Math-Net.Ru

Общероссийский математический портал

Г. И. Ивченко, Оценивание в схеме Маркова–Пойа, *Матем. заметки*, 1998, том 64, выпуск 3, 373–382

DOI: <https://doi.org/10.4213/mzm1407>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением

<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 3.80.181.102

26 апреля 2023 г., 12:50:55

УДК 519

ОЦЕНИВАНИЕ В СХЕМЕ МАРКОВА–ПОЙЯ

Г. И. Ивченко

Рассматривается урновая схема Маркова–Пойя, когда из урны, содержащей заданное число шаров каждого из $N \geq 2$ различных цветов, последовательно и равновероятно извлекаются шары, причем каждый раз извлеченный шар возвращается в урну с добавлением s шаров того же цвета. В предположении, что общее число шаров в урне в начальный момент известно, но ее состав, характеризуемый вектором $\bar{\theta} = (\theta_1, \dots, \theta_N)$ долей цветов, является неизвестным параметром модели, обсуждаются различные подходы к оцениванию $\bar{\theta}$ по информации $\bar{\eta}(n) = (\eta_1(n), \dots, \eta_N(n))$, где η_j есть число шаров j -го цвета, наблюдавшихся при n извлечениях. Обсуждается связь полученных результатов с известными результатами для полиномиального и многомерного гипергеометрического распределений.

Библиография: 8 названий.

1. Введение. Мы рассматриваем целочисленный случайный вектор $\bar{\eta}(n) = (\eta_1(n), \dots, \eta_N(n))$ с распределением

$$P\{\bar{\eta}(n) = \bar{n}\} = \prod_{j=1}^N \binom{\alpha_j + n_j - 1}{n_j} / \binom{\alpha + n - 1}{n}, \quad \bar{n} \in K_{Nn}, \quad (1)$$

где $\alpha_j > 0, j = 1, \dots, N, \alpha = \alpha_1 + \dots + \alpha_N, N \geq 2$ и

$$K_{Nn} = \{\bar{n} = (n_1, \dots, n_N) : n_j = 0, 1, \dots, j = 1, \dots, N, n_1 + \dots + n_N = n\}.$$

Напомним, что для любого вещественного x и целого $k > 0$

$$\binom{x}{k} = \frac{(x)_k}{k!} = \frac{x(x-1) \cdots (x-k+1)}{k!}, \quad \binom{x}{0} = 1.$$

Распределение (1) названо в [1] *распределением Маркова–Пойя* (обозначается символически $\mathcal{L}(\bar{\eta}(n)) = \text{МП}(n; \bar{\alpha}), \bar{\alpha} = (\alpha_1, \dots, \alpha_N)$) ввиду того, что впервые оно встречается (для случая $N = 2$) в работах этих авторов еще в начале нашего столетия. Это одно из наиболее интересных и популярных в литературе по дискретным вероятностным задачам распределений, история и вероятностные свойства которого детально

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, грант № 97-01-00387.

описаны в [1]. Наиболее известно применение МП-распределения к процессу извлечения шаров в урновой схеме со случайно изменяющимся составом урны (урновая схема Маркова–Пойя). Именно, пусть имеется урна, содержащая шары N различных цветов $A_1, \dots, A_N$, причем первоначально число шаров цвета A_j равно a_j , $j = 1, \dots, N$, так что $a = a_1 + \dots + a_N$ есть общее число шаров в урне в начальный момент. В каждый последующий момент из урны наугад извлекается один шар, фиксируется его цвет, и шар возвращается обратно в урну с добавлением $s \geq 1$ новых шаров того же цвета, что и извлеченный шар. Если $\eta_j(n)$ обозначает число A_j -шаров, наблюдавшихся в этом процессе при n извлечениях, то вектор $\bar{\eta}(n) = (\eta_1(n), \dots, \eta_N(n))$ будет иметь распределение (1) при $\alpha_j = a_j/s$, $j = 1, \dots, N$. Формально здесь можно положить также $s = 0$ и $s = -1$, и тогда в первом из этих случаев (1) сводится к хорошо известному полиномиальному (мультиномиальному) распределению $M(n, \bar{p})$ с $\bar{p} = \bar{a}/a$, описывающему схему случайной повторной выборки из конечной совокупности, а во втором – к многомерному гипергеометрическому распределению $МГГР(n; \bar{a})$, описывающему (при $n \leq a$) схему случайной бесповторной выборки.

Далее мы предполагаем, что вектор $\bar{\theta} = \bar{a}/\alpha$ является неизвестным параметром модели $МП(n; \alpha \bar{\theta})$ с параметрическим множеством $\Theta = \{\bar{\theta} = (\theta_1, \dots, \theta_N) : \theta_j > 0, j = 1, \dots, N, \theta_1 + \dots + \theta_N = 1\}$, и рассматриваем проблему его оценивания по наблюдению над вектором $\bar{\eta}(n)$. В терминах урновой схемы это означает, что нам известно общее число a шаров в урне в начальный момент, но не известен ее состав $\bar{a} = (a_1, \dots, a_N)$, и мы хотим оценить вектор $\bar{\theta} = \bar{a}/a$ долей цветов по наблюдению частот шаров соответствующих цветов в n опытах (число s добавляемых шаров при этом также считается известным). Ниже рассматриваются различные подходы к оцениванию параметра $\bar{\theta}$, при этом качество произвольной оценки $\bar{T} = (T_1, \dots, T_N)$ измеряется с помощью функции риска

$$R(\bar{T}, \bar{\theta}) = E|\bar{T} - \bar{\theta}|^2 = \sum_{j=1}^N (T_j - \theta_j)^2. \quad (2)$$

В п. 2 находятся несмещенная оценка $\bar{\theta}^*$ и оценка с постоянным риском $\bar{T}_c$ и проводится их сравнение в терминах меры (2). В п. 3 рассматривается байесовский подход к оцениванию $\bar{\theta}$: определяется класс сопряженных априорных распределений и находится общий вид соответствующих байесовских оценок, которые оказываются линейными относительно наблюдений. В п. 4 при некоторых условиях на параметры модели находится минимаксная оценка, которая оказывается совпадающей с $\bar{T}_c$, и выписывается явный вид наименее благоприятного априорного распределения. Наконец, в п. 5 и 6 демонстрируется, как из полученных результатов следуют соответствующие (известные) результаты для моделей $M(n; \bar{\theta})$ и $МГГР(n; a\bar{\theta})$.

2. Оценки: несмещенная и с постоянным риском. Известно [1], что в модели (1)

$$E\eta_j(n) = n \frac{\alpha_j}{\alpha}, \quad D\eta_j(n) = \frac{n\alpha_j(\alpha - \alpha_j)(\alpha + n)}{\alpha^2(\alpha + 1)}, \quad j = 1, \dots, N. \quad (3)$$

Отсюда следует, что несмещенной оценкой для $\bar{\theta} = \bar{a}/\alpha$ является статистика $\bar{\theta}^* = \bar{\eta}(n)/n$, а ее риск (2) есть

$$R(\bar{\theta}^*, \bar{\theta}) = \sum_{j=1}^N D\theta_j^* = c(1 - |\bar{\theta}|^2), \quad c = c(\alpha, n) = \frac{\alpha + n}{n(\alpha + 1)}. \quad (4)$$

Если ввести вектор $\bar{e} = (1/N, \dots, 1/N) \in \Theta$, то (4) можно записать в виде

$$R(\bar{\theta}^*, \bar{\theta}) = c \left(\frac{N-1}{N} - |\bar{\theta} - \bar{e}|^2 \right), \quad (5)$$

откуда следует, что

$$m(\bar{\theta}^*) \equiv \max_{\bar{\theta}} R(\bar{\theta}^*, \bar{\theta}) = c \frac{N-1}{N}. \quad (6)$$

Рассмотрим теперь класс линейных оценок вида

$$\bar{T}(d) = d\bar{\theta}^* + (1-d)\bar{e}, \quad d \in [0, 1].$$

Для таких оценок функция риска (2) есть

$$R(\bar{T}(d), \bar{\theta}) = E|d(\bar{\theta}^* - \bar{\theta}) - (1-d)(\bar{\theta} - \bar{e})|^2 = d^2 R(\bar{\theta}^*, \bar{\theta}) + (1-d)^2 |\bar{\theta} - \bar{e}|^2,$$

что с учетом (4) дает окончательную формулу

$$R(\bar{T}(d), \bar{\theta}) = ((1-d)^2 - cd^2) |\bar{\theta}|^2 + cd^2 - \frac{(1-d)^2}{N}. \quad (7)$$

Отсюда следует, что при $d = (1 + \sqrt{c})^{-1}$ риск оценки $\bar{T}(d)$ не зависит от параметра $\bar{\theta}$; следовательно, статистика

$$\bar{T}_c = \bar{T}((1 + \sqrt{c})^{-1}) = \frac{1}{1 + \sqrt{c}} \bar{\theta}^* + \frac{c}{1 + \sqrt{c}} \bar{e} \quad (8)$$

является оценкой с постоянным риском, который есть

$$R(\bar{T}_c, \bar{\theta}) = \frac{c}{(1 + \sqrt{c})^2} \cdot \frac{N-1}{N}. \quad (9)$$

Из (6) и (9), в свою очередь, следует, что $R(\bar{T}_c, \bar{\theta}) < m(\bar{\theta}^*)$, т.е. по принципу минимакса оценка $\bar{T}_c$ лучше несмещенной оценки $\bar{\theta}^*$. Зона предпочтения оценки $\bar{T}_c$ перед $\bar{\theta}^*$ имеет вид (см. (5))

$$G_c = \{\bar{\theta} : R(\bar{T}_c, \bar{\theta}) < R(\bar{\theta}^*, \bar{\theta})\} = \left\{ \bar{\theta} : |\bar{\theta} - \bar{e}|^2 < \frac{N-1}{N} \varphi(c) \right\}, \quad \varphi(c) = \frac{\sqrt{c}(2 + \sqrt{c})}{(1 + \sqrt{c})^2}.$$

Функция $\varphi(c)$ монотонно возрастает, поэтому область G_c расширяется с увеличением c . Но (см. (4))

$$c = \frac{1}{n} + \frac{n-1}{n} \cdot \frac{1}{\alpha+1} = \frac{1}{\alpha+1} + \frac{\alpha}{\alpha+1} \cdot \frac{1}{n},$$

т.е. c убывает как при увеличении n , так и при увеличении α , и если оба эти параметра одновременно велики, то c мало, и в этом случае область G_c представляет собой малую окрестность (радиуса порядка $\sqrt{c}$) точки $\bar{e}$.

Рассмотрим теперь функцию $\psi(c) = c(1 + \sqrt{c})^{-2}$ в (9). Эта функция также монотонно возрастает. Поэтому риск оценки $\bar{T}_c$ убывает с увеличением как n , так и α . В частности,

если n фиксировано, а $\alpha \rightarrow \infty$, то $c \downarrow n^{-1}$, $\psi(c) \downarrow (1 + \sqrt{n})^{-2}$, и “предельная” оценка принимает вид

$$\bar{T}_{n-1} = \frac{\sqrt{n}}{1 + \sqrt{n}} \bar{\theta}^* + \frac{1}{1 + \sqrt{n}} \bar{e} = \left(\frac{n_j(n) + \sqrt{n}/N}{n + \sqrt{n}}, j = 1, \dots, N \right). \quad (10)$$

Но это есть хорошо известная минимаксная оценка параметра $\bar{\theta}$ полиномиального распределения $M(n; \bar{\theta})$ (см. [2]). Это является отражением того факта, что распределение МП($n; \alpha \bar{\theta}$) при $\alpha \rightarrow \infty$ сходится к $M(n; \bar{\theta})$ (непосредственно следует из (1)) и, таким образом, задача оценивания $\bar{\theta}$ сводится в пределе к аналогичной задаче для полиномиального распределения, решение которой хорошо известно. В свете этого можно было бы предположить, что и в допредельной ситуации (при фиксированном α) оценка с постоянным риском $\bar{T}_c$ также обладает свойством минимаксности. Однако, доказать это удастся лишь при некотором дополнительном условии на параметры n и α (см. п. 4). Эта “экзотическая” особенность модели МП($n; \alpha \bar{\theta}$) для случая $N = 2$ была обнаружена ранее в [3].

3. Байесовские оценки. В соответствии с байесовским принципом неизвестный параметр модели рассматривается как случайная величина с заданным априорным распределением $\mathcal{L}(\bar{\theta}) = G(\bar{\theta})$ на параметрическом множестве (для упрощения записи формул мы обозначаем одним и тем же символом как случайную величину, так и принимаемое ею значение), и при таком подходе ищется байесовская оценка $\bar{T}_G = \arg \min r(\bar{T})$, т.е. оценка, минимизирующая байесовский риск

$$r(\bar{T}) = \int_{\Theta} R(\bar{T}, \bar{\theta}) dG(\bar{\theta});$$

при этом, как известно, в случае квадратичной функции потерь $|\bar{T} - \bar{\theta}|^2$ оценка $\bar{T}_G$ совпадает с апостериорным средним $E(\bar{\theta} | \bar{\eta}(n))$.

Важную роль в байесовской теории играют сопряженные априорные распределения, для которых апостериорные распределения $\mathcal{L}(\bar{\theta} | \bar{\eta}(n) = \bar{n})$ принадлежат тому же семейству, что и $\mathcal{L}(\bar{\theta})$. Вид такого семейства для рассматриваемой модели МП($n; \alpha \bar{\theta}$) дается в следующей лемме.

ЛЕММА. Для модели $\mathcal{L}(\bar{\eta}(n)) = \text{МП}(n, \alpha \bar{\theta})$ сопряженное семейство задается соотношением $\mathcal{L}(\alpha \bar{\theta} - \bar{\gamma}) = \text{МП}(m, \bar{\gamma})$, где $\bar{\gamma} = (\gamma_1, \dots, \gamma_N)$ — любой положительный вектор такой, что $m = \alpha - \gamma$ есть целое положительное число, $\gamma = \gamma_1 + \dots + \gamma_N$. При этом апостериорное распределение имеет вид

$$\mathcal{L}(\alpha \bar{\theta} - \bar{\gamma} | \bar{\eta}(n) = \bar{n}) = \text{МП}(m, \bar{\gamma} + \bar{n}). \quad (11)$$

Доказательство следует непосредственно из определения (1) МП-модели. Действительно, если $\bar{k} \in K_{Nm}$, то апостериорная вероятность $P\{\alpha \bar{\theta} - \bar{\gamma} = \bar{k} | \bar{\eta}(n) = \bar{n}\}$ пропорциональна произведению

$$P\{\alpha \bar{\theta} - \bar{\gamma} = \bar{k}\} P\{\bar{\eta}(n) = \bar{n} | \bar{\theta} = \alpha^{-1}(\bar{\gamma} + \bar{k})\},$$

что, в свою очередь, пропорционально

$$\prod_{j=1}^N \binom{\gamma_j + k_j - 1}{k_j} \binom{\gamma_j + k_j + n_j - 1}{n_j} \quad \text{или} \quad \prod_{j=1}^N \binom{\gamma_j + k_j + n_j - 1}{k_j}.$$

Но это и означает, что апостериорное распределение имеет вид (11).

Таким образом, в рассматриваемой проблеме сопряженными являются распределения также МП-типа. Но это дискретные распределения с конечным носителем, т.е. каждое такое распределение приписывает ненулевые вероятности лишь конечному числу точек параметрического множества Θ . В связи с этим возникает вопрос, насколько целесообразно использование таких распределений. По-видимому, для общего случая, когда в (1) вектор $\bar{\alpha}$ никак не специализируется, такие априорные распределения вряд ли подходящи (к тому же при $\alpha \leq 1$ условие леммы на выбор вектора $\bar{\gamma}$ не может быть выполненным вообще). Но если речь идет об урновой МП-схеме, то параметр $\bar{\theta} = \bar{\alpha}/a$, т.е. его допустимые значения ограничены (при известном a) лишь конечным числом точек с рациональными координатами, и потому при $\alpha = a/s > 1$ указанные в лемме априорные распределения представляются вполне уместными. Основное же достоинство сопряженных распределений заключается, как известно, в удобстве вычисления байесовских оценок, в чем мы убеждаемся ниже.

В нашем случае (см. (3)) из (11) следует, что

$$E(\alpha \bar{\theta} - \bar{\gamma} | \bar{\eta}(n) = \bar{n}) = m \frac{\bar{\gamma} + \bar{n}}{\gamma + n},$$

и потому байесовская оценка, соответствующая указанному в лемме априорному распределению есть

$$\bar{T}(\bar{\gamma}) = E(\bar{\theta} | \bar{\eta}(n)) = \alpha^{-1} \left(\frac{\alpha + n}{\gamma + n} \bar{\gamma} + \frac{m}{\gamma + n} \bar{\eta}(n) \right). \quad (12)$$

Отметим, что это линейные относительно наблюдений $\bar{\eta}(n)$ оценки, которые можно записать также в виде

$$\bar{T}(\bar{\gamma}) = w \bar{\theta}^* + (1 - w) \bar{\gamma}', \quad \bar{\gamma}' = \frac{\bar{\gamma}}{\gamma} \in \Theta, \quad w = \frac{(\alpha - \gamma)n}{\alpha(\gamma + n)} \in (0, 1). \quad (13)$$

Таким образом, $\bar{T}(\bar{\gamma})$ лежит на отрезке, соединяющем точки $\bar{\theta}^*$ и $\bar{\gamma}'$. Дополним сказанное еще указанием функции риска оценки $\bar{T}(\bar{\gamma})$:

$$R(\bar{T}(\bar{\gamma}), \bar{\theta}) = cw^2(1 - |\bar{\theta}|^2) + (1 - w)^2 |\bar{\gamma}' - \bar{\theta}|^2, \quad (14)$$

а также ее байесовского риска:

$$r(\bar{T}(\bar{\gamma})) = [cw^2\gamma(\alpha + 1) + (1 - w)^2(\alpha - \gamma)] \frac{1 - |\bar{\gamma}'|^2}{\alpha(\gamma + 1)} = \frac{(\alpha + n)(\alpha - \gamma)\gamma}{\alpha^2(\gamma + n)(\gamma + 1)} (1 - |\bar{\gamma}'|^2) \quad (15)$$

(рутинные вычисления, основанные на формулах (3), мы опускаем).

Выделим интересный частный случай, соответствующий урновой МП-схеме с числом шаров $a > N$ и параметром $s = 1$. Тогда $\bar{\alpha} = \bar{a}$, $\alpha = a$, $\bar{\theta} = \bar{a}/a$ и при выборе в лемме $\bar{\gamma} = N\bar{e} = (1, \dots, 1)$ сопряженное априорное распределение принимает вид $\mathcal{L}(a\bar{\theta} - N\bar{e}) = \text{МП}(a - N, N\bar{e})$. Но (см. (1)) это есть равномерное распределение на множестве $K_{N, a-N}$, приписывающее одинаковые вероятности $\binom{a-1}{N-1}^{-1}$ всем допустимым значениям вектора состава $\bar{a}$ с положительными координатами. Для этого случая байесовская оценка (13) принимает вид

$$\bar{T}(N\bar{e}) = w_0\bar{\theta}^* + (1 - w_0)\bar{e}, \quad w_0 = \frac{n(a - N)}{a(n + N)}.$$

Ее функция риска (14) есть

$$((1 - w_0)^2 - cw_0^2)|\bar{\theta}|^2 + cw_0^2 - \frac{(1 - w_0)^2}{N},$$

а байесовский риск (15) равен

$$\frac{(a + n)(a - N)(N - 1)}{a^2(N + n)(N + 1)}.$$

4. Минимаксное оценивание. Если отсутствует информация об априорном распределении $\mathcal{L}(\theta)$ или байесовский подход вообще не применим, то в таких случаях для сравнения “качества” различных оценок (при выбранной функции потерь) можно использовать максимальный риск $m(\bar{T}) = \sup_{\bar{\theta}} R(\bar{T}, \bar{\theta})$. Если существует оценка $\bar{T}^* = \arg \min m(\bar{T})$, т.е. минимизирующая максимальный риск, то ее называют минимаксной оценкой. Такая оценка гарантирует от чрезмерных потерь: наихудший ожидаемый ущерб, связанный с использованием этой оценки, настолько мал, насколько это в принципе возможно. В ряде случаев такое свойство оценки может быть предпочтительным и потому минимаксная оценка может быть практически важной. Для отыскания минимаксной оценки часто используется следующий прием: если удастся найти такое априорное распределение параметра $G^*(\bar{\theta})$, для которого функция риска соответствующей байесовской оценки $\bar{T}_{G^*}$ постоянна: $R(\bar{T}_{G^*}, \bar{\theta}) \equiv \text{const}$ (такое распределение G^* называется *наименее благоприятным априорным распределением*), то в этом случае $\bar{T}^* = \bar{T}_{G^*}$; при этом G^* часто находится в сопряженном семействе. Применим эти общие принципы в нашей задаче.

В п. 2 была построена оценка с постоянным риском $\bar{T}_c$ (см. (8)), которая находится на отрезке, соединяющем точки $\bar{\theta}^*$ и $\bar{e}$. Но на этом же отрезке лежат и все байесовские оценки $\bar{T}(\gamma\bar{e})$ (см. (13)), где γ , $0 < \gamma < \alpha$, такое, что $\alpha - \gamma$ есть целое положительное число. Следовательно, байесовскую оценку с постоянным риском можно получить, выбирая γ так, чтобы $\bar{T}(\gamma\bar{e})$ совпало с $\bar{T}_c$ (если это, конечно, возможно). Это приводит к уравнению

$$\frac{(\alpha - \gamma)n}{\alpha(\gamma + n)} = \frac{1}{1 + \sqrt{c}},$$

которому удовлетворяет единственное значение

$$\gamma^* = \frac{\alpha n \sqrt{c}}{\alpha + n(1 + \sqrt{c})} = \frac{\alpha}{1 + \sqrt{(1 + \alpha)(1 + \alpha/n)}}. \quad (16)$$

Таким образом, если $m = \alpha - \gamma^*$ является целым положительным числом, то $\bar{T}(\gamma^* \bar{e}) = \bar{T}_c$ и, стало быть, оценка $\bar{T}_c$ минимаксная. Этому значению γ^* соответствует сопряженное распределение $G^*(\bar{\theta})$, задаваемое соотношением $\mathcal{L}(\alpha \bar{\theta} - \gamma^* \bar{e}) = \text{МП}(m; \gamma^* \bar{e})$, т.е.

$$P\{\bar{\theta} = \alpha^{-1}(\gamma^* \bar{e} + \bar{k})\} = \prod_{j=1}^N \binom{\gamma^*/N + k_j - 1}{k_j} / \binom{\gamma^* + m - 1}{m}, \quad \bar{k} \in K_{Nm}. \quad (17)$$

Подведем итог в виде следующего утверждения.

ТЕОРЕМА. Если параметры α и n модели $\text{МП}(n, \alpha \bar{\theta})$ такие, что $m = \alpha - \gamma^*$, где γ^* дано в (16), есть целое положительное число, то оценка $\bar{T}_c$, данная в (8), является минимаксной. Она является также байесовской оценкой для $\bar{\theta}$ при априорном распределении G^* , данным в (17).

Подчеркнем также, что условие теоремы не зависит от размерности N модели $\text{МП}(n, \alpha \bar{\theta})$ и совпадает с соответствующим условием, полученным в [3] для случая $N = 2$.

Добавим к этому еще следующее. Если α есть целое число, то для применимости теоремы γ^* также должно быть целым. В свою очередь, это будет так лишь тогда, когда α и n имеют представления

$$\alpha = k \left(k + 2 + \frac{k(k+1)^2}{l} \right), \quad n = k^2 + l, \quad (18)$$

где k – произвольное натуральное число, а l – любой делитель числа $k(k+1)^2$ (это непосредственно следует из представления (16)). Если имеет место представление (18), то $\gamma^* = k$, а

$$m = \alpha - \gamma^* = k(k+1) \left(1 + \frac{k(k+1)}{l} \right).$$

Например, для $k = 1$ условию (18) удовлетворяют следующие пары значений (α, n) : (7, 2), (5, 3), (4, 5), а для $k = 2$ – пары (44, 5), (26, 6), (20, 7), (14, 10), (12, 13), (10, 22).

5. Полиномиальное распределение. Из изложенных результатов предельным переходом при $\alpha \rightarrow \infty$ можно получить соответствующие результаты для полиномиальной модели $M(n; \bar{\theta})$, которая, как отмечено в п. 2, является предельной для модели $\text{МП}(n, \alpha \bar{\theta})$. Действительно, как следует из [1, теорема 2], для указанных в лемме сопряженных распределений при $\alpha \rightarrow \infty$ выполняется предельное соотношение

$$\mathcal{L} \left(\frac{\alpha \bar{\theta} - \bar{\gamma}}{\alpha} \right) \rightarrow \mathcal{L}(\bar{\theta}) = D(\bar{\gamma}),$$

где $D(\bar{\gamma})$ есть распределение Дирихле, являющееся, как известно, сопряженным для $M(n; \bar{\theta})$.

Далее, байесовская оценка (12) сводится в пределе к виду $(\bar{\eta}(n) + \bar{\gamma}) / (n + \gamma)$ – байесовской оценке для $\bar{\theta}$ в модели $M(n; \bar{\theta})$, соответствующей априорному распределению $D(\bar{\gamma})$ (при этом снимается ограничение на выбор параметра γ).

Наконец, поскольку (см. (16)) $\gamma^* \rightarrow \sqrt{n}$, минимаксная оценка $\bar{T}(\gamma \bar{e})$ и наименее благоприятное априорное распределение (17) сводятся в пределе соответственно к оценке (10) и распределению $D(\sqrt{n} \bar{e})$ – также хорошо известным результатам для модели $M(n; \bar{\theta})$ (см. [2]).

6. Многомерное гипергеометрическое распределение. В этом пункте мы кратко обсудим задачу оценивания параметра $\bar{\theta} = (\theta_1, \dots, \theta_N)$ в модели МГГР($n; a\bar{\theta}$), $\bar{\theta} = \bar{a}/a$, которая, как отмечено в п. 1, соответствует случаю выбора без возвращения из урны состава $\bar{a} = (a_1, \dots, a_N)$. Эта модель является двойственной к МП-модели (получается из нее подстановкой $\alpha_j \mapsto -a_j$, $j = 1, \dots, N$) и, как подчеркивается (и демонстрируется на ряде примеров) в [1], многие результаты для МГГР могут быть получены по принципу двойственности формальной переформулировкой соответствующих результатов для МП-модели. Именно под таким углом зрения мы и рассмотрим примыкающие по теме статистические вопросы для МГГР, хотя задача оценивания неизвестного состава $\bar{a}$ в схеме бесповторной выборки ранее уже рассматривалась в литературе [4]–[7]. Далее предполагается, что $n < a$ (случай $n = a$ тривиален).

Прежде всего отметим, что несмещенная оценка $\bar{\theta}^* = \bar{\eta}(n)/n$ и оценка с постоянным риском $\bar{T}_c$, данная в (8), являются таковыми же и в модели МГГР($n; a\bar{\theta}$) с учетом того, что теперь величина $c = c(-a, n) = (a - n)/(n(a - 1))$. Отличие состоит лишь в том, что теперь риски этих оценок (см. (4) и (9)) возрастают с ростом a , и потому при $a \rightarrow \infty$ оценка $\bar{T}_c$ сходится к оценке $\bar{T}_{n-1}$, данной в (10), так сказать, “снизу”.

Также сохраняют свой вид и байесовские оценки $\bar{T}(\bar{\gamma})$, данные в (13), но теперь с $w = (n(a + \gamma))/(a(n + \gamma))$. Однако модель МГГР($n; a\bar{\theta}$) в этом аспекте имеет и свою специфику: для нее семейство сопряженных распределений имеет вид $\mathcal{L}(\bar{a}) = \text{МП}(a; \bar{\gamma})$ (без ограничений на параметр $\gamma = \gamma_1 + \dots + \gamma_N$), при этом для апостериорных распределений выполняется соотношение

$$\mathcal{L}(\bar{a} - \bar{n} | \bar{\eta} = \bar{n}) = \text{МП}(a - n; \bar{\gamma} + \bar{n}).$$

Это проверяется непосредственно стандартными рассуждениями и является новым результатом.

Что же касается минимаксной оценки $\bar{T}^*$, то она также совпадает с оценкой постоянного риска: $\bar{T}^* = \bar{T}_c$ и имеет теперь вид

$$\bar{T}^* = \left(\bar{\theta}^* + \sqrt{\frac{a - n}{n(a - 1)}} \bar{e} \right) / \left(1 + \sqrt{\frac{a - n}{n(a - 1)}} \right).$$

Она же является байесовской оценкой $\bar{T}(\gamma^* \bar{e})$ с γ^* , данным в (16) (но уже без ограничений на его арифметические свойства), т.е. теперь при любых a и $n < a$

$$\gamma^* = a \sqrt{\frac{n(a - n)}{a - 1}} / \left(a - n - \sqrt{\frac{n(a - n)}{a - 1}} \right),$$

а наименее благоприятное априорное распределение есть $\mathcal{L}(\bar{a}) = \text{МП}(a; \gamma^* \bar{e})$. Здесь следует заметить, что при $n = a - 1$ величина $\gamma^* = \infty$, и в этом случае под $\text{МП}(a; \gamma^* \bar{e})$ следует понимать предел для $\text{МП}(a; \gamma \bar{e})$ при $\gamma \rightarrow \infty$, который, как отмечалось выше, есть полиномиальное распределение $M(a; \bar{e})$ (вид же оценки $\bar{T}^*$ от γ^* не зависит).

Наконец, отметим также следующий интересный факт. Поскольку риск оценки $\bar{T}_{n-1}$, равный (см. (7))

$$cd^2(1 - |\bar{\theta}|^2) + (1 - d^2) \left(|\bar{\theta}|^2 - \frac{1}{N} \right), \quad d = \frac{\sqrt{n}}{1 + \sqrt{n}},$$

очевидно, возрастает с ростом a до предельного значения

$$r = \frac{N-1}{N}(1 + \sqrt{n})^{-2},$$

а байесовский риск оценки $\bar{T}^*$ (см. (9)) удовлетворяет соотношению

$$r(\bar{T}^*) = \frac{c}{(1 + \sqrt{c})^2} \cdot \frac{N-1}{N} \rightarrow r, \quad a \rightarrow \infty,$$

отсюда, как известно (см., например, [8, с. 232]), следует, что статистика $\bar{T}_{n-1}$ является минимаксной оценкой параметра $\bar{\theta}$ в случае, когда a – общее число шаров в урне в начальный момент – также не известно. Этот результат (для более общей функции потерь) ранее установлен в [6].

7. Заключение. Вопросы оценивания, включая байесовский и минимаксный подходы, для таких популярных дискретных распределений, как полиномиальное $M(n; \bar{p})$ и гипергеометрическое МГГР($n; \bar{a}$), неоднократно обсуждались в статистической литературе. В то же время, как подчеркивается в [1], эти распределения укладываются в более общую схему распределения Маркова–Пойа МП($n; \bar{\alpha}$) при соответствующей трактовке его параметров $\bar{\alpha} = (\alpha_1, \dots, \alpha_N)$. Нашей целью было рассмотреть некоторые статистические вопросы для всех этих распределений одновременно, в рамках общего подхода. Это представляет, помимо прочего, определенный методический интерес, так как при этом выясняются как общие факторы, так и специфические особенности различных конкретных моделей.

Для упрощения технической стороны задачи мы ограничились при этом простейшим (но практически важным) вариантом функции потерь $L(\bar{T}, \bar{\theta}) = |\bar{T} - \bar{\theta}|^2$ хотя, конечно, вся идеология может быть проведена и в случае общей функции потерь вида $L(\bar{T}, \bar{\theta}) = (\bar{T} - \bar{\theta})' L(\bar{T} - \bar{\theta})$ с произвольной неотрицательно определенной матрицей L (в духе работ [5]–[7]).

Если отвлечься от конкретных деталей, которые оказались сами по себе достаточно интересными, то итог можно изобразить следующей условной диаграммой на языке шкалы параметра α с выколотым нулем (см. рис. 1).

Рис. 1

На рис. 1 зона всех положительных значений α соответствует собственно МП-модели, множество всех целых отрицательных значений α – модели МГГР, а бесконечно удаленные (в обоих направлениях) точки соответствуют мультиномиальной модели. При этом ряду результатов, имеющих место в рамках одной из этих моделей, соответствуют их “естественные” аналоги для двух других. Но это соответствие не всегда и не во всем буквальное и в ряде аспектов окрашено специфическими особенностями этих моделей.

Таким образом, можно сказать, что эти три модели имеют как общие родовые черты, позволяющие угадывать направление поиска аналогов результатов, полученных в одной из них, в других, так и индивидуальность, придающую этому процессу характер творческой, но отнюдь не “механической” работы.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

- [1] Ивченко Г. И. Об урновой схеме Маркова–Пойа: от 1917 г. до наших дней // Обзорение прикладной и промышл. матем. Сер. дискретн. матем. 1996. Т. 3. № 4. С. 484–511.
- [2] Steinhaus H. The problem of estimation // Ann. Math. Statist. 1957. V. 28. P. 633–648.
- [3] Styszynski A. On minimax estimation of the parameter in the Polya urn scheme // Zastos. Mat. 1984. V. 18. № 2. P. 225–229.
- [4] Trybula S. Some problems of simultaneous minimax estimation // Ann. Math. Statist. 1958. V. 29. P. 245–253.
- [5] Trybula S. Minimax estimation of the parameters of the multivariate hypergeometric distribution // Zastos. Mat. 1985. V. 18. № 4. P. 559–570.
- [6] Trybula S., Wilczynski M. Estimation under sampling from a finite population // Zastos. Mat. 1985. V. 18. № 4. P. 571–576.
- [7] Wilczynski M. Minimax estimation for the multinomial and multivariate hypergeometric distributions // Sankhyā Ser. A. 1985. V. 47. № 1. P. 128–132.
- [8] Леман Э. Теория точечного оценивания. М.: Наука, 1991.

Московский государственный институт электроники и математики

Поступило
25.06.97