Дискретная математика

том 24 выпуск 4 * 2012

УДК 519.2

Статистические задачи для случайных подстановок с цензурированными данными

© 2012 г. Г. И. Ивченко, М. В. Солдаткина

В d-мерной параметрической модели случайных подстановок степени n решаются задачи асимптотического при $n \to \infty$ оценивания неизвестных параметров и проверки различных гипотез о них по неполным данным, когда в подстановке наблюдаются лишь циклы некоторой ограниченной длины. Также рассматривается случай больших выборок и строится критерий однородности для него.

1. Введение

В [2] была введена следующая d-параметрическая модель случайных n-подстановок, то есть взаимно однозначных отображений множества $X_n = \{1, 2, \ldots, n\}$ в себя.

Пусть задано некоторое разбиение множества X_n:

$$X_n = \bigcup_{j=1}^{d} A_j, \quad A_i \cap A_j = \varnothing, \quad i \neq j, \tag{1}$$

и пусть $\mathbf{c}(n) = (c_1, c_2, \ldots, c_n)$ есть цикловая структура n-подстановки s, именно, c_i есть число циклов длины i, $i = 1, \ldots, n$, причем

$$\sum_i i c_i = n.$$

Циклы подстановки s, длины которых являются элементами подмножества A_j, называются A_j-циклами, их число обозначается

$$C_{A_j}(n) = \sum_{i \in A_j} c_i, \qquad j = 1, \ldots, d.$$

Если подмножество A_j имеет вид

$$A_j = \{k : k = ld + j,\ l \geqslant 0\} \tag{2}$$

для некоторых целых $d \geqslant 2$ и $1 \leqslant j \leqslant d$, то говорят о конгруэнтных циклах (см. [1], с. 187).

Далее, на множестве всех n-подстановок $S_n = \{s\}$ задается параметрическая вероятностная мера вида

$$\mathbf{P}_\theta(s) = I\left(\sum_{i=1}^{n} i c_i = n\right) \frac{1}{H_n(\theta)} \prod_{j=1}^{d} \theta_j^{C_{A_j}(n)}, \tag{3}$$

где $I(\cdot)$ — индикатор, $\theta = (\theta_1, \ldots, \theta_d)$, $\theta_j \geqslant 0$, $j = 1, \ldots, d$,— произвольные параметры, не равные нулю одновременно, и $H_n(\theta)$ — необходимый нормирующий множитель, имеющий вид

$$H_n(\theta) = n! \, [z^n] \exp\left\{\sum_{j=1}^{d} \theta_j \sum_{i \in A_j} \frac{z^i}{i}\right\}; \tag{4}$$

здесь и далее

$$[z^n] f(z) = \mathrm{coef}_{z^n} f(z).$$

В [2] доказана асимптотическая, при $n \to \infty$, нормальность вектора

$$\mathbf{C}(n) = (C_{A_1}(n), \ldots, C_{A_d}(n))$$

чисел конгруэнтных циклов случайной подстановки в такой модели и решены соответствующие задачи проверки статистических гипотез о параметре $\theta = (\theta_1, \ldots, \theta_d)$.

Подчеркнем, что в этих исследованиях предполагается известной вся цикловая последовательность $\mathbf{c}(n) = (c_1, c_2, \ldots, c_n)$, то есть имеется полная информация о наблюдаемой подстановке, и соответствующие выводы имеют асимптотический (при $n \to \infty$) характер. Но в этом случае не всегда является реалистичным предположение о том, что мы можем наблюдать всю цикловую последовательность $\mathbf{c}(n)$. Может быть и так, что наблюдению доступно лишь какое-то ограниченное число k ее первых членов $c_1, c_2, \ldots, c_k$, в этом случае говорят о цензурированных (неполных) данных. Статистические задачи для случайных подстановок с неполными данными в рамках однопараметрической модели Эванса, когда подстановка $s \in S_n$ наблюдается с вероятностью, пропорциональной $\theta^{c(n)}$, где

$$c(n) = c_1 + c_2 + \ldots + c_n$$

есть общее число циклов подстановки s, рассматривались в [4]. В настоящей работе аналогичный подход применяется к описанной выше d-параметрической модели с конгруэнтными циклами.

Именно, будем предполагать, что в наблюдаемой подстановке $s \in S_n$ для каждого $j = 1, \ldots, d$ доступно подсчету лишь число A_j-циклов с длинами, не превосходящими заданного уровня K_j. В таком случае, пусть

$$\xi_j(n) = \sum_{i \in A_j} c_i I(i \leqslant K_j), \qquad j = 1, \ldots, d, \tag{5}$$

есть наши исходные данные (количества наблюдаемых A_j-циклов). Рассмотрим различные вопросы статистического вывода для модели (1)–(4) именно по таким неполным данным (5). При этом будем предполагать, что порядок подстановки $n \to \infty$, а параметры цензурирования K_j, $j = 1, \ldots, d$, фиксированы.

Основой для дальнейших выводов будет служить следующее утверждение об асимптотическом распределении наблюдаемых статистик $c_1, c_2, \ldots$, то есть начальных членов цикловой структуры подстановки.

Теорема 1. *Для случайной подстановки s в модели* (1)–(4) *при* $n \to \infty$ *числа циклов ограниченной длины асимптотически независимы, и при этом число A_j-циклов длины i имеет в пределе распределение Пуассона* $\Pi(\theta_j / i)$.

Доказательство этой теоремы приведено в приложении.

Следствие 1. *Наблюдаемые статистики* (5) *асимптотически независимы, и*

$$\mathcal{L}(\xi_j(n)) \to \Pi(\theta_j \lambda_j), \qquad \lambda_j = \sum_{i \in A_j} \frac{1}{i} I(i \leqslant K_j) = \sum_{l \leqslant (K_j - j)/d} \frac{1}{ld + j}. \tag{6}$$

Из этих результатов следует, во-первых, что статистические выводы о каждом из параметров θ_j можно делать независимо по наблюдению лишь соответствующей статистики $\xi_j(n)$, и, во-вторых, исходная проблема в асимптотике сводится к соответствующим статистическим задачам для пуассоновской модели с неизвестным параметром, решение которых достаточно хорошо известно. Используя известные результаты для пуассоновской модели, мы в разделе 2 приводим решение задач точечного и доверительного оценивания параметров θ_j нашей модели, и в разделе 3 — задач проверки статистических гипотез. В разделе 4 анализируется многовыборочный случай (наблюдается большое число независимых подстановок), и в разделе 5 решается задача проверки гипотезы однородности для этой ситуации.

2. Оценивание параметров

Пусть имеется N независимых подстановок $s_1, \ldots, s_N$, полученных при одном и том же значении неизвестного параметра $\theta = (\theta_1, \ldots, \theta_d)$, и, тем самым, имеется выборка (набор наблюдаемых статистик (5))

$$(\xi_1{}^{(k)}(n), \ldots, \xi_d{}^{(k)}(n)), \qquad k = 1, \ldots, N. \tag{7}$$

Как отмечено во введении, компоненты этих векторов асимптотически, при $n \to \infty$ и фиксированных уровнях цензурирования K_j, независимы и распределены в соответствии с (6), поэтому для суммарного числа наблюдаемых A_j-циклов

$$T_j = T_j(N, n) = \sum_{k=1}^{N} \xi_j{}^{(k)}(n) \tag{8}$$

будет выполняться предельное соотношение

$$\mathcal{L}(T_j(N, n)) \to \Pi(N\theta_j \lambda_j). \tag{9}$$

Это соотношение сводит задачу оценивания параметров θ_j модели (1)–(4) к задаче оценивания неизвестного параметра пуассоновской модели.

Известно (см., например, [6, §3.4]), что оптимальной, то есть несмещенной с минимальной дисперсией, оценкой сходящегося при всех $\theta > 0$ степенного ряда

$$\tau(\theta) = \sum_{i \geqslant 0} a_i \theta^i$$

по наблюдению над случайной величиной X с пуассоновским распределением $\Pi(\theta)$ является статистика

$$\tau^* = \sum_{i \geqslant 0} a_i (X)_i,$$

где

$$(X)_i = X(X-1)\cdots(X-i+1), \quad i \geqslant 1, \quad (X)_0 = 1.$$

Учитывая это и соотношение (9), мы можем сформулировать следующее общее утверждение для нашей модели.

Предложение 1. *Асимптотически оптимальной оценкой сходящегося при всех $\theta_j > 0$ степенного ряда*

$$\tau(\theta_j) = \sum_{i \geqslant 0} a_i \theta_j^i$$

является статистика

$$\tau^* = \sum_{i \geqslant 0} \frac{a_i (T_j(N,n))_i}{(N\lambda_j)^i};$$

в частности, асимптотически оптимальная оценка параметра θ_j имеет вид

$$\theta_j^* = \frac{T_j(N,n)}{N\lambda_j}. \tag{10}$$

Соответствующий же γ-доверительный интервал для θ_j асимптотически имеет вид

$$\left(\frac{1}{2N\lambda_j} \chi^2_{(1-\gamma)/2, 2T_j}, \frac{1}{2N\lambda_j} \chi^2_{(1+\gamma)/2, 2T_j+2} \right), \tag{11}$$

где $\chi^2_{p,r}$ есть p-квантиль распределения хи-квадрат с r степенями свободы.

3. Проверка гипотез

Если требуется проверить какие-то гипотезы о частных значениях параметров θ_j в обсуждаемой ситуации, то на основании соотношения (9) надо воспользоваться общей теорией для пуассоновской модели, учитывая при этом специфику тестовых статистик $T_j(N,n)$, с тем лишь замечанием, что соответствующие алгоритмы в нашем случае будут иметь характер асимптотических утверждений. Продемонстрируем это на конкретном примере проверки простой гипотезы $H_0: \theta_j = \theta_{j0}$ против правосторонней альтернативы H_1^+: $\theta_j > \theta_{j0}$. Известно (см. [6], §5.3), что в задачах такого типа существует равномерно наиболее мощный (РНМ) критерий, который строится следующим образом. Обозначим в нашем случае

$$\theta_0 = \theta_{j0} N\lambda_j,$$

и при заданной вероятности ошибки первого рода α определим целое число t_α условием

$$\alpha'' \equiv \sum_{m=t_\alpha+1}^{\infty} \frac{\theta_0^m}{m!} e^{-\theta_0} < \alpha \leqslant \sum_{m=t_\alpha}^{\infty} \frac{\theta_0^m}{m!} e^{-\theta_0} \equiv \alpha'. \tag{12}$$

Если здесь $\alpha = \alpha'$, то искомый критерий является нерандомизированным и асимптотически задается критической областью

$$\mathscr{X}_\alpha^+ = \{T_j(N,n) \geqslant t_\alpha\}. \tag{13}$$

Если же в (12) имеет место строгое неравенство $\alpha < \alpha'$, то критерий является рандомизированным и задается критической функцией

$$\Phi_\alpha(T_j) = \begin{cases} 1, & T_j > t_\alpha, \\ \dfrac{\alpha - \alpha''}{\alpha' - \alpha''}, & T_j = t_\alpha, \\ 0, & T_j < t_\alpha. \end{cases}$$

В любом случае мощность этого критерия при произвольной альтернативе

$$\theta_1 = \theta_j N \lambda_j > \theta_0$$

вычисляется по формуле

$$W(\theta_1) \sim \sum_{m=t_\alpha+1}^{\infty} \frac{\theta_1^m}{m!} e^{-\theta_1} + (\alpha - \alpha'') \left(\frac{\theta_1}{\theta_0}\right)^{t_\alpha} e^{\theta_0 - \theta_1}. \tag{14}$$

Замечание 1. В случае $\alpha < \alpha'$, вместо рандомизированного критерия можно также использовать нерандомизированные РНМ критерии

$$\mathscr{X}_{\alpha'}^+ = \{T_j \geqslant t_\alpha\}$$

или

$$\mathscr{X}_{\alpha''}^+ = \{T_j \geqslant t_\alpha + 1\}$$

с уровнями значимости, асимптотически равными соответственно α' и α''.

Аналогично анализируется задача (H_0, H_1^-) с левосторонней альтернативой, а при двусторонней альтернативе критерий задается объединением двух односторонних критических областей, то есть имеет вид

$$\mathscr{X}_\alpha = \{T_j \geqslant t_{\alpha_1}\} \cup \{T_j \leqslant t_{\alpha_2}\}, \qquad \alpha_1 + \alpha_2 = \alpha.$$

Замечание 2. Случай $\theta_{j0} = 1$, $j = 1, \ldots, d$, соответствует наиболее важной для приложений гипотезе о равновероятности подстановок, и изложенная методика дает новые критерии проверки этой гипотезы, учитывающие широкий класс специальных альтернатив, при которых $\theta_j \neq 1$ для некоторых j.

4. Большие выборки

Если число N наблюдаемых подстановок велико, то можно применить теорию больших выборок (при $N \to \infty$) и получить более сильные выводы. В этом случае из (8)–(9) следует, что для нормированной статистики

$$\widetilde{T}_j = \widetilde{T}_j(N,n) = \frac{T_j(N,n)}{N\lambda_j} \tag{15}$$

справедлива нормальная аппроксимация

$$\mathcal{L}(\tilde{T}_j) \sim \mathcal{N}\left(\theta_j, \frac{\theta_j}{N\lambda_j}\right), \tag{16}$$

и, тем самым, статистические выводы о параметрах θ_j можно получать, используя соответствующую теорию нормальной модели [4]. Аналогичная ситуация рассматривалась в [3], поэтому, следуя ей, можно сформулировать для нашего случая следующие утверждения.

Предложение 2. *Статистика $\tilde{T}_j$ является асимптотически несмещенной и асимптотически эффективной оценкой параметра θ_j; такими же свойствами обладает статистика $\tau(\tilde{T}_j)$ как оценка $\tau(\theta_j)$ для любой дифференцируемой функции τ.*

Предложение 3. *Асимптотическим γ-доверительным интервалом для параметрической функции $\tau(\theta_j)$ с непрерывной производной $\tau'(\theta_j)$ является интервал*

$$\left(\tau(\tilde{T}_j) \mp z_\gamma \tau'(\tilde{T}_j)\sqrt{\frac{\tilde{T}_j}{N\lambda_j}}\right), \tag{17}$$

где

$$z_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

и $\Phi^{-1}(t)$ — обратная функция к стандартной нормальной функции распределения $\Phi(x)$.

Предложение 4. *Критерий уровня значимости α для проверки гипотезы $H_0\colon \theta_j = 1$ при левосторонней альтернативе $H_1^-\colon \theta_j < 1$ асимптотически задается критической областью*

$$\mathcal{K}_\alpha^-(N) = \{\tilde{T}_j < 1 - u_\alpha/\sqrt{N\lambda_j}\}, \qquad u_\alpha = \Phi^{-1}(1-\alpha); \tag{18}$$

пороговыми здесь являются альтернативы вида

$$H_{1N}^-\colon \theta_j = \theta_j(N) = 1 - \frac{t}{\sqrt{N\lambda_j}}, \qquad t > 0,$$

и мощность критерия (18) *при таких близких альтернативах удовлетворяет при $N \to \infty$ соотношению*

$$W_N(\theta_j) = \mathbf{P}(\mathcal{K}_\alpha^-(N)) \to \Phi(t - u_\alpha).$$

Аналогично, в задаче $(H_0, H_1^+\colon \theta_j > 1)$ критерий уровня значимости α асимптотически имеет вид

$$\mathcal{K}_\alpha^+(N) = \{\tilde{T}_j > 1 + u_\alpha/\sqrt{N\lambda_j}\}, \tag{19}$$

и его мощность при близких альтернативах

$$H_{1N}^+\colon \theta_j = \theta_j(N) = 1 + \frac{t}{\sqrt{N\lambda_j}}, \qquad t > 0,$$

асимптотически равна $\Phi(t-u_\alpha)$.

Наконец, в задаче с двусторонней альтернативой $(H_0, H_1 = H_1^- \cup H_1^+)$ *критерий асимптотически имеет вид*

$$\mathscr{K}_\alpha(N) = \left\{ |\widetilde{T}_j - 1| > \frac{u_{\alpha/2}}{\sqrt{N\lambda_j}} \right\}, \tag{20}$$

и его мощность при близких альтернативах

$$H_{1N}: \theta_j = \theta_j(N) = 1 + \frac{t}{\sqrt{N\lambda_j}}, \qquad t \neq 0,$$

асимптотически равна $\Phi(t-u_{\alpha/2}) + \Phi(-t-u_{\alpha/2})$.

Сформулированные в предложении 4 критерии позволяют контролировать значения отдельных координат параметрического вектора $\theta = (\theta_1, \ldots, \theta_d)$. Вместе с тем, желательно уметь также проверять гипотезы и о полном этом векторе. Важнейшей такой гипотезой является утверждение

$$H_0: \theta_1 = \theta_2 = \ldots = \theta_d = 1,$$

означающее, что подстановки выбираются равновероятно из S_n.

В случае, если эта гипотеза справедлива, то из (15)–(16) следует, что статистики

$$T_j^*(N) = \frac{T_j - N\lambda_j}{\sqrt{N\lambda_j}}, \qquad j = 1, \ldots, d,$$

асимптотически распределены по стандартному нормальному закону и асимптотически независимы. Следовательно, сумма их квадратов асимптотически распределена по закону $\chi^2(d)$:

$$\mathscr{L}\left(T^2(N) \equiv \sum_{j=1}^{d} \frac{(T_j - N\lambda_j)^2}{N\lambda_j} \right) \to \chi^2(d). \tag{21}$$

Этот результат позволяет сформулировать следующий критерий согласия для гипотезы равновероятности H_0, учитывающий всю исходную информацию.

Предложение 5. *При заданном уровне значимости* α,

$$H_0 \text{ отвергается} \iff \{T^2(N) > \chi^2_{1-\alpha,d}\}.$$

5. Гипотеза однородности

Выше предполагалось, что в процессе наблюдения над подстановками $s_1, \ldots, s_N$ параметр $\theta = (\theta_1, \ldots, \theta_d)$ модели (1)–(4) остается неизменным, что на самом деле является гипотезой, которая сама должна быть подвергнута статистической проверке; такая гипотеза и называется гипотезой однородности. Здесь предлагается асимптотический (для больших выборок) вариант соответствующего критерия однородности.

Итак, теперь считаем, что каждый из векторов (7) получен, вообще говоря, при своем (но неизвестном) значении параметра θ. Обозначим значение этого параметра для подстановки s_k через $\theta^{(k)} = (\theta_1^{(k)}, \ldots, \theta_d^{(k)})$. Тогда гипотеза однородности есть утверждение

$$H_0: \theta^{(1)} = \ldots = \theta^{(N)}.$$

В силу предположения о независимости подстановок и соотношения (6), можно считать, что при гипотезе H_0

$$(\xi_j^{(k)}(n), k = 1, \ldots, N) = (X_{1j}, \ldots, X_{Nj}) \tag{22}$$

представляют собой независимую выборку из распределения Пуассона $\Pi(\theta_j \lambda_j)$ при некотором $\theta_j > 0$; далее для краткости полагаем

$$\alpha_j = \theta_j \lambda_j;$$

при этом, в силу следствия 1 из введения, для разных j выборки (22) асимптотически независимы.

Построим выборочные среднее и дисперсию выборки (22):

$$\bar{X}_j = \frac{1}{N} \sum_{i=1}^{N} X_{ij},$$

$$S_j^2 = \frac{1}{N} \sum_{i=1}^{N} (X_{ij} - \bar{X}_j)^2,$$

и введем статистику

$$T_{Nj} = \frac{1}{\bar{X}_j} \sqrt{\frac{N-1}{2}} \left(\bar{X}_j - \frac{N}{N-1} S_j^2 \right). \tag{23}$$

Как показано в [4], эта статистика при гипотезе H_0 распределена асимптотически по стандартному нормальному закону $\mathcal{N}(0, 1)$, а так как для разных j такие статистики асимптотически независимы, то объединенная статистика

$$T_N^2 = \sum_{j=1}^{d} T_{Nj}^2 \tag{24}$$

при гипотезе H_0 будет асимптотически иметь распределение $\chi^2(d)$.

Тем самым, основываясь на статистике T_N^2, можно предложить следующий критерий согласия для гипотезы H_0.

Предложение 6. *При заданном уровне значимости α, гипотеза однородности H_0 отвергается тогда и только тогда, когда*

$$T_N^2 \geqslant \chi_{1-\alpha,d}^2.$$

Приложение

В теории случайных n-подстановок хорошо известен факт асимптотической, при $n \to \infty$, независимости и асимптотической пуассоновости начальных членов цикловой последовательности $\mathbf{c}(n) = (c_1, c_2, \ldots, c_n)$ случайной равновероятной подстановки, при этом

$$\mathcal{L}(c_i) \to \Pi\left(\frac{1}{i}\right)$$

для любого конечного i.

Аналогичное свойство имеет место и для модели Эванса, причем в этом случае

$$\mathscr{L}(c_i) \to \Pi\left(\frac{\theta}{i}\right).$$

Это свойство асимптотической независимости и пуассоновости чисел циклов конечной длины сохраняется и для общей параметрической модели, введенной в [5], согласно которой произвольная подстановка $s \in S_n$ наблюдается с вероятностью, пропорциональной $\prod_i \theta_i^{c_i}$, где $\theta = (\theta_1, \ldots, \theta_n)$, $\theta_\bullet \geqslant 0$ — параметр меры. В этом случае для производящей функции цикловой структуры $\mathbf{c}(n)$ справедливо представление

$$\mathbf{E}_\theta \prod_{i=1}^{n} t_i^{c_i} = \frac{H_n(t\theta)}{H_n(\theta)}, \tag{25}$$

где

$$H_n(\theta) = n!\,[z^n] \exp\{a(z;\theta)\}$$

и

$$a(z;\theta) = \sum_{i=1}^{n} \theta_i \frac{z^i}{i}.$$

Отсюда следует, что производящая функция для произвольного конечного числа k начальных членов $c_1, c_2, \ldots, c_k$ имеет вид

$$\mathbf{E}_\theta \prod_{i=1}^{k} t_i^{c_i} = \frac{n!}{H_n(\theta)}[z^n] \exp\left\{a(z;\theta) + \sum_{i=1}^{k} \theta_i (t_i - 1)\frac{z^i}{i}\right\}.$$

Далее, используя стандартную технику метода перевала, нетрудно получить, что при $n \to \infty$ и фиксированных $\theta_\bullet$ справедливо соотношение

$$\mathbf{E}_\theta \prod_{i=1}^{k} t_i^{c_i} \to \exp\left\{\sum_{i=1}^{k} \frac{\theta_i}{i}(t_i - 1)\right\},$$

которое означает асимптотическую независимость величин $c_1, \ldots, c_k$ и их пуассоновскую сходимость:

$$\mathscr{L}(c_i) \to \Pi\left(\frac{\theta_i}{i}\right). \tag{26}$$

Утверждение теоремы, сформулированной во введении, является следствием этого результата, поскольку для всех A_j-циклов параметры $\theta_\bullet$ в рассматриваемой модели одинаковы.

Список литературы

1. Сачков В. Н., *Введение в комбинаторные методы дискретной математики*. МЦНМО, Москва, 2004.

2. Соболева М. В., Асимптотическая нормальность чисел конгруэнтных циклов в случайных подстановках. *Дискретная математика* (2012) **24**, №1, 123–131.

3. Ивченко Г. И., Медведев Ю. И., Статистика параметрической модели случайных подстановок. *Труды по дискретной математике* (2004) **8**, 116–127.

4. Ивченко Г. И., Медведев Ю. И., Статистические выводы для случайных подстановок по неполным данным. *Труды по дискретной математике* (2006) **9**, 66–76.

5. Ивченко Г. И., Медведев Ю. И., Случайные подстановки: общая параметрическая модель. *Дискретная математика* (2006) **18**, №4, 105–112.

6. Ивченко Г. И., Медведев Ю. И., *Введение в математическую статистику*. ЛКИ/URSS, Москва, 2010.

7. Ивченко Г. И., Медведев Ю. И., О случайных подстановках. *Труды по дискретной математике* (2002) **5**, 73–92.

Статья поступила 28.08.2012.