Специализированные программные системы для проведения аналитических выкладок как инструмент решения вероятностных задач со случайным разбиением интервала[1]

А.Л.Резник[2], В.М.Ефимов[2], А.А.Соловьев[2], А.В.Торгов[2]

[2]Институт автоматики и электрометрии, Сибирское Отделение Российской Академии Наук, 63090 г. Новосибирск, проспект академика Коптюга, 1, Россия,
E-mail: reznik@iae.nsk.su, torgov@iae.nsk.su, solowey@rambler.ru

Многочисленные научные и технические приложения, относящиеся к обработке случайных дискретных изображений, теории массового обслуживания, технической диагностике и другим дисциплинам [1-4], приводят к следующей вероятностной задаче, связанной со случайным разбиением интервала:

"Пусть n точек случайно брошены на интервал (0,1), т.е. имеется n независимых испытаний случайной величины, равномерно распределенной на интервале (0,1). Необходимо найти вероятность $P_{n,k}(\varepsilon)$ того, что внутри интервала (0,1) не содержится ни одного подынтервала Ω_ε длиной ε, содержащего более k точек."

Несмотря на кажущуюся простоту поставленной задачи, ее аналитическое решение известно (Parzen, E. (1960) Modern Probability Theory and Its Applications, John Wiley and Sons, Inc., New York-London) [5] лишь для случая, когда k=1:

$$P_{n,1}(\varepsilon) = (1-(n-1)\varepsilon)^n, \quad (0 \le \varepsilon \le 1/(n-1)). \tag{1}$$

Наиболее доступный способ получения этой формулы состоит в представлении искомой вероятности в виде повторного интеграла

$$P_{n,1}(\varepsilon) = n! \int_{(n-1)\varepsilon}^{1} dx_n \left\{ \int_{(n-2)\varepsilon}^{x_n-\varepsilon} dx_{n-1} \ldots \left\{ \int_{2\varepsilon}^{x_4-\varepsilon} dx_3 \left\{ \int_{\varepsilon}^{x_3-\varepsilon} dx_2 \left\{ \int_{0}^{x_2-\varepsilon} dx_1 \right\} \right\} \right\} \right\} \tag{2}$$

с дальнейшим последовательным интегрированием по всем переменным x_1, x_2,..., x_n (множитель $n!$ перед интегралом связан с количеством перестановок из n элементов, поскольку запись (2) предполагает, что точки x_1, x_2,..., x_n упорядочены по возрастанию).

К сожалению, для значений k>1 вероятность $P_{n,k}(\varepsilon)$ не удается представить в такой же компактной форме, как при k=1. Более того, в общем случае вероятность $P_{n,k}(\varepsilon)$, в отличие от случая k=1, описывающегося формулой (1), представляет собой не единое полиномиальное выражение, а набор полиномов от параметра ε, имеющих степень n и непрерывно «сшитых» в нескольких местах.

Из-за сложностей, возникающих при попытке найти общий вид соотношений для вероятности $P_{n,k}(\varepsilon)$, нами были предприняты усилия по поэтапному решению

[1] Настоящая работа частично поддержана Российским фондом фундаментальных исследований (проект № 10-01-00458), Президиумом РАН (проект № 228/2009) и Президиумом СО РАН (интеграционный проект № 71/2009).

сформулированной задачи. В первую очередь эти усилия были направлены на то, чтобы попытаться найти частные формулы для вероятности $P_{n,k}(\varepsilon)$ при фиксированных значениях n и k, которые в дальнейшем могли бы помочь в отыскании общих закономерностей их образования и тем самым «навести» исследователя на правильный ответ.

Вообще говоря, общее решение задачи может быть представлено в виде многомерного интеграла

$$P_{n,k}(\varepsilon) = n! \int\limits_{D_{n,k}(\varepsilon)} \dots \int dx_1 \dots dx_n, \tag{3}$$

где область интегрирования $D_{n,k}(\varepsilon) \subset R^n$ описывается системой линейных неравенств

$$\begin{cases} 0 < x_1 < x_2 < \dots < x_{n-1} < x_n < 1, \\ x_{k+1} - x_1 > \varepsilon, \\ x_{k+2} - x_2 > \varepsilon, \\ \vdots \\ x_n - x_{n-k} > \varepsilon. \end{cases} \tag{4}$$

Суть первого из предложенных нами подходов, который мы назвали методом прямого аналитического интегрирования, такова. Интеграл (3) по области (4) записывается в эквивалентной форме:

$$P_{n,k}(\varepsilon) = n! \int\limits_{-\infty}^{+\infty} \dots \int 1[x_1]1[x_2] \dots 1[x_n - x_{n-1}]1[1 - x_n]1[x_{k+1} - x_1 - \varepsilon] \times$$

$$\times 1[x_{k+2} - x_2 - \varepsilon] \times \dots \times 1[x_n - x_{n-k} - \varepsilon] dx_1 \dots dx_n, \tag{5}$$

т.е. интегрирование уже ведется по всему пространству, а индикаторная функция области интегрирования $D_{n,k}(\varepsilon)$ трансформируется в произведение единичных функций $1[z] = \begin{cases} 0, & z \le 0, \\ 1, & z > 0, \end{cases}$ входящих в подынтегральное выражение многократного интеграла (5).

Затем n-мерный интеграл (5) с помощью последовательного применения соотношения

$$\left(\prod_{j=1}^{l} 1[x_r - \alpha_j] \right) \left(\prod_{i=1}^{m} 1[\beta_i - x_r] \right) =$$

$$= \sum_{j=1}^{l} \sum_{i=1}^{m} 1[x_r - \alpha_j]1[\beta_i - x_r] \left\{ 1[\beta_i - \alpha_j] \left(\prod_{\substack{q=1 \\ q \ne j}}^{l} 1[\alpha_j - \alpha_q] \right) \left(\prod_{\substack{s=1 \\ s \ne i}}^{m} 1[\beta_s - \beta_i] \right) \right\} \tag{6}$$

сводится к набору повторных интегралов с уже расставленными пределами интегрирования и с указанием на границы изменения параметра ε, к которым относится каждый из них. На заключительном этапе нужно лишь провести собственно интегрирование и объединить результаты.

Предложенный подход позволяет конструктивно вычислять частные формулы $P_{n,k}(\varepsilon)$ при конкретных значениях n и k, но вычислительная сложность алгоритма такова, что для значений $n>4$ практически невозможно провести расчеты «вручную». Поэтому изложенный метод был нами абсолютно формализован и на его основе в полном соответствии с описанной выше схемой создан пакет программ для *компьютерного* проведения всех необходимых аналитических преобразований, включая расчет пределов интегрирования, проверку промежуточных систем неравенств на совместность, упорядочение по мультииндексам и непосредственное интегрирование. Используя их, были вычислены формулы $P_{n,k}(\varepsilon)$ вплоть до значений n =14 (некоторые примеры таких расчетов приведены в Таблице 1).

Табл.1

n	k	Область изменения ε	$P_{n,k}(\varepsilon)$
3	2	(0,1)	$1-3\varepsilon^2+2\varepsilon^3$
4	2	(0,1/2)	$1-12\varepsilon^2+24\varepsilon^3-14\ \varepsilon^4$
		(1/2,1)	$2-8\varepsilon+12\varepsilon^2-8\varepsilon^3+2\varepsilon^4$
	3	(0,1)	$1-4\varepsilon^3+3\varepsilon^4$
….			
14	2	(0,1/8)	$1 - 1092\varepsilon^2 + 16744\varepsilon^3 + 156156\varepsilon^4 - 6498492\varepsilon^5 + 63159096\varepsilon^6$ $- 56566224\varepsilon^7 + 295079404\varepsilon^8 + 338060148\varepsilon^9 + 1741401056\varepsilon^{10}$ $- 79779000\varepsilon^{11} - 522127532\varepsilon^{12} - 701324068\varepsilon^{13} + 1460968152\varepsilon^{14}$
		(1/8,1/7)	$65 - 6552\varepsilon + 309036\varepsilon^2 - 8975512\varepsilon^3 + 178510332\varepsilon^4 - 2564958396\varepsilon^5$ $+ 27418999608\varepsilon^6 - 221377405392\varepsilon^7 + 1356313562604\varepsilon^8$ $- 6275548683404\varepsilon^9 + 21590662949856\varepsilon^{10} - 53575501829304\varepsilon^{11}$ $+ 90670532458324\varepsilon^{12} - 93708297788196\varepsilon^{13} + 44634761108184\varepsilon^{14}$
		(1/7,1/6)	$429 - 36036\varepsilon + 1405404\varepsilon^2 - 33729696\varepsilon^3 + 556539984\varepsilon^4 - 6678479808\varepsilon^5$ $+ 60106318272\varepsilon^6 - 412157611008\varepsilon^7 + 2163827457792\varepsilon^8$ $- 8655309831168\varepsilon^9 + 25965929493504\varepsilon^{10} - 56652937076736\varepsilon^{11}$ $- 84979405615104\varepsilon^{12} - 78442528260096\varepsilon^{13} + 36182263971840\varepsilon^{14}$

Пример работы программы прямого аналитического интегрирования.

Для определения асимптотического поведения вероятности $P_{n,k}(\varepsilon)$ на первом участке (при $n\varepsilon$ <<1) нами был разработан еще один метод, в основу которого положено простое соображение, что для восстановления полинома степени n достаточно знать его значение и значения всех его производных до степени n в какой-либо одной точке (например, в нуле), т.е. алгоритм вычислений основан на полиномиальном тождестве

$$P(\varepsilon) = \sum_{i=0}^{n} \frac{1}{i!} \frac{d^i P}{d\varepsilon^i}(0)\varepsilon^i . \qquad (7)$$

Значение самого полинома $P_{n,k}(\varepsilon)$ в точке ε=0, естественно, равно 1, так как первоначальная система неравенств (4) при ε=0 совместна и налагает ограничения лишь

на порядок следования переменных x_1, x_2,..., x_n. Далее находим первую производную

$$\frac{dP(\varepsilon)}{d\varepsilon} = n!\int...\int \frac{d}{d\varepsilon}\{1[...]\times...\times 1[...]\}dx_1...dx_n . \qquad (8)$$

Для этого нам потребуются следующие простые соотношения:

$$\frac{d}{dz}1[z] = \delta(z) , \qquad (9)$$

$$\int_{-\infty}^{+\infty} \delta(z)f(z)dz = f(0) , \qquad (10)$$

где δ (z) – обычная дельта-функция.

Проводя дифференцирование подынтегрального выражения в формуле (8) по обычным правилам дифференцирования произведения функций и используя при этом (9), исходное соотношение для первой производной переводится в сумму интегралов, каждый из которых можно упростить, осуществив (с использованием (10)) интегрирование по одной из переменных, входящих в аргумент соответствующей дельта-функции. Такая схема замечательна тем, что не требует расчета пределов интегрирования, а сам вычислительный процесс сводится к многократно повторяющимся операциям подстановки. Поскольку нас интересует лишь значение производной полинома в точке ε=0, фактическое интегрирование заменяется проверкой каждого из подынтегральных выражений на совместность: если нет нарушения в порядке следования переменных, то интеграл равен единице, в противном случае он равен нулю. В результате последовательного выполнения таких преобразований выражение (8) принимает вид суммы интегралов, кратность которых на единицу меньше по сравнению с первоначальным интегралом. Поскольку при этом структура подынтегральных выражений остается неизменной, сохраняется возможность повторного использования описанной процедуры для вычисления производных более высокого порядка без изменения принципиальной схемы.

Для нахождения общего решения $P_{n,2}(\varepsilon)$ на всем интервале изменения параметра ε нами был разработан специальный рекурсивно-комбинаторный метод. Краткая суть алгоритма такова. Интервал (0,*L*) интерпретируется как совокупность *r* равных дискретов. Случайное бросание *n* точек на интервал (0,*L*) интерпретируется как случайное бросание *n* неразличимых шаров по *r* ящикам. Множество из *l* смежных ячеек служит аналогом подынтервала длиной ε. Исход бросания, когда ни один из таких *l*-подынтервалов, содержащихся внутри исходного *r*-интервала (0,*L*), не имеет более 2 точек, считается “успешным”, а отношение общего числа “успешных” бросаний *Q(r,n,l)* к общему числу исходов опыта *Q(r,n)* принимается в качестве целочисленного аналога вероятности $P_{n,2}(\varepsilon)$. В принципе, нахождение замкнутого аналитического выражения для общего числа “успешных бросаний” эквивалентно полному решению задачи для *k=2*, поскольку в этом случае вероятность $P_{n,2}(\varepsilon)$ вычисляется простым предельным переходом

$$P_{n,2}(\varepsilon) = \lim_{\substack{r\to\infty \\ (l/r)\to\varepsilon}} \frac{Q(r,n,l)}{Q(r,n)}, \qquad (11)$$

а выражение для общего числа бросаний *Q(r,n)* известно (Feller, W. (1967) An Introduction to Probability Theory and Its Applications (in Russian), Mir, Moscow) [6]:

$$Q(r,n) = C_{n+r-1}^{r-1} = \frac{(n+r-1)!}{n!(r-1)!}. \qquad (12)$$

Для реализации этой идеи нами были получены рекуррентные соотношения, позволившие осуществить расчет формул *Q(r,n,l)* при последовательно возрастающих значениях *n*. Вся процедура, как и в предыдущих алгоритмах, была полностью формализована и запрограммирована.

С использованием всех описанных алгоритмов нами были проведены расчеты, с помощью которых установлено, что для произвольных *n* на интервале *0 < ε <1/(entier((n+1)/2))* вероятность $P_{n,2}(\varepsilon)$ (т.е. вероятность $P_{n,k}(\varepsilon)$ при *k=2*) представляется в виде:

$$\begin{aligned} P_{n,2}(\varepsilon) &= C_n^0 + C_n^2(-n+2)\varepsilon^2 + C_n^3(4n-10)\varepsilon^3 + C_n^4(3n^2-37n+86)\varepsilon^4 + C_n^5(-40n^2+394n- \\ &922)\varepsilon^5 + C_n^6(-15n^3+625n^2-5171n-12086)\varepsilon^6 + C_n^7(420n^3-10724n^2+79996n-187002)\varepsilon^7 \\ &+ C_n^8(105n^4-10570n^3+205499n^2-1426841n+3336406)\varepsilon^8 + C_n^9(5040n^4-155708n^3 \\ &+ 2267664n^2 - 17317506n^1 + 52315558) + C_n^{10}(-945n^5+189000n^4-15794625n^3 \\ &+ 389687181n^2 - 3798029823n + 12998966646)\varepsilon^{10} + o(\varepsilon^{10}). \end{aligned} \qquad (13)$$

Кроме того, проведенные расчеты позволили усмотреть общую закономерность и высказать гипотезу [4], заключающуюся в том, что. при *k=2* для четных *n* на интервале *1/(n/2) < ε <1/((n/2)-1)* справедлива формула

$$P_{n,2}(\varepsilon) = (2/n)C_n^{(n/2)-1}(1-((n/2)-1)\varepsilon)^n. \qquad (14)$$

Впоследствии нам удалось провести строгое математическое доказательство этой формулы [7].

Чтобы продвинуться далее в этой задаче, в качестве дополнительного метода нами предложен еще один алгоритмический подход к вычислению вероятностей $P_{n,k}(\varepsilon)$. Суть его в том, что с помощью прямого интегрирования непосредственно в общей аналитической форме последовательно отыскиваются формулы для вероятностей $P_{n,n-1}(\varepsilon)$, $P_{n,n-2}(\varepsilon)$, $P_{n,n-3}(\varepsilon)$ и т.д. Проведенные в соответствии с разработанным алгоритмом расчеты позволили установить следующие формулы:

$$P_{n,n-1}(\varepsilon) = 1 - \varepsilon^n - n\varepsilon^{n-1}(1-\varepsilon). \qquad (15)$$

$$P_{n,n-2}(\varepsilon)=\begin{cases}1-2C_n^2\varepsilon^{n-2}(1-\varepsilon)^2-2\varepsilon^n, & 0\le\varepsilon\le(1/2);\\ 1-2\varepsilon^n+(2\varepsilon-1)^n-2C_n^2\varepsilon^{n-2}(1-\varepsilon)^2, & (1/2)\le\varepsilon\le 1\end{cases} \tag{16}$$

$$P_{\substack{n,n-3\\(n>6)}}(\varepsilon)=\begin{cases}1-2\varepsilon^n+C_n^1(6\varepsilon^n-4\varepsilon^{n-1})+C_n^2(-3\varepsilon^n+\varepsilon^{n-2})+\\ \quad+C_n^3(9\varepsilon^n-18\varepsilon^{n-1}+12\varepsilon^{n-2}-3\varepsilon^{n-3}),\\ \qquad 0\le\varepsilon\le(1/2);\\ 1-2\varepsilon^n+(2\varepsilon-1)^n+C_n^1(1-\varepsilon)(-2\varepsilon^{n-1}+2(2\varepsilon-1)^{n-1})+\\ \quad+C_n^2(1-\varepsilon)^2(\varepsilon^{n-2}+(2\varepsilon-1)^{n-2})-3C_n^3\varepsilon^{n-3}(1-\varepsilon)^3,\\ \qquad (1/2)\le\varepsilon\le 1.\end{cases} \tag{17}$$

Отметим, что, в отличие всех ранее приводившихся соотношений, формулы (15)-(17) рассчитаны вручную. Формализация алгоритма для получения аналитических соотношений $P_{n,n-4}(\varepsilon)$, $P_{n,n-5}(\varepsilon)$,… программным путем представляется весьма сложной (так как выкладки необходимо программировать при двух свободных параметрах – n и ε), но вполне осуществимой задачей.

Заключение

Проведенные нами исследования показали, что, несмотря на внешнюю простоту поставленной задачи, ее решение представляет собой сложно формализуемый и весьма трудоемкий в вычислительном плане процесс. В ходе проделанной работы нам удалось провести идею, высказанную в свое время Дж. фон Нейманом: исследователь сталкивается с задачей, которую не в состоянии решить, привлекает ЭВМ для проведения трудоемких расчетов, которые способны натолкнуть его на «правильный» ответ, и в случае удачи (т.е. зная подсказанное компьютером решение) проводит строгое и конструктивное доказательство. В нашем случае формула (14) была «подсказана» компьютером, а лишь впоследствии строго доказана. Интересно также отметить, что коэффициенты, входящие в полученную нами формулу (14), являются числами Каталана, которые впервые встречаются в работах Л.Эйлера и возникают при решении огромного числа вероятностно-комбинаторных и прикладных задач.

Список литературы:

[1] *Ефимов В.М., Искольдский А.М., Лившиц З.А., Крендель Ю.М.* О характеристиках различных методов считывания изображений дискретных структур // Автометрия, 1973, №1, С.3-7.

[2] *Ефимов В.М., Резник А.Л.* Аналитическое вычисление на ЭВМ объемов, ограниченных системой гиперплоскостей в *n*-мерном пространстве //Автометрия, 1976, №1, С.116-119.
[3] *Ефимов В.М., Резник А.Л.* Аналитическое определение с помощью ЭВМ статистических характеристик процесса щелевого сканирования потока Бернулли. //Автометрия, 1977, №4, С.49-51.
[4] *Резник А.Л.* Моделирование на ЭВМ непрерывного считывания изображений дискретной структуры // Автометрия, 1981, №6, С.3-6.
[5] *Parzen, E.* Modern Probabillity Theory and Its Applications, John Wiley and Sons, Inc., New York-London, 1960
[6] *В.Феллер*. Введение в теорию вероятностей и ее приложения. М.: Мир, 1966.
[7] *A. L. Reznik, V. M. Efimov, A. V. Torgov, and A. A. Solov'ev*. Computer Analytics in Problems with a Random Partition of the Interval // Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications, 2011, Vol.21, No.2, pp.202-205.