Этот алгоритм основан на применении неоднородных по времени цепей Маркова.

2. Основная часть

Разделимыми статистиками называются случайные величины следующего вида:

$$\zeta = \sum_{j=1}^{N} f_j(\nu_j(T, \mathbf{p})) \tag{1}$$

где $f_1(x), \ldots, f_N(x)$ — заданные функции, а $\nu_1(T, \mathbf{p}), \ldots, \nu_N(T, \mathbf{p})$ — абсолютные частоты исходов $1, \ldots, N$ в T испытаниях по полиномиальной схеме с вектором $\mathbf{p}$ вероятностей исходов.

Например, если $f_j(x) = \frac{(x-Tq_j)^2}{Tq_j}$ $(j = 1, \ldots, N)$, то

$$\zeta = \sum_{j=1}^{N} f_j(\nu_j) = \sum_{j=1}^{N} \frac{(\nu_j - Tq_j)^2}{Tq_j} \tag{2}$$

является статистикой Пирсона; если $\mathbf{q} = \text{const}$, $\min\{q_1, \ldots, q_N\} > 0$ и $T \to \infty$, то при $\mathbf{p} = \mathbf{q}$ распределения ζ сходятся к распределению хи-квадрат с $N-1$ степенями свободы, а в схеме серий с $\mathbf{q} = \text{const}$ и $\mathbf{p} = \mathbf{q} + \mathbf{a}/\sqrt{T}$ распределения ζ аппроксимируются нецентральными распределениями хи-квадрат.

При небольших значениях T и N таблицы распределения разделимой статистики (1) можно составлять с помощью ЭВМ, перебирая все C_{T+N-1}^{N-1} возможных значений вектора частот $(\nu_1, \ldots, \nu_N)$: если $x_0 = -\infty < x_1 < \ldots < x_M = \infty$, то при любом $m = 1, \ldots, M$

$$\mathbf{P}\left\{x_{m-1} < \sum_{j=1}^{N} f_j(\nu_j) \leqslant x_m\right\} = \sum_{\substack{k_1,\ldots,k_N \geqslant 0,\ k_1+\ldots+k_N=T \\ x_{m-1} < \sum_{j=1}^{N} f_j(k_j) \leqslant x_m}} \frac{T!}{k_1! \ldots k_N!} \prod_{j=1}^{N} p_j^{k_j}.$$

Но вычислительная сложность такого перебора ограничивает применимость этого метода только небольшими значениями параметров N и T.

Для построения более эффективного алгоритма вычисления точных распределений разделимых статистик рассмотрим неоднородную по времени конечную двумерную цепь Маркова $\zeta_n = (\zeta_n^{(1)}, \zeta_n^{(2)})$, $n = 0, 1, \ldots, N$, с множеством состояний $\{0, \ldots, N\} \times S$, где S — множество всех возможных значений $\sum_{j=1}^{n} f_j(k_j)$, $n = 1, \ldots, N$, $k_1, \ldots, k_N \geq 0$, $k_1 + \ldots + k_N = T$.

Таким образом,

$$\begin{aligned}
&\left(\zeta_0^{(1)}, \zeta_0^{(2)}\right) = (0, 0)\,, \\
&\left(\zeta_n^{(1)}, \zeta_n^{(2)}\right) = \left(\sum_{j=1}^{n} \nu_j, \sum_{j=1}^{n} f_j\left(\nu_j(T, \mathbf{p})\right)\right), n = 1, 2, \ldots, N-1, \\
&\left(\zeta_N^{(1)}, \zeta_N^{(2)}\right) = (T, \zeta(T, N))\,.
\end{aligned} \tag{3}$$

Тогда $\mathbf{P}\{\zeta_0 = (0,0)\} = 1$ и переходные вероятности цепи (3) имеют вид

$$p_{n-1}((u,v),(x,y)) = \mathbf{P}\{\zeta_n = (x,y) | \zeta_{n-1} = (u,v)\} = \tag{4}$$
$$= \begin{cases} C_{T-u}^{x-u} \left(\frac{p_n}{P_n}\right)^{x-u} \left(1 - \frac{p_n}{P_n}\right)^{T-x}, & 0 \leqslant u \leqslant x \leqslant T,\ y = v + f_n(x-u), \\ 0, & \text{в остальных случаях,} \end{cases}$$

где $P_n = \sum_{k=n}^{N} p_k$, $n = 1, \ldots, N$.

Распределение разделимой статистики (1) совпадает с распределением компоненты $\zeta_N^{(2)}$ в цепи Маркова (3).

В общем случае (когда функции $f_j : \{0, 1, \ldots\} \to \mathbb{R}$ могут принимать произвольные действительные значения) непосредственное вычисление распределения ζ по формулам (4) может оказаться практически невозможным, так как для хранения промежуточных распределений ζ_n требуются массивы объема $O(N \max\{C_{T+k}^k \colon k = 1, \ldots, N-1\})$, что даже при умеренных значениях T и N потребует слишком большого объема оперативной памяти. Достаточный для вычислений объем памяти существенно уменьшается, если все функции f_j — целочисленные и неотрицательные.

Например, при вычислении распределения статистики Пирсона в случае, когда $q_1 = q_2 = \ldots = q_N = 1/N$, можно преобразовать формулу (2) к виду

$$X_N^2 = \sum_{j=1}^{N} \frac{(\nu_j - T/N)^2}{T/N} = \frac{N}{T} \left(\sum_{j=1}^{N} \left(\nu_j - \left\langle \frac{T}{N} \right\rangle \right)^2 - N \left(\frac{T}{N} - \left\langle \frac{T}{N} \right\rangle \right)^2 \right), \tag{5}$$

где $\langle x \rangle = [x + \frac{1}{2}]$ обозначает ближайшее целое к x. Тогда все слагаемые суммы $\sum_{j=1}^{N} (\nu_j - \langle T/N \rangle)^2$ — целые числа, ее максимальное значение не превосходит NT^2, следовательно, распределение этой суммы можно вычислять по формулам (4), используя память объема $O(NT^3)$, за $O(N^2T^3)$ арифметических операций.

Если функции f_j неотрицательны, то траектории цепи Маркова ζ_n монотонно не убывают по каждой координате. Поэтому при вычислении функции распределения статистики Пирсона в области значений, не превышающих заданное значение M, для сокращения времени вычислений и объема используемой памяти можно не учитывать значения сумм вида $\sum_{j \geq 1} (\nu_j - T/N)^2$, превышающие MT/N. Достаточный объем памяти тогда оценивается величиной $O(MT^2/N)$ при вычислениях по формуле (5).

Несколько примеров конкретных разделимых статистик, к которым применим описанный алгоритм, имеются в [5], [6]. Результаты

точных вычислений распределений статистики Пирсона для различных полиномиальных схем приведены в [7], [8], [9]. Алгоритмы были реализованы в виде программ на языке C++. В частности, для равновероятных схем они позволяют вычислять распределения с числом исходов порядка нескольких сотен, а испытаний — нескольких тысяч.

Алгоритм был применен также для вычисления других статистик в полиномиальных схемах размещений частиц по ячейкам. Например, предельными распределениями числа пустых ячеек $\mu_0(N,T)$ могут быть распределение Пуассона или нормальное распределение в зависимости от предельных соотношений между параметрами N и T (см. [4]). Для вычисления точного распределения числа пустых ячеек можно воспользоваться приведенным выше алгоритмом, рассмотрев статистику $\mu_0(N,T)$ как сумму функций-индикаторов $f_j(\nu_j) = I\{\nu_j = 0\}$, где ν_j — число частиц в j-й ячейке. При $f_j(\nu_j) = I\{\nu_j = r\}$ статистика $\mu_r(N,T)$ представляет число ячеек, в которых находится ровно r частиц, где $r = 0, 1, 2, \ldots$

Проведенные вычисления показали, что при некоторых сочетаниях параметров хвосты точных распределений отличаются от хвостов обычно используемых аппроксимаций в несколько раз.

3. Заключение

Предложенный алгоритм позволяет вычислять точные распределения разделимых статистик при конкретных значениях параметров вероятностных схем. Это может быть полезным для уточнений статистических выводов в случаях, когда объем выборки ограничен и критические значения, выбранные по предельному и точному распределениям, существенно различаются. Применение алгоритма не ограничивается приведенными в данной работе статистиками и может быть существенно расширено.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 14-50-00005).

Литература

1. *Holzman G. I., Good I. J.* The Poisson and chi-squared approximation as compared with the true upper-tail probability of Pearson's χ^2 for equiprobable multinomials // J. Statist. Plann. Inference. — 1986. — Vol.13, no. 3. — P. 283–295.
2. *Good I. J., Gover T. N., Mitchell G. J.* Exact distributions for χ^2 and for likelihood-ratio statistic for the equiprobable multinomial distribution // J. Amer. Statist. Assoc. — 1970. — Vol. 65. — P. 267–283.

3. *Селиванов Б. И.* О вычислении допредельных распределений разделимых статистик полиномиальной схемы // Дискретная математика. — 2006. — Т. 18, вып. 3. — С. 85–94.
4. *Колчин В. Ф., Севастьянов Б. А., Чистяков В. П.* Случайные размещения. — М.: Наука, 1976.
5. *Зубков А. М.* Рекуррентные формулы для распределений функционалов от дискретных случайных величин // Обозр. прикл. промышл. математики. — 1996. — Т. 3, вып. 4. — С. 567–573.
6. *Зубков А. М.* Методы расчета распределений сумм случайных величин // Труды по дискретной математике. — М.: Физматлит, 2002. — Т. 5. — С. 51–60.
7. *Filina M. V., Zubkov A. M.* Exact computation of Pearson statistics distribution and some experimental results // Austrian J. Statist. — 2008. — Vol. 37, no. 1. — P. 129–135.
8. *Filina M. V., Zubkov A. M.* Tail properties of Pearson statistics distributions // Austrian J. Statist. — 2011. — Vol. 40, no. 1 & 2. — P. 47–54.
9. *Filina M. V., Zubkov A. M.* Some remarks on the noncentral Pearson statistics distributions // Computer Data Analysis and Modeling. Proc. XI Int. Conf., Minsk, 2016. — Publ. center BSU. — P. 155–158.

UDC 519.213.42+519.712

Algorithm of exact computation of divisible statistics distributions and its applications

A.M. Zubkov[*], **M.V. Filina**[*]

[*] *Steklov Mathematical Institute of Russian Academy of Sciences, 8 Gubkina St., Moscow 119991, Russia*

Computationally efficient algorithm realizing exact computation of divisible statistics distributions for multinomial scheme is described. The algorithm is based on the embedded nonhomogeneous Markov chain. It was applied for computation of the Pearson statistics distribution and distributions of some statistics of random allocation of particles into cells. Comparisons of exact numerical values of distribution functions of statistics with usually used approximations from corresponding limit theorems show that exact tail probabilities may be in several times larger than that of approximating distributions.

Keywords: exact distributions of statistics, exact computation of distributions.