Math-Net.Ru

Общероссийский математический портал

Г. И. Ивченко, Ю. И. Медведев, Статистические задачи, связанные с организацией контроля за процессами генерации дискретных случайных последовательностей, *Дискрет. матем.*, 2000, том 12, выпуск 2, 3–24

DOI: <https://doi.org/10.4213/dm330>

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: 54.210.77.194

26 апреля 2023 г., 12:24:57

УДК 519.2

Статистические задачи, связанные с организацией контроля за процессами генерации дискретных случайных последовательностей

© 2000 г. Г. И. Ивченко, Ю. И. Медведев

Работа посвящена анализу основных статистических задач оценивания и проверки гипотез в рамках общей модели Маркова–Пойа, вероятностный анализ которой проведен авторами ранее. Статистический анализ схемы Маркова–Пойа представляет интерес в связи с исследованием возможных разладок в работе генераторов случайных последовательностей, а также в связи с вопросами безопасности локальных компьютерных сетей.

1. Введение

В силу различного рода технических, организационных и других причин статистические закономерности процессов генерации дискретных случайных последовательностей могут в некоторые моменты времени значительно отклоняться от предъявляемых требований (появление разладок). При этом в дальнейшем разладки могут устраняться самопроизвольно или под действием автоматического управляющего контроля. Моменты возникновения и устранения разладок, в свою очередь, носят случайный характер.

Система статистического контроля должна организовываться таким образом, чтобы была возможность быстрого выявления значительных отклонений процесса генерации от нормативных требований. Необходима также процедура быстрого выявления существующих длительное время разладок, хотя и приводящих к относительно небольшим отклонениям от нормы.

В связи со сказанным процедура статистического контроля предусматривает вычисление и анализ ряда статистик-функционалов на отрезках последовательностей.

Имеющиеся до последнего времени данные, основанные на обобщении опыта эксплуатации генераторов дискретных случайных последовательностей, позволяют делать вывод о том, что в качестве соответствующих теоретико-вероятностных схем могут быть взяты полиномиальная схема или цепь Маркова. В целом использование этих схем сохраняет актуальность и по сей день.

Однако, к настоящему времени некоторые особенности построения генераторов и процедуры выработки последовательностей в целом создают необходимость рас-

ширения класса теоретико-вероятностных схем, моделирующих процессы генерации при наличии разладок. В число такого рода схем, в частности, включена схема Маркова–Пойа, дающая возможность более детально моделировать процессы, происходящие в генераторах при возникновении в них разладок.

Статистический анализ схемы Маркова–Пойа представляет интерес не только в связи с исследованием возможных разладок в работе генераторов случайных последовательностей, но и в связи с вопросами безопасности в локальных компьютерных сетях.

Известно, что схема Маркова–Пойа моделирует процессы вирусного заражения. Не исключено, что эта схема будет моделировать и какие-либо типы процессов машинного вирусного заражения.

Настоящая работа посвящена систематическому исследованию основных статистических задач оценивания и проверки гипотез в рамках общей модели Маркова–Пойа, вероятностный анализ которой проведен в предыдущей работе авторов [1].

2. Оценивание неизвестных параметров модели Маркова–Пойа

Данный раздел посвящен анализу несмещенных и минимаксных оценок неизвестных параметров схемы Маркова–Пойа. Рассматриваются различные ситуации, включая большие выборки, схемы серий и многовыборочные схемы.

2.1. Несмещенное оценивание в модели Маркова–Пойа

2.1.1. Напомним, что многомерным распределением Маркова–Пойа (далее, кратко, МП-распределение или МП-модель) называется распределение целочисленного случайного вектора $\bar{\eta}(n) = (\eta_1(n), \dots, \eta_N(n))$ на множестве

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

задаваемое вероятностями

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

Здесь $\bar{\alpha} = (\alpha_1, \dots, \alpha_N)$, $\alpha_j > 0$, $j = 1, \dots, N$, — произвольные вещественные числа, являющиеся параметрами модели, и $\alpha = \alpha_1 + \dots + \alpha_N$. В дальнейшем это распределение кратко обозначается символом МП($n, \bar{\alpha}$).

Аналитические свойства этой модели как в точной, так и в асимптотической постановках (при $n \rightarrow \infty$) детально рассмотрены в [1]. В частности, в [1] отмечается, что распределение (1) описывает процесс извлечения шаров в урновой схеме Маркова–Пойа. Именно, предположим, что имеется урна, содержащая шары N различных цветов $A_1, \dots, A_N$, причем первоначально число A_j -шаров равно a_j , $j = 1, \dots, N$, так что общее число шаров в урне в стартовый момент есть $a = a_1 + \dots + a_N$. Далее начинается процесс последовательного и равновероятного извлечения из урны по одному шару, причем каждый раз извлеченный шар возвращается обратно в урну с одновременным добавлением заданного числа c новых шаров

того же цвета. Если $\eta_j(n)$ обозначает число A_j -шаров после n опытов, то вектор частот $\bar{\eta}(n) = (\eta_1(n), \dots, \eta_N(n))$ будет иметь распределение МП($n, \bar{\alpha}$) (1) с $\alpha_j = a_j/c$, $j = 1, \dots, N$.

В приложениях этой модели, которая включает в себя как частные случаи (при определенной интерпретации, см. [1]) такие известные модели, как полиномиальная и многомерная гипергеометрическая, параметры $\bar{\alpha}$ (или часть из них) могут быть неизвестны, и тогда возникает необходимость их статистического оценивания по наблюдаемым данным, например, по наблюдениям над вектором $\bar{\eta}(n)$. Так, нам может быть известным общее число шаров a в урне в начальный момент, но неизвестно их распределение по цветам $A_1, \dots, A_N$, т.е. неизвестен вектор $\bar{a} = (a_1, \dots, a_N)$. Для статистического оценивания этого вектора или, что в данном случае эквивалентно, вектора долей $\bar{\theta} = (\theta_1, \dots, \theta_N) = \bar{a}/a$, организуется статистический эксперимент, состоящий в последовательном извлечении шаров из урны по описанной выше схеме, где c является управляемым параметром, итогом которого является наблюдаемая реализация вектора $\bar{\eta}(n)$. В этом случае задача состоит в построении статистических оценок для неизвестного параметра $\bar{\theta}$ по наблюдениям над $\bar{\eta}(n)$ и исследовании их свойств с целью отыскания наилучших, в том или ином смысле, оценок, в частности, исследовании влияния управляемого параметра c на свойства этих оценок. Такого рода задачи и будут предметом нашего изучения в настоящем разделе.

Итак, далее мы рассматриваем статистическую МП-модель, т.е. модель, в которой параметр $\bar{\theta} = (\theta_1, \dots, \theta_N) = \bar{\alpha}/\alpha$ неизвестен и является некоторой точкой параметрического множества

$$\Theta = \{\bar{\theta}: \theta_j > 0, j = 1, \dots, N, \theta_1 + \dots + \theta_N = 1\},$$

величина α далее считается известной. В качестве оценок $\bar{\theta}$ мы будем использовать те или иные классы статистик $\bar{T} = \bar{T}(\bar{\eta}(n))$, измеряя их качество с помощью функции риска

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

2.1.2. Рассмотрим сначала несмещенные оценки, т.е. оценки, удовлетворяющие условию

$$\mathbf{E}\bar{T}(\bar{\eta}(n)) = \bar{\theta}, \quad \bar{\theta} \in \Theta. \quad (3)$$

Как показано в [1], для МП-распределения (1) имеют место следующие формулы для моментов:

$$\begin{aligned} \mathbf{E}\eta_j(n) &= \frac{n\alpha_j}{\alpha}, & j = 1, \dots, N, \\ \mathbf{D}\eta_j(n) &= \frac{n\alpha_j(\alpha - \alpha_j)(\alpha + n)}{\alpha^2(\alpha + 1)}, & (4) \\ \text{cov}(\eta_i(n), \eta_j(n)) &= -\alpha_i\alpha_j \frac{n(\alpha + n)}{\alpha^2(\alpha + 1)}, & i \neq j. \end{aligned}$$

Отсюда следует, что $\mathbf{E}\eta_j(n)/n = \alpha_j/\alpha = \theta_j$, $j = 1, \dots, N$, т.е. статистика $\bar{\theta}^* = \bar{\eta}(n)/n$ удовлетворяет условию (3) и потому является несмещенной оценкой для $\bar{\theta}$.

При этом, в силу (4)

$$\mathbf{D}\eta_j^* = \mathbf{D}\frac{\eta_j(n)}{n} = e\theta_j(1 - \theta_j), \quad e = e(\alpha, n) = \frac{\alpha + n}{n(\alpha + 1)},$$

и поэтому в (2)

$$R(\bar{\theta}^*, \bar{\theta}) = \sum_{j=1}^N \mathbf{D}\theta_j^* = e(1 - |\bar{\theta}|^2). \quad (5)$$

Введем, далее, вектор $\bar{\theta}^0 = (1/N, \dots, 1/N)$. Тогда, так как

$$|\bar{\theta}|^2 = |\bar{\theta} - \bar{\theta}^0|^2 + 1/N,$$

для (5) можно записать также представление

$$R(\bar{\theta}^*, \bar{\theta}) = e((N - 1)/N - |\bar{\theta} - \bar{\theta}^0|^2), \quad (6)$$

из которого, очевидно, следует, что

$$\max_{\bar{\theta} \in \Theta} R(\bar{\theta}^*, \bar{\theta}) = R(\bar{\theta}^*, \bar{\theta}^0) = e(N - 1)/N. \quad (7)$$

Суммируя эти рассуждения, мы получаем следующее утверждение.

Теорема 1. Несмещенной оценкой параметра $\bar{\theta}$ в МП-модели $\text{МП}(n, \alpha\bar{\theta})$, $\bar{\theta} \in \bar{\Theta}$, является статистика $\bar{\theta}^* = \bar{\eta}(n)/n$. Риск этой оценки дается формулой (5) и его максимальное значение указано в (7).

Подчеркнем следующее обстоятельство. Величина $e = e(\alpha, n)$ в (5) не стремится к нулю при $n \rightarrow \infty$, ее предел равен

$$\lim_{n \rightarrow \infty} e(\alpha, n) = 1/(\alpha + 1),$$

поэтому оценка $\bar{\theta}^*$ не является состоятельной: $\bar{\theta}^*$ не сходится к $\bar{\theta}$ по вероятности с ростом объема выборки. В то же время, эта оценка является наилучшей (по критерию (2)) в классе всех несмещенных оценок (это следует из полноты статистики $\bar{\eta}(n)$, см. раздел 2.2), и тем самым в МП-модели не существует состоятельных несмещенных оценок параметра $\bar{\theta}$ (по одному наблюдению над $\bar{\eta}(n)$). Отсюда, в частности, следует, что необходимо расширить класс используемых оценок, включив в рассмотрение и смещенные оценки, с целью поиска более точных (по критерию (2)), нежели $\bar{\theta}^*$, оценок. Решение этой задачи можно искать на пути построения минимаксной оценки, чему посвящен следующий раздел.

Замечание 1. Отсутствие свойства состоятельности у оценки $\bar{\theta}^* = \bar{\eta}(n)/n$ является отражением того специфического свойства МП-модели, что в ней (см. [1]) вектор относительных частот $\bar{\eta}(n)/n$ сходится с вероятностью 1 при $n \rightarrow \infty$ к случайному вектору $\bar{p} = (p_1, \dots, p_N)$ (а не к $\mathbf{E}\bar{\eta}(n)/n = \bar{\alpha}/\alpha$), имеющему распределение Дирихле $\mathcal{D}(\bar{\alpha})$, задаваемое плотностью

$$g(\bar{p}, \bar{\alpha}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_N)} \prod_{j=1}^N p_j^{\alpha_j - 1}, \quad \bar{p} \in \mathcal{P}_N = \Theta. \quad (8)$$

2.2. Оценка с постоянным риском

Рассмотрим класс оценок вида

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

При $a = 1$ получаем несмещенную оценку $\bar{\theta}^*$, при $a \neq 1$ оценка $\bar{T}$ будет, следовательно, смещенной. Таким образом, мы допускаем и смещенные оценки, имея цель отыскать среди них оценки с меньшим риском.

Вычислим риск (2) произвольной оценки $\bar{T}(a, \bar{b})$. Согласно (5)

$$\begin{aligned} R(\bar{T}(a, \bar{b}), \bar{\theta}) &= \mathbf{E}|a(\bar{\theta}^* - \bar{\theta}) + (1-a)(\bar{\theta} - \bar{b})|^2 = a^2 R(\bar{\theta}^*, \bar{\theta}) + (1-a)^2 |\bar{\theta} - \bar{b}|^2 \\ &= ea^2(1 - |\bar{\theta}|^2) + (1-a)^2 |\bar{\theta} - \bar{b}|^2 \\ &= ((1-a)^2 - ea^2)|\bar{\theta}|^2 - 2(1-a)^2(\bar{\theta}, \bar{b}) + (1-a)^2 |\bar{b}|^2 + ea^2. \end{aligned}$$

Отсюда легко видеть, что при $a = 1/(1 + \sqrt{e})$, $\bar{b} = \bar{\theta}^0$ риск оценки $\bar{T}(a, \bar{b})$ не зависит от $\bar{\theta}$ и равен

$$-2(1-a)^2 \frac{1}{N} + (1-a)^2 |\bar{\theta}^0|^2 + ea^2 = (1-a)^2 \frac{N-1}{N} = \frac{e(N-1)}{(1+\sqrt{e})^2 N}.$$

Введем обозначение

$$\bar{T}_e = \bar{T}\left(\frac{1}{1+\sqrt{e}}, \bar{\theta}^0\right) = \frac{1}{1+\sqrt{e}}\bar{\theta}^* + \frac{\sqrt{e}}{1+\sqrt{e}}\bar{\theta}^0. \quad (9)$$

Как видно из предыдущих выкладок, справедливо следующее утверждение.

Теорема 2. В классе оценок вида

$$\bar{T}(a, \bar{b}) = a\bar{\theta}^* + (1-a)\bar{b}, \quad a \in [0, 1], \quad \bar{b} \in \Theta,$$

имеется единственная оценка $\bar{T}_e$, указанная в (9), риск которой $R(\bar{T}_e, \bar{\theta})$ не зависит от $\bar{\theta}$ и равен $e(N-1)/((1+\sqrt{e})^2 N)$.

Исследуем более детально свойства этой оценки $\bar{T}_e$. Прежде всего, из (7) следует, что

$$R(\bar{T}_e, \bar{\theta}) < \max_{\bar{\theta}} R(\bar{\theta}^*, \bar{\theta}),$$

т.е. по принципу минимакса оценка $\bar{T}_e$ лучше несмещенной оценки $\bar{\theta}^*$ (напомним, что согласно принципу минимакса из двух оценок лучшей считается та, для которой максимальное значение риска меньше).

Вычислим теперь зону предпочтения оценки $\bar{T}_e$ перед оценкой $\bar{\theta}^*$, т.е. множество тех значений параметра $\bar{\theta}$, где $R(\bar{T}_e, \bar{\theta}) < R(\bar{\theta}^*, \bar{\theta})$. Согласно (6)

$$G_e = \{\bar{\theta}: R(\bar{T}_e, \bar{\theta}) < R(\bar{\theta}^*, \bar{\theta})\} = \{\bar{\theta}: |\bar{\theta} - \bar{\theta}^0|^2 < \psi(e)(N-1)/N\},$$

где

$$\psi(e) = \sqrt{e}(2 + \sqrt{e})(1 + \sqrt{e})^{-2}.$$

Легко убедиться, что функция $\psi(e)$ монотонно возрастает, поэтому область G_e расширяется с ростом величины $e = (\alpha + n)/(n(\alpha + 1))$. В свою очередь,

$e \downarrow 1/(\alpha + 1)$ при $n \rightarrow \infty$, т.е. с ростом объема выборки n (при фиксированном α) зона G_e сужается. Также сужается зона G_e при росте α (и фиксированном n), так как тогда $e \downarrow 1/n$. Вообще же

$$e = e(\alpha, n) = \frac{1}{n} + \frac{n-1}{n} \frac{1}{\alpha+1} = \frac{1}{\alpha+1} + \frac{\alpha}{(\alpha+1)n},$$

т.е.

$$\max\left(\frac{1}{n}, \frac{1}{\alpha+1}\right) < e < \frac{1}{n} + \frac{1}{\alpha+1},$$

поэтому если одновременно n и α велики, то e мало и область предпочтений G_e также мала: она представляет собой малую окрестность точки $\bar{\theta}^0$.

Таким образом, если $\bar{\theta} \in G_e$, то оценка $\bar{T}_e$ точнее оценки $\bar{\theta}^*$, если же $\bar{\theta} \in \Theta \setminus G_e$, то более точной является оценка $\bar{\theta}^*$.

Наконец, проанализируем поведение риска оценки $\bar{T}_e$ в зависимости от изменения величин n и α . Так как риск $\bar{T}_e$ пропорционален функции $\varphi(e) = e(1 + \sqrt{e})^{-2}$, а эта функция монотонно возрастает по e , то минимальный риск имеет оценка $\bar{T}_{e'}$, где $e' = \inf e(\alpha, n)$. В частности, при заданном n величина $e' = 1/n$ и (см. (9))

$$\bar{T}_{e'} = \frac{\sqrt{n}}{\sqrt{n}+1} \bar{\theta}^* + \frac{1}{\sqrt{n}+1} \bar{\theta}^0 = \left(\frac{\eta_j(n) + \sqrt{n}/N}{n + \sqrt{n}}, j = 1, \dots, N \right). \quad (10)$$

Замечание 2. Оценка (10) фактически является предельным значением для оценки (9) при $\alpha \rightarrow \infty$. Но, как отмечается в [1], при таком предельном переходе, сохраняющем фиксированными значения параметров $\theta_j = \alpha_j/\alpha, j = 1, \dots, N$, МП-модель переходит в полиномиальную модель $M(n, \bar{\theta})$, для которой, как известно (10) определяет минимаксную оценку параметра $\bar{\theta}$. Этот факт подсказывает гипотезу, что оценка (9) является минимаксной и в допредельной ситуации, т.е. при любом заданном α .

В заключение отметим, что если $n, \alpha \rightarrow \infty$ одновременно, то $e \rightarrow 0$ и тогда $\bar{T}_e = \bar{\theta}^*(1 + o_p(1))$, а оценка $\bar{\theta}^*$ становится состоятельной.

Теперь уже легко ответить на важный вопрос о влиянии управляемого параметра c на свойства рассматриваемых оценок. Поскольку $\alpha = a/c$, где a — начальное заполнение урны, с увеличением числа c добавляемых шаров параметр α уменьшается, а величина $e = (\alpha + n)n^{-1}(\alpha + 1)^{-1}$ и функция $\varphi(e) = e(1 + \sqrt{e})^{-2}$ возрастают. Тем самым, риски оценок $\bar{\theta}^*$ и $\bar{T}_e$ увеличиваются с увеличением c , т.е. статистические свойства оценок ухудшаются. Отсюда следует вывод, что в условиях активного эксперимента, т.е. когда выбор параметра c зависит от исследователя, наилучшим является выбор минимального из возможных значений этого управляемого параметра. Идеальным же с этих позиций является вариант $c = 0$, т.е. следует вообще отказаться от добавления новых шаров: в этом случае мы имеем дело с урной постоянного состава $\bar{a} = (a_1, \dots, a_N)$, и задача сводится к оцениванию параметра $\bar{\theta} = \bar{a}/a$ полиномиальной модели $M(n, \bar{\theta})$. Полученное выше решение (10) как раз и соответствует этому случаю (см. замечание 2).

2.3. Оценивание в многовыборочной модели

В статистических задачах итоговые выводы, как правило, существенно зависят от структуры данных, т.е. от того, что наблюдается. Выше мы рассмотрели одновыборочную схему, а именно, предполагали, что наблюдается лишь одна реализация

случайного вектора $\bar{\eta}(n)$. Было показано, что в этом случае несмещенная оценка $\bar{\theta}^* = \bar{\eta}(n)/n$ оказывается несостоятельной (при $n \rightarrow \infty$ и фиксированном α). Таким образом, в такой схеме наблюдений даже при больших n информации в данных оказывается недостаточно для состоятельного оценивания параметра $\bar{\theta}$ в модели $\text{МП}(n, \alpha\bar{\theta})$. Тем не менее, состоятельности в рассматриваемой задаче можно все же добиться, но на другом пути, а именно, если схему наблюдений изменить так, чтобы можно было получать независимые реализации вектора $\bar{\eta}(n)$ и притом в большом числе. Этот случай мы и рассмотрим в настоящем разделе.

Итак, пусть имеется K независимых наблюдений

$$\bar{\eta}_i(n), \quad i = 1, \dots, K,$$

над случайным вектором $\bar{\eta}(n)$, имеющим распределение $\text{МП}(n, \alpha\bar{\theta})$. Тогда мы можем построить K несмещенных и независимых в совокупности оценок параметра $\bar{\theta}$, а именно,

$$\bar{\theta}_i^* = \bar{\eta}_i(n)/n, \quad i = 1, \dots, K.$$

Их среднее арифметическое

$$\bar{\theta}^*(K) = \frac{1}{K} \sum_{i=1}^K \bar{\theta}_i^* = \frac{1}{Kn} \sum_{i=1}^K \bar{\eta}_i(n) \quad (11)$$

также, очевидно, несмещенно оценивает $\bar{\theta}$.

Вычислим дисперсионную матрицу этой оценки. В силу независимости и одинаковой распределенности векторов $\bar{\eta}_i(n)$ из (4)

$$\mathbf{D}\bar{\theta}^*(K) = \frac{1}{Kn^2} \mathbf{D}\bar{\eta}(n) = \frac{e(\alpha, n)}{K} (\text{diag } \bar{\theta} - \bar{\theta}\bar{\theta}')$$

(здесь $\text{diag } \bar{\theta}$ обозначает диагональную матрицу с диагональю $\bar{\theta}$, векторы понимаются как вектор-столбцы, штрих означает транспонирование). Отсюда следует, что риск (2) оценки $\bar{\theta}^*(K)$ равен

$$R(\bar{\theta}^*(K), \bar{\theta}) = \sum_{j=1}^K \mathbf{D}\theta_j^*(K) = \frac{e(\alpha, n)}{K} (1 - |\bar{\theta}|^2),$$

т.е. он в K раз меньше риска одновыборочной оценки $\bar{\theta}^* = \bar{\theta}^*(1)$ (см. (5)). Если число наблюдений $K \rightarrow \infty$, то риск K -выборочной оценки (11) стремится к 0, и тем самым эта оценка является состоятельной.

Подчеркнем, что в рассматриваемом случае параметры модели n и α остаются фиксированными, а общий ресурс испытаний, равный Kn , неограниченно растет за счет возрастания числа K независимых реализаций вектора $\bar{\eta}(n)$. В одновыборочном же случае, хотя число испытаний n также неограниченно растет, доставляемая ими информация в силу зависимости испытаний нарастает недостаточно быстро для того, чтобы обеспечить свойство состоятельности соответствующей несмещенной оценки. Тем самым, мы имеем пример ситуации, когда наличие фактора независимости в структуре данных оказывает решающее влияние на повышение качества соответствующей оценки.

3. Проверка гипотез в модели Маркова–Пойа

В этом разделе рассматриваются различные задачи проверки статистических гипотез в МП-модели: гипотезы независимости (простые и сложные), гипотезы о наличии рандомизации, гипотезы о смешивающей мере Дирихле, гипотезы однородности, а также смежные вопросы доверительного оценивания неизвестных параметров; при этом анализ проводится как в классической схеме больших выборок, так и в условиях схемы серий, а также многовыборочной схемы.

3.1. Перестановочные последовательности и гипотезы для них

Важнейшим свойство МП-модели является свойство перестановочности, состоящее в следующем [1] (далее мы используем язык урновой схемы). Пусть X_i обозначает номер цвета i -го извлеченного из урны шара, $i = 1, 2, \dots$. Тогда совместное распределение любого конечного набора $(X_{i_1}, \dots, X_{i_k})$ этих случайных величин не зависит от их индексов, а зависит лишь от их числа k . Такие случайные последовательности называются перестановочными, тем самым введенная последовательность $\{X_i\}$ является перестановочной.

Известно, что каждая перестановочная последовательность является смесью независимых одинаково распределенных переменных, порождаемой некоторой смешивающей мерой P на множестве соответствующих функций распределения. Если рассматриваемые случайные величины дискретны и принимают значения в конечном множестве $\{1, 2, \dots, N\}$ (как в нашем случае), то смешивающая мера есть распределение случайного вектора $\bar{p} = (p_1, \dots, p_N)$, где p_j есть вероятность появления исхода j , $j = 1, \dots, N$, $p_1 + \dots + p_N = 1$. Другими словами, любая перестановочная последовательность со значениями в $\{1, 2, \dots, N\}$ может быть получена рандомизацией с помощью некоторой меры P на симплексе

$$\mathcal{P}_N = \{\bar{p} = (p_1, \dots, p_N), p_j \geq 0, j = 1, \dots, N, p_1 + \dots + p_N = 1\}$$

полиномиальной последовательности с распределением $\bar{p}$.

В частности, если смешивающая мера P соответствует распределению Дирихле $\mathcal{D}(\alpha)$, (см. (8)), то мы имеем последовательность Маркова–Пойа $\{X_i, i \geq 1\}$ (см. [1]). Таким образом, введенная МП-последовательность $\{X_i\}$ является рандомизированной полиномиальной последовательностью, параметр $\bar{p}$ которой является случайным вектором с распределением $\mathcal{L}(\bar{p}) = \mathcal{D}(\alpha)$, $\bar{p} \in \mathcal{P}_N$.

Подчеркнем, что вырожденным (атомическим) мерам P (и только им) соответствуют последовательности независимых величин (полиномиальные последовательности). Напомним также (см. [1]), что если $\alpha, \alpha_j \rightarrow \infty$ так, что

$$\alpha_j/\alpha \rightarrow p_j^0, \quad 0 < p_j^0 < 1, \quad j = 1, \dots, N,$$

то распределение $\mathcal{D}(\bar{\alpha})$ сходится к вырожденному в точке $\bar{p}^0$ распределению, и, следовательно, МП-последовательность в пределе становится полиномиальной с распределением $\bar{p}^0$.

Теперь, с учетом изложенного, мы можем перейти к формулировке статистических задач.

Пусть наблюдается выборка $(X_1, \dots, X_n)$ из последовательности $\{X_i\}_i^\infty$ дискретных случайных величин со значениями в множестве $\{1, \dots, N\}$. По этой информации требуется проверить следующие гипотезы:

$H_0(\bar{p}^0): \{X_i\}$ — полиномиальная последовательность с заданным распределением $\bar{p}^0 \in \mathcal{P}_N$;

$H_0: \{X_i\}$ — полиномиальная последовательность с некоторым (неизвестным) распределением;

$H_1(\bar{\alpha}): \{X_i\}$ является МП-последовательностью с заданной смешивающей мерой $\mathcal{D}(\bar{\alpha})$;

$H_1: \{X_i\}$ — МП-последовательность с неизвестным параметром $\bar{\alpha}$;

$H_2: \{X_i\}$ — перестановочная последовательность с некоторой (неизвестной) невырожденной смешивающей мерой P .

Подчеркнем, что гипотезы $H_0(\bar{p}^0)$ и $H_1(\bar{\alpha})$ простые, а остальные гипотезы сложные, при этом $H_1(\bar{\alpha}) \subset H_1 \subset H_2$. Представляет также интерес различие двух простых гипотез $H_1(\bar{\alpha}^{(1)})$ и $H_1(\bar{\alpha}^{(2)})$, $\bar{\alpha}^{(1)} \neq \bar{\alpha}^{(2)}$.

Таким образом, речь далее будет идти о проверке некоторых естественных гипотез в рамках общей модели перестановочных дискретных последовательностей. В этом контексте гипотезы $H_0(\bar{p}^0)$ и H_0 можно трактовать как гипотезы независимости, гипотезы $H_1(\bar{\alpha})$ и H_1 как МП-гипотезы, а H_2 как общую гипотезу о наличии рандомизации или, другими словами, гипотезу о том, что случайные величины $\{X_i\}$ условно независимы и (условно) одинаково распределены.

На практике часто требуется проверить гипотезу о том, что наблюдения $(X_1, \dots, X_n)$ или, что эквивалентно, вектор частот $\bar{\eta}(n) = (\eta_1(n), \dots, \eta_N(n))$ соответствуют равновероятной полиномиальной схеме, т.е. гипотезу $H_0(\bar{p}^0)$ при $\bar{p}^0 = (1/N, \dots, 1/N)$. При этом в качестве альтернативы выбирают, как правило некоторую другую (неравновероятную) полиномиальную схему $H_0(\bar{p})$, $\bar{p} \neq \bar{p}^0$. Однако такой подход не всегда применим, так как обычно альтернатива нам неизвестна. В таких ситуациях естественно рассматривать некоторую усредненную альтернативу, получаемую усреднением всех неравновероятных схем с N исходами по некоторой вероятностной мере на симплексе $\mathcal{P}_N$, т.е. когда вектор вероятностей исходов $\bar{p} = (p_1, \dots, p_N)$ рассматривается как случайный параметр с некоторым распределением $\mathcal{L}(\bar{p}) = \mathbf{P}$ на $\mathcal{P}_N$. При таком подходе мы приходим к модели перестановочной последовательности, в которой проверяется гипотеза независимости $H_0(\bar{p}^0)$. При дополнительном априорном предположении, что смешивающая мера P порождается распределением Дирихле $\mathcal{D}(\bar{\alpha})$, мы конкретизируем альтернативную гипотезу до МП-гипотезы $H_1(\bar{\alpha})$.

3.2. Правдоподобие данных и достаточная статистика

В общем случае, если $(X_1, \dots, X_n)$ — наблюдения над перестановочными случайными величинами, то, как отмечено в разделе 2.1, правдоподобие данных есть

$$\mathbf{P}\{(X_1, \dots, X_n) = (j_1, \dots, j_n)\} = \int_{\mathcal{P}_N} \prod_{j=1}^N p_j^{n_j} d\mathbf{P}(\bar{p}), \quad (12)$$

где $\mathbf{P}$ — соответствующая смешивающая мера, а n_j — число реализаций исхода j в выборке, $j = 1, \dots, N$. Из (12) следует, что вектор

$$\bar{\eta}(n) = \left(\eta_j(n) = \sum_{i=1}^n I(X_i = j), j = 1, \dots, N \right)$$

является достаточной статистикой выборки $(X_1, \dots, X_n)$. При этом, если $\mathbf{P} = \mathcal{D}(\bar{\alpha})$, то $\mathcal{L}(\bar{\eta}(n)) = \text{МП}(n, \bar{\alpha})$ (см. (1)).

Известно (см. [1]), что при $n \rightarrow \infty$ (большие выборки) $\bar{\eta}(n)/n \rightarrow \bar{p}$ (п. н.), где $\bar{p}$ — случайный вектор с распределением $\mathbf{P}$ на симплексе $\mathcal{P}_N$. Таким образом, для больших выборок задача проверки гипотез о законе распределения перестановочной последовательности $\{X_i\}$ асимптотически сводится к проверке соответствующих гипотез о смешивающей мере $\mathbf{P}$ по наблюдениям над вектором $\bar{p} \approx \bar{\eta}(n)/n$.

Установим еще важное свойство полноты достаточной статистики $\bar{\eta}(n)$, именно, что

$$\mathbf{E}_{\mathbf{P}} T(\bar{\eta}(n)) = 0 \quad \forall \mathbf{P} \implies T \equiv 0$$

(на $\mathcal{X}_{N_n}$, см. (1)). Действительно, по формуле полного математического ожидания

$$\mathbf{E}_{\mathbf{P}} T(\bar{\eta}(n)) = \mathbf{E}_{\mathbf{P}} (\mathbf{E}_{\bar{p}} (T(\bar{\eta}(n) \mid \bar{p}))).$$

Из условия, что это математическое ожидание равно нулю для любой меры $\mathbf{P}$ на $\mathcal{P}_N$ следует его равенство нулю для всех атомических мер, т.е. равенство нулю $\mathbf{E}_{\bar{p}} (T(\bar{\eta}(n) \mid \bar{p}))$ для всех $\bar{p} \in \mathcal{P}_N$. Но для вырожденной меры $\mathbf{P}$ вектор $\bar{\eta}(n)$ имеет полиномиальное распределение, а, как известно, в полиномиальной модели вектор частот $\bar{\eta}(n)$ обладает свойством полноты, т.е.

$$\mathbf{E}_{\bar{p}} T(\bar{\eta}(n) \mid \bar{p}) = 0 \quad \forall \mathbf{P} \implies T \equiv 0,$$

что и требовалось доказать.

Замечание 3. Свойство полноты достаточная статистика $\bar{\eta}(n)$ сохраняет и в более узкой модели Маркова–Пояа, т.е. когда класс смешивающих мер совпадает с классом мер Дирихле $\{\mathcal{D}(\bar{\alpha})\}$, так как (как упоминалось выше) каждое полиномиальное распределение является пределом для некоторой последовательности МП-распределений.

3.3. Проверка гипотез в случае больших выборок

3.3.1. Рассмотрим сначала задачу проверки простой гипотезы $H_0(\bar{p}^0)$ при конкурирующих гипотезах вида $H_1(\bar{\alpha})$. При гипотезе $H_0(\bar{p}^0)$ с вероятностью 1 $\bar{\eta}(n)/n \rightarrow \bar{p}^0$ при $n \rightarrow \infty$, а при альтернативе $H_1(\bar{\alpha})$ соответствующий предел есть случайная величина с распределением $\mathcal{D}(\bar{\alpha})$, поэтому отсюда следует, что гипотеза $H_0(\bar{p}^0)$ идентифицируется в этой ситуации с вероятностью, стремящейся к 1 при $n \rightarrow \infty$. Именно, если задать критическую область, например, в виде

$$\mathcal{G}_{\varepsilon}(\bar{p}^0) = \left\{ \left| \frac{\bar{\eta}(n)}{n} - \bar{p}^0 \right| > \varepsilon \right\}, \quad \varepsilon > 0,$$

то вероятность ошибки первого рода

$$\mathbf{P}\{\mathcal{G}_{\varepsilon}(\bar{p}^0) \mid H_0(\bar{p}^0)\} \rightarrow 0$$

для любого $\varepsilon > 0$ при $n \rightarrow \infty$.

Вероятность же ошибки второго рода будет удовлетворять соотношению

$$\mathbf{P}\{\bar{\mathcal{G}}_{\varepsilon}(\bar{p}^0) \mid H_1(\bar{\alpha})\} \rightarrow \int_{|\bar{p} - \bar{p}^0| \leq \varepsilon} g(\bar{p}, \bar{\alpha}) d\bar{p}$$

(см. (8)), что выбором достаточно малого ε интеграл в правой части соотношения может быть сделан сколь угодно малым.

Таким образом, для больших выборок гипотезы $H_0(\bar{p}^0)$ и $H_1(\bar{\alpha})$ при любом $\bar{\alpha}$ различаются почти достоверно. Очевидно, такой же вывод справедлив и в задаче $(H_0(\bar{p}^0), H_2)$ проверки $H_0(\bar{p}^0)$ внутри H_2 .

Замечание 4. Для различения гипотез $H_0(\bar{p}^0)$ и $H_1(\bar{\alpha})$ в данной ситуации нетрудно построить и асимптотический вариант критерия Неймана–Пирсона. Именно, статистика отношения правдоподобия $l_n = l_n(H_1; H_0)$ равна в этом случае (см. [1])

$$l_n = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_N) \Gamma(\alpha + n)} \prod_{j=1}^N \frac{\Gamma(\alpha_j + \eta_j(n))}{(p_j^0)^{\eta_j(n)}},$$

и с использованием формулы Стирлинга для гамма-функции нетрудно получить, что при $n \rightarrow \infty$ с вероятностью, стремящейся к 1,

$$\frac{1}{n} \ln l_n \sim \sum_{j=1}^N \frac{\eta_j(n)}{n} \ln \frac{\eta_j(n)}{np_j^0} \rightarrow \begin{cases} 0 & \text{при } H_0(\bar{p}^0), \\ F(\bar{p}) \equiv \sum_{j=1}^N p_j \ln(p_j/p_j^0) & \text{при } H_1(\bar{\alpha}). \end{cases}$$

Следовательно, критерий асимптотически имеет вид: если $F(\bar{\eta}(n)/n) > z$, то принимаем $H_1(\bar{\alpha})$, в противном случае принимаем $H_0(\bar{p})$, и его ошибки суть

$$\begin{aligned} P\{H_1(\alpha)|H_0(\bar{p}^0)\} &\sim P\{F(\bar{\eta}(n)/n) > z|H_0(\bar{p}^0)\} \rightarrow 0 \quad \forall z > 0, \\ P\{H_0(\bar{p}^0)|H_1(\bar{\alpha})\} &\sim \int_{F(\bar{p}) \leq z} g(\bar{p}, \bar{\alpha}) d\bar{p}, \end{aligned}$$

и последний интеграл мал при малом z .

3.3.2. Рассмотрим теперь задачу проверки сложной гипотезы независимости H_0 при МП-альтернативах типа $H_1(\bar{\alpha})$. Теперь уже при гипотезе H_0 нам неизвестна точка вырождения смешивающей меры, и потому по одной выборке (по одной реализации $\bar{\eta}(n)$) идентифицировать H_0 невозможно. Однако, если в нашем распоряжении есть две независимые (и однородные) выборки

$$(X_1^{(1)}, \dots, X_{n_1}^{(1)}), \quad (X_1^{(2)}, \dots, X_{n_2}^{(2)}),$$

то при гипотезе H_0 соответствующие статистики $\bar{\eta}^{(1)}(n_1)/n_1$ и $\bar{\eta}^{(2)}(n_2)/n_2$ сходятся при $n_1, n_2 \rightarrow \infty$ к одной и той же постоянной. Следовательно, с вероятностью 1

$$\left| \frac{\bar{\eta}^{(1)}(n_1)}{n_1} - \frac{\bar{\eta}^{(2)}(n_2)}{n_2} \right| \rightarrow 0.$$

Если же справедлива некоторая альтернатива $H_1(\bar{\alpha})$, то эта разность с вероятностью 1 имеет пределом $|\bar{p}^{(1)} - \bar{p}^{(2)}|$, где $p^{(1)}$ и $p^{(2)}$ — независимые копии случайной величины $\bar{p}$ с распределением $\mathcal{D}(\bar{\alpha})$.

Теперь уже ясно, как построить критерий, различающий гипотезы H_0 и $H_1(\bar{\alpha})$. Именно, если выбрать критическую область в виде

$$\mathcal{G}_\varepsilon = \left\{ \left| \frac{\bar{\eta}^{(1)}(n_1)}{n_1} - \frac{\bar{\eta}^{(2)}(n_2)}{n_2} \right| > \varepsilon \right\}, \quad (13)$$

то $\mathbf{P}(\mathcal{E}_\varepsilon | H_0) \rightarrow 0$ при $n_1, n_2 \rightarrow \infty$ для любого $\varepsilon > 0$, а

$$\mathbf{P}\{\bar{\mathcal{E}}_\varepsilon | H_1(\bar{\alpha})\} \rightarrow \int_{|\bar{p}^{(1)} - \bar{p}^{(2)}| \leq \varepsilon} g(\bar{p}^{(1)}; \bar{\alpha}) g(\bar{p}^{(2)}; \bar{\alpha}) d\bar{p}^{(1)} d\bar{p}^{(2)},$$

и последний интеграл может быть сделан сколь угодно малым за счет выбора достаточно малого ε .

Таким образом, в рассматриваемой модели гипотезы независимости (как простая $H_0(\bar{p}^0)$, так и сложная H_0) могут быть идентифицированы при наличии соответствующей информации (данных) почти достоверно.

3.3.3. Пусть, наконец, требуется различить две простые МП-гипотезы $H_1(\bar{\alpha}^{(i)})$, $i = 1, 2$, по большой выборке $(X_1, \dots, X_n)$. Здесь $\bar{\eta}(n)/n \rightarrow \bar{p}$ (п. н.) и

$$\mathcal{L}(\bar{p} | H_1(\bar{\alpha}^{(i)})) = \mathcal{D}(\bar{\alpha}^{(i)}), \quad i = 1, 2.$$

Поэтому в пределе задача сводится к различению двух распределений Дирихле по одному наблюдению над соответствующей величиной $\bar{p}$.

Статистика наиболее мощного критерия Неймана–Пирсона для предельной задачи имеет здесь вид

$$l(\bar{p}) = \frac{g(\bar{p}; \bar{\alpha}^{(1)})}{g(\bar{p}; \bar{\alpha}^{(2)})} = \frac{A_1}{A_2} \prod_{j=1}^N p_j^{\alpha_j^{(1)} - \alpha_j^{(2)}},$$

где

$$A_i = \frac{\Gamma(\alpha^{(i)})}{\prod_{j=1}^N \Gamma(\alpha_j^{(i)})}, \quad i = 1, 2,$$

или

$$l'(\bar{p}) \equiv \ln l(\bar{p}) = \sum_{j=1}^N (\alpha_j^{(1)} - \alpha_j^{(2)}) \ln p_j + \ln \frac{A_1}{A_2}. \quad (14)$$

Поэтому критерий Неймана–Пирсона асимптотически выглядит в данной задаче следующим образом: если $l'(\bar{\eta}(n)/n) > z$, то принимаем $H_1(\bar{\alpha}^{(1)})$, в противном случае принимаем $H_1(\bar{\alpha}^{(2)})$. Ошибки такого критерия вычисляются по следующим формулам: вероятность ошибки первого рода

$$\mathbf{P}\{H_1(\bar{\alpha}^{(2)}) | H_1(\bar{\alpha}^{(1)})\} = \mathbf{P}\{l'(\bar{\eta}(n)/n) \leq z | H_1(\bar{\alpha}^{(1)})\} \rightarrow \int_{l'(\bar{p}) \leq z} g(\bar{p}; \bar{\alpha}^{(1)}) d\bar{p}$$

при $n \rightarrow \infty$, а вероятность ошибки второго рода

$$\mathbf{P}\{H_1(\bar{\alpha}^{(1)}) | H_1(\bar{\alpha}^{(2)})\} = \mathbf{P}\{l'(\bar{\eta}(n)/n) > z | H_1(\bar{\alpha}^{(2)})\} \rightarrow \int_{l'(\bar{p}) > z} g(\bar{p}; \bar{\alpha}^{(2)}) d\bar{p}$$

при $n \rightarrow \infty$.

3.3.4. Рассмотрим применение описанной выше методики к урновой МП-схеме. Пусть известен первоначальный состав урны $\bar{a} = (a_1, \dots, a_N)$, но неизвестен механизм извлечения шаров, т. е. неизвестен параметр c . Это означает, что при $\bar{p}^0 = \bar{a}/a$ гипотеза $H_0(\bar{p}^0)$ соответствует тому, что $c = 0$, а гипотеза, фиксирующая значение $c > 0$, есть гипотеза $H_1(\bar{\alpha})$ с $\bar{\alpha} = \bar{a}/c$. Общая же гипотеза H_1 означает, что $c > 0$, без фиксации конкретного значения этого параметра.

Замечание 5. Случай бесповторного выбора, т. е. гипотеза $c = -1$, в рамках данного подхода идентифицируется при $n \geq a$ фактом исчерпания урны (обрывом на a -м шаге последовательности $\{X_i\}$).

В свете сказанного $\mathcal{L}(\bar{\eta}(n)|c = 0) = M(n; \bar{p}^0)$, а $\mathcal{L}(\bar{\eta}(n) | c > 0) = \text{МП}(n; \bar{\alpha})$, и согласно преддущему, критерий $\mathcal{G}_\varepsilon(\bar{p}^0)$ при достаточно малом $\varepsilon > 0$ будет почти достоверно различать гипотезу $c = 0$ и любую простую альтернативу $c > 0$.

Если же и состав урны $\bar{a}$ неизвестен, то речь идет о проверке сложной гипотезы H_0 при конкурирующей также сложной гипотезе H_1 . В этой ситуации для различения данных гипотез требуется иметь две независимые реализации одной и той же урновой схемы и использовать критерий $\mathcal{G}_\varepsilon$, определенный формулой (13).

Наконец, если требуется различить две простые гипотезы о МП-механизме: $c = c_1 > 0$ и $c = c_2 > 0$ (состав $\bar{a}$ неизвестен), то статистика Неймана–Пирсона (14) принимает вид

$$l'(\bar{p}) = \left(\frac{1}{c_1} - \frac{1}{c_2} \right) \sum_{j=1}^N a_j \ln p_j + \ln \frac{A_1}{A_2},$$

и далее работает общая схема критерия Неймана–Пирсона.

3.4. Оценивание параметров в схеме серий

Выше предполагалось, что параметры $\bar{a} = (a_1, \dots, a_N)$ гипотезы $H_1(\bar{\alpha})$ фиксированы, в то время как объем выборки $n \rightarrow \infty$. Рассмотрим теперь ситуацию типа схемы серий, когда $\alpha = \alpha(n) \rightarrow \infty$ при $n \rightarrow \infty$. Соответствующие предельные теоремы для достаточной статистики $\bar{\eta}(n) = (\eta_1(n), \dots, \eta_N(n))$, необходимые для конструирования критериев в такой ситуации, приведены в [1]. Мы выделим здесь для детального анализа наиболее практически интересный случай схемы серий, характеризуемый наличием нормальной аппроксимации для распределения статистики $\bar{\eta}(n)$. Именно, далее будем считать, что выполняются следующие условия (A):

- (1) величина $\alpha = \alpha_1 + \dots + \alpha_N$ известна и $\alpha = \alpha(n) \rightarrow \infty$ при $n \rightarrow \infty$ так, что $d = n/\alpha \leq c < \infty$;
- (2) параметры $p_j = \alpha_j/\alpha$, $j = 1, \dots, N$, неизвестны ($\bar{p} = (p_1, \dots, p_N) \in \mathcal{P}_N$) и существует такое $\delta > 0$, что $p_j \in (\delta, 1 - \delta)$ для всех $j = 1, \dots, N$.

В терминах урновой МП-схемы эти предположения означают, что в стартовый момент времени в урне находится много шаров и общее их число известно, при этом количества шаров различных цветов сравнимы между собой, но их точные значения (или доли $p_1, \dots, p_N$) неизвестны. Относительно же параметров N и $c > 0$ далее мы будем считать, что они фиксированы и известны.

В этой ситуации мы рассмотрим соответствующие вопросы проверки гипотез и доверительного оценивания неизвестных параметров $\bar{p}$. В дальнейшем мы также

будем использовать обозначение $\bar{p}_{N-1} = (p_1, \dots, p_{N-1})$ для укороченного параметрического вектора.

3.4.1. Прежде всего отметим, что в условиях (A)

$$e(\alpha, n) = \frac{\alpha + n}{n(\alpha + 1)} = \frac{1 + d}{n(1 + \alpha^{-1})} \rightarrow 0,$$

и потому статистика $\bar{p}^* = \bar{\eta}(n)/n$ является несмещенной и состоятельной оценкой для параметра $\bar{p}$ (см. раздел 1.1). Более того, в рассматриваемых условиях эта статистика обладает свойством асимптотической нормальности, что следует из теоремы 3 в [1]. Переформулировка этой теоремы в наших обозначениях имеет следующий вид.

Теорема 3. Пусть $n \rightarrow \infty$ и выполняются условия (A). Тогда

$$\mathcal{L}(\sqrt{n}(\bar{p}_{N-1}^* - \bar{p}_{N-1})) \rightarrow \mathcal{N}(\bar{0}, \Sigma_{N-1}),$$

где

$$\Sigma_{N-1} = \|\sigma_{ij} = (1 + d)p_i(\delta_{ij} - p_j)\|_1^{N-1}$$

и δ_{ij} — символ Кронекера. Кроме того, если функция $\tau(\bar{p})$ непрерывно дифференцируема, то

$$\mathcal{L}(\sqrt{n}(\tau_n^* - \tau(\bar{p}))/\sigma_n^*) \rightarrow \mathcal{N}(0, 1),$$

где $\tau_n^* = \tau(\bar{p}^*)$, $\sigma_n^* = \sigma(\bar{p}^*)$ и

$$\sigma^2(\bar{p}) = \bar{b}' \Sigma_{N-1} \bar{b}, \quad \bar{b} = \bar{b}(\bar{p}) = \left(\frac{\partial \tau(\bar{p})}{\partial p_1}, \dots, \frac{\partial \tau(\bar{p})}{\partial p_{N-1}} \right),$$

при этом предполагается, что $\sigma(\bar{p}) \neq 0$ при $\bar{p} \in \mathcal{P}_N$.

Замечание 6. Утверждение теоремы относительно параметрических функций $\tau(\bar{p})$ является следствием общих теорем сходимости для функций от случайных величин (см., например, [2], с. 335–339).

Основываясь на теореме 3 и применяя стандартную статистическую технику, можно рассчитывать асимптотические доверительные интервалы (области) для неизвестных параметров $\bar{p}$ и функций от них. Так, для произвольной непрерывно дифференцируемой функции $\tau(\bar{p})$ асимптотически несмещенной и состоятельной оценкой в силу теоремы 3 является статистика $\tau_n^* = \tau(\bar{p}^*)$, а асимптотический γ -доверительный интервал для $\tau(\bar{p})$ имеет вид

$$(\tau_n^* \pm c_\gamma \sigma_n^* / \sqrt{n}), \tag{15}$$

где $c_\gamma = \varphi^{-1}((1 + \gamma)/2)$ и φ^{-1} есть обратная функция к

$$\varphi(x) = 1/\sqrt{2\pi} \int_{-\infty}^x e^{-t^2/2} dt.$$

В качестве примера применения этого общего результата рассмотрим задачу оценивания отдельной доли, скажем, p_1 . Пусть $\tau(\bar{p}) = \tau(p_1)$, тогда

$$\sigma^2(p_1) = (1 + d)p_1(1 - p_1)(\tau'(p_1))^2,$$

и следовательно, интервал (15) сводится к интервалу

$$(\tau(p_1^*) \pm c_\gamma \tau'(p_1^*) \sqrt{(1+d)p_1^*(1-p_1^*)/n}), \quad p_1^* = \eta_1(n)/n.$$

В частности, при $\tau(p_1) = \arcsin \sqrt{p_1}$ этот интервал имеет вид

$$(\tau(p_1^*) \pm (c_\gamma/2) \sqrt{(1+d)/n}),$$

т. е. он имеет фиксированную длину $c_\gamma \sqrt{(1+d)/n}$, а асимптотический γ -доверительный интервал для самого параметра p_1 есть

$$(p_1^* \pm c_\gamma \sqrt{(1+d)p_1^*(1-p_1^*)/n}).$$

3.4.2. Рассмотрим теперь задачу одновременного оценивания всех координат вектора долей $\bar{p}$. В этом случае требуется при заданном доверительном уровне γ построить такое случайное подмножество $\mathcal{G}_{\gamma,n} = \mathcal{G}_\gamma(\bar{\eta}(n)) \subset \mathcal{P}_N$, для которого при $n \rightarrow \infty$ выполняется соотношение

$$\mathbf{P}\{\bar{p} \in \mathcal{G}_{\gamma,n}\} \rightarrow \gamma.$$

Такое асимптотическое γ -доверительное подмножество также может быть построено на основании теоремы 3.

Действительно, из этой теоремы следует, что статистика

$$X_n^2 = n(\bar{p}_{N-1}^* - \bar{p}_{N-1})' \Sigma_{N-1}^{-1} (\bar{p}_{N-1}^* - \bar{p}_{N-1})$$

(здесь векторы понимаются как векторы-столбцы, штрих означает транспонирование) имеет в пределе (при $n \rightarrow \infty$) распределение $\chi^2(N-1)$. Учитывая, что (см. [1])

$$\Sigma_{N-1}^{-1} = \frac{1}{1+d} \left\| \frac{\delta_{ij}}{p_i} + \frac{1}{p_j} \right\|_1^{N-1},$$

нетрудно получить представление

$$X_n^2 = X_n^2(\bar{p}) = \frac{1}{1+d} \sum_{j=1}^N \frac{(\eta_j(n) - np_j)^2}{np_j} = \frac{1}{1+d} \sum_{j=1}^N \frac{(p_j^* - p_j)^2}{p_j}. \quad (16)$$

Обозначая теперь $\chi_{\gamma,N-1}^2$ γ -квантиль распределения $\chi^2(N-1)$ и определяя множество равенством

$$\mathcal{G}_{\gamma,n} = \left\{ \bar{p}: \sum_{j=1}^N \frac{(p_j^* - p_j)^2}{p_j} \leq \frac{1+d}{n} \chi_{\gamma,N-1}^2, \sum_{j=1}^N p_j = 1 \right\}, \quad (17)$$

получаем, что

$$\mathbf{P}(\bar{p} \in \mathcal{G}_{\gamma,n}) = \mathbf{P}(X_n^2 \leq \chi_{\gamma,N-1}^2) \rightarrow \gamma.$$

Таким образом, указанное в (17) множество является искомым (асимптотическим) γ -доверительным множеством для $\bar{p}$.

Замечание 7. Если в (17) параметры p_j в знаменателях квадратичной формы заменить их состоятельными оценками $p_j^* = \eta_j(n)/n$, то полученное при такой модификации множество $\mathcal{G}_{\gamma,n}^*$ также будет асимптотическим γ -доверительным множеством для $\bar{p}$. Оно представляет собой пересечение случайного эллипсоида с центром в точке $\bar{p}^*$ с гиперплоскостью $p_1 + \dots + p_N = 1$. В частности, при $N = 2$ множество $\mathcal{G}_{\gamma,n}^*$ сводится к доверительному интервалу для p_1 , построенному выше.

Аналогично решается и более общая задача одновременного оценивания заданных линейных комбинаций параметров p_j . Пусть требуется построить доверительное множество для $r \leq N - 1$ линейных комбинаций

$$\bar{l} = (l_1, \dots, l_r): \bar{l} = L\bar{p}_{N-1},$$

где L — заданная $r \times (N - 1)$ матрица ранга r . Тогда из теоремы 3 следует, что для $\bar{l}^* = L\bar{p}_{N-1}^*$ выполняется асимптотическое (при $n \rightarrow \infty$) соотношение

$$\mathcal{L}(\sqrt{n}(\bar{l}^* - \bar{l})) \rightarrow \mathcal{N}(0, \Sigma_L),$$

где $\Sigma_L = L\Sigma_{N-1}L'$. Отсюда следует, что

$$\mathcal{L}(n(\bar{l}^* - \bar{l})'\Sigma_L^{-1}(\bar{l}^* - \bar{l})) \rightarrow \chi^2(r),$$

при этом здесь матрицу $\Sigma_L^{-1} = \Sigma_L^{-1}(\bar{p})$ можно заменить на

$$\Sigma_L^{*-1} \equiv \Sigma_L^{-1}(\bar{p}^*).$$

На основании этого, как и выше, получаем следующий асимптотический γ -доверительный эллипсоид для $\bar{l}$:

$$\mathcal{G}_{L,\gamma,n} = \{\bar{l}: n(\bar{l}^* - \bar{l})'\Sigma_L^{*-1}(\bar{l}^* - \bar{l}) \leq \chi_{\gamma,r}^2\}. \quad (18)$$

Отметим несколько частных случаев результата (18). Рассмотрим одну линейную комбинацию

$$l = \bar{\lambda}'\bar{p}_{N-1} = \lambda_1 p_1 + \dots + \lambda_{N-1} p_{N-1}.$$

Тогда эллипсоид (18) сводится к интервалу

$$(\bar{\lambda}'\bar{p}_{N-1}^* \pm c_\gamma \sqrt{(\bar{\lambda}'\Sigma_{N-1}(\bar{p}^*)\bar{\lambda})/n}).$$

В качестве второго примера построим доверительное множество для части координат вектора $\bar{p}$. Рассмотрим произвольный подвектор

$$\bar{p}(r) = (p_{j_1}, \dots, p_{j_r}), \quad 1 \leq j_1 < \dots < j_r < N, \quad r < N - 1.$$

Тогда его можно представить в виде $L\bar{p}_{N-1}$, где матрица L имеет в m -й строке j_m -й элемент, равный единице, а остальные элементы, равные нулю, $m = 1, \dots, r$. Но в этом случае матрица $\Sigma_L = \Sigma_{N-1}(j_1, \dots, j_r)$ совпадает с минором матрицы Σ_{N-1} , получающимся в пересечении строк и столбцов с указанными номерами. Следовательно, из (18) получаем, что асимптотический γ -доверительный эллипсоид для $\bar{p}(r)$ имеет вид $\{\bar{p}(r): n(\bar{p}^*(r) - \bar{p}(r))'\Sigma_L^{*-1}(j_1, \dots, j_r)(\bar{p}^*(r) - \bar{p}(r)) \leq \chi_{\gamma,r}^2\}$.

3.4.3. В общем случае построение доверительных эллипсоидов типа (17), (18) представляет собой трудную вычислительную задачу. Поэтому для практических применений предпочтительнее иметь результат, позволяющий строить доверительные интервалы для каждой из компонент оцениваемого параметрического вектора, с заданной вероятностью одновременно покрывающего соответствующие координаты, или, другими словами, систему T совместных доверительных интервалов (доверительных параллелепипедов). Соответствующий алгоритм нетрудно построить, следуя известному в дисперсионном анализе методу Шеффе [3]. Для его вывода нам понадобится следующее утверждение из теории квадратичных форм.

Лемма. Пусть B — положительно определенная матрица и $\bar{t}, \bar{\lambda}$ — вектор-столбцы. Тогда

$$\bar{t}' B \bar{t} = \max_{\bar{\lambda}} \frac{(\bar{\lambda}' \bar{t})^2}{\bar{\lambda}' B^{-1} \bar{\lambda}}. \quad (19)$$

Доказательство. Для доказательства (19) надо положить $\bar{X} = H' \bar{t}$, $\bar{Y} = H^{-1} \bar{\lambda}$, где $B = H H'$ (здесь $H = U \Lambda^{1/2}$ и U — ортогональная матрица, приводящая B к диагональному виду Λ). Тогда

$$\bar{X}' \bar{Y} = \bar{\lambda}' \bar{t}, \quad \bar{X}' \bar{X} = \bar{t}' B \bar{t}, \quad \bar{Y}' \bar{Y} = \bar{\lambda}' B^{-1} \bar{\lambda},$$

и по неравенству Коши-Буняковского

$$(\bar{X}' \bar{Y})^2 \leq (\bar{X}' \bar{X})(\bar{Y}' \bar{Y}),$$

то есть $\bar{t}' B \bar{t} \geq (\bar{\lambda}' \bar{t})^2 / (\bar{\lambda}' B^{-1} \bar{\lambda})$ для всех $\bar{\lambda}$. Лемма доказана.

Положим теперь в (19)

$$\bar{t} = \bar{p}_{N-1}^* - \bar{p}_{N-1}, \quad B = \Sigma_{N-1}^{*-1},$$

тогда эллипсоид (18) при единичной матрице L может быть записан в эквивалентной форме как

$$\left\{ \bar{p}_{N-1} : |\bar{\lambda}' (\bar{p}_{N-1}^* - \bar{p}_{N-1})| \leq \sqrt{\chi_{\gamma, N-1}^2 (\bar{\lambda}' \Sigma_{N-1}^* \bar{\lambda}) / n, \forall \bar{\lambda}} \right\}.$$

Таким образом, в итоге мы находим, что при $n \rightarrow \infty$

$$\mathbf{P} \left\{ |\bar{\lambda}' (\bar{p}_{N-1}^* - \bar{p}_{N-1})| \leq \sqrt{\chi_{\gamma, N-1}^2 (\bar{\lambda}' \Sigma_{N-1}^* \bar{\lambda}) / n, \forall \bar{\lambda}} \right\} \rightarrow \gamma. \quad (20)$$

Соотношение (20) дает решение нашей задачи, так как оно означает, что для всех линейных комбинаций $\{\bar{\lambda}' \bar{p}_{N-1}\}$ система доверительных интервалов

$$\left\{ (\bar{\lambda}' \bar{p}_{N-1}^* \pm \sqrt{\chi_{\gamma, N-1}^2 (\bar{\lambda}' \Sigma_{N-1}^* \bar{\lambda}) / n}) \right\}$$

имеет асимптотически (при $n \rightarrow \infty$) доверительный уровень γ .

На практике обычно требуется оценивать конечное число заданных линейных комбинаций $\bar{\lambda}'_i \bar{p}_{N-1}$, $i = 1, \dots, m$. В таком случае из (20) находим, что асимптотически

$$\mathbf{P} \{ \bar{\lambda}'_i \bar{p}_{N-1}^* - u_{\gamma i} / \sqrt{n} < \bar{\lambda}' \bar{p}_{N-1} < \bar{\lambda}'_i \bar{p}_{N-1}^* + u_{\gamma i} / \sqrt{n}, i = 1, \dots, m \} \geq \gamma,$$

где

$$u_{\gamma i} = (\chi_{\gamma, N-1}^2 (\bar{\lambda}'_i \Sigma_{N-1}^* \bar{\lambda}_i))^{1/2}, \quad i = 1, \dots, m.$$

В частности, отсюда следует, что система совместных доверительных интервалов для самих долей p_j (всех или части) имеет вид

$$\left\{ (p_j^* \pm \sqrt{\chi_{\gamma, N-1}^2 (1+d) p_j^* (1-p_j^*) / n}) \right\},$$

их доверительный уровень при больших n не меньше γ .

В качестве еще одного примера укажем, что интервалы

$$\left\{ (p_i^* - p_j^* \pm \sqrt{\chi_{\gamma, N-1}^2 (1+d) (p_i^* + p_j^* - (p_i^* - p_j^*)^2) / n}), j < i \right\}$$

образуют асимптотическую систему совместных доверительных интервалов уровня, не меньшего γ , для всех разностей $\{p_i - p_j, j < i\}$.

3.5. Проверка гипотез в схеме серий

В условиях (А) предыдущего раздела, базируясь на теореме 3, можно разработать и теорию (типа хи-квадрат) проверки соответствующих гипотез. В данном случае гипотезы о параметрах модели удобно параметризовать вектором $\bar{p} = (p_1, \dots, p_N)$ и обозначать символом $H(\bar{p})$ соответствующую простую гипотезу.

3.5.1. Прежде всего, рассмотрим вопрос о проверке простой гипотезы $H(\bar{p}^0)$, где вектор $\bar{p}^0$ удовлетворяет требованиям условия (А). Статистику (16), соответствующую этому значению параметра $\bar{p}$, будем далее обозначать $\hat{X}_n^2$. Тогда на основании теоремы 3 можно заключить, что при $n \rightarrow \infty$

$$\mathcal{L}(\hat{X}_n^2 | \bar{p}^0) \rightarrow \chi^2(N-1).$$

Отсюда следует стандартное правило построения критерия согласия хи-квадрат: при заданном уровне значимости ω гипотеза $H(\bar{p}^0)$ отклоняется тогда и только тогда, когда осуществляется событие $\{\hat{X}_n^2 > \chi_{1-\omega, N-1}^2\}$. При таком задании критического множества ошибки первого рода (отклонить $H(\bar{p}^0)$, когда она верна) удовлетворяет при $n \rightarrow \infty$ соотношению

$$\mathbf{P}\{\hat{X}_n^2 > \chi_{1-\omega, N-1}^2 | \bar{p}^0\} \rightarrow \omega.$$

Покажем, что этот критерий в условиях (А) является состоятельным, т.е. для любой фиксированной альтернативы $\bar{p} \neq \bar{p}^0$ его мощность $W_n(\bar{p})$ удовлетворяет при $n \rightarrow \infty$ асимптотическому соотношению

$$W_n(\bar{p}) = \mathbf{P}\{\hat{X}_n^2 > \chi_{1-\omega, N-1}^2 | \bar{p}\} \rightarrow 1. \quad (21)$$

Для этого исследуем асимптотическое поведение среднего и дисперсии статистики $\overset{\circ}{X}_n^2$ при гипотезе $H(\bar{p})$.

Записав представление (см. (16)) в виде

$$\frac{1+d}{n} \overset{\circ}{X}_n^2 = \sum_{j=1}^N \frac{(p_j^* - p_j)^2}{p_j^0} + 2 \sum_{j=1}^N (p_j^* - p_j) \frac{p_j - p_j^0}{p_j^0} + \sum_{j=1}^N \frac{(p_j - p_j^0)^2}{p_j^0}$$

и учитывая формулы

$$\mathbf{E}_{\bar{p}}(p_j^* - p_j) = 0, \quad \mathbf{D}_{\bar{p}} p_j^* = \frac{1+d}{n(1+\alpha^{-1})} p_j(1-p_j),$$

находим, что

$$\mathbf{E}_{\bar{p}} \overset{\circ}{X}_n^2 = \frac{n}{1+d} \sum_{j=1}^N \frac{(p_j - p_j^0)^2}{p_j^0} + \frac{1}{1+\alpha^{-1}} \sum_{j=1}^N \frac{p_j(1-p_j)}{p_j^0} = O(n).$$

На этом же пути, исходя из соотношения

$$\begin{aligned} \frac{(1+d^2)}{n^2} \mathbf{D}_{\bar{p}} \overset{\circ}{X}_n^2 &= \mathbf{E}_{\bar{p}} \left(\sum_{j=1}^N \frac{(p_j^* - p_j)^2}{p_j^0} + 2 \sum_{j=1}^N (p_j^* - p_j) \frac{p_j - p_j^0}{p_j^0} \right)^2 \\ &\quad - \frac{(1+d)^2}{n^2(1+\alpha^{-1})^2} \left(\sum_{j=1}^N \frac{p_j(1-p_j)}{p_j^0} \right)^2, \end{aligned}$$

можно вывести и точную формулу для дисперсии. Однако, ввиду ее громоздкости, мы делать этого не будем, поскольку нам для доказательства состоятельности критерия нужна лишь ее асимптотическая оценка, которая имеет вид

$$\mathbf{D}_{\bar{p}} \overset{\circ}{X}_n^2 = O(n).$$

Эту оценку нетрудно получить, если заметить, что в рассматриваемых условиях для любых целых фиксированных $r_1, \dots, r_N$ с $r = r_1 + \dots + r_N \geq 2$

$$\mathbf{E}_{\bar{p}} \prod_j (p_j^* - p_j)^{r_j} = O\left(\frac{1}{n}\right),$$

что, в свою очередь, следует из формул (см. [1])

$$\mathbf{E}_{\bar{p}} \prod_j (\eta_j(n))_{r_j} = \frac{(n)_{r_j}}{(\alpha + r - 1)_{r_j}} \prod_j (\alpha_j + r_j - 1)_{r_j} = n^r \prod_j p_j^{r_j} \left(1 + O\left(\frac{1}{n}\right)\right).$$

Таким образом, при любой альтернативе $\bar{p} \neq \bar{p}^0$ среднее и дисперсия статистики $\overset{\circ}{X}_n^2$ при $n \rightarrow \infty$ неограниченно возрастают как величины порядка $O(n)$. Отсюда, по неравенству Чебышева при $t = \chi_{1-\omega, N-1}^2$

$$\mathbf{P}\{\overset{\circ}{X}_n^2 \leq t | \bar{p}\} \leq \mathbf{P}\{|\overset{\circ}{X}_n^2 - \mathbf{E}_{\bar{p}} \overset{\circ}{X}_n^2| \geq |\mathbf{E}_{\bar{p}} \overset{\circ}{X}_n^2 - t| | \bar{p}\} \leq \frac{\mathbf{D}_{\bar{p}} \overset{\circ}{X}_n^2}{(\mathbf{E}_{\bar{p}} \overset{\circ}{X}_n^2 - t)^2} = O\left(\frac{1}{n}\right),$$

что и доказывает (21).

Дополним анализ свойств рассматриваемого критерия исследованием асимптотического поведения его мощности при близких альтернативах, в качестве которых возьмем значения параметра вида

$$\bar{p}(n) = \bar{p}^0 + \bar{\beta}/\sqrt{n},$$

где

$$\bar{\beta} = (\beta_1, \dots, \beta_N) \neq \bar{0}, \quad \beta_1 + \dots + \beta_N = 0,$$

— произвольный фиксированный вектор, задающий отклонение от гипотезы $H(\bar{p}^0)$.

Если истинным является значение параметра $\bar{p}(n)$, то, как и при доказательстве теоремы 3 в [1], можно показать, что при $n \rightarrow \infty$ справедливо предельное соотношение $\mathcal{L}(\sqrt{n}(\bar{p}_{N-1}^* - \bar{p}_{N-1}(n)) \mid \bar{p}(n)) \rightarrow \mathcal{N}(\bar{0}, \overset{\circ}{\Sigma}_{N-1})$, где $\overset{\circ}{\Sigma}_{N-1} = \Sigma_{N-1}(\bar{p}^0)$. Но поскольку $\sqrt{n}(\bar{p}_{N-1}^* - \bar{p}_{N-1}^0) = \sqrt{n}(\bar{p}_{N-1}^* - \bar{p}_{N-1}(n)) + \bar{\beta}_{N-1}$, $\bar{\beta}_{N-1} = (\beta_1, \dots, \beta_{N-1})$, отсюда следует, что

$$\mathcal{L}(\sqrt{n}(\bar{p}_{N-1}^* - \bar{p}_{N-1}) \mid \bar{p}(n)) \rightarrow \mathcal{N}(\bar{\beta}_{N-1}, \overset{\circ}{\Sigma}_{N-1}).$$

В свою очередь это означает, что статистика $\overset{\circ}{X}_n^2$ имеет при гипотезе $\bar{p}(n)$ в пределе нецентральное распределение хи-квадрат $\chi^2(N-1; \lambda)$ с параметром нецентральности

$$\lambda = \bar{\beta}'_{N-1} \overset{\circ}{\Sigma}_{N-1}^{-1} \bar{\beta}_{N-1} = \frac{1}{1+d} \sum_{j=1}^N \beta_j^2 / p_j^0. \quad (22)$$

Суммируя изложенное, можно сформулировать следующий результат.

Теорема 4. В условиях (А) критерий согласия для произвольной гипотезы $H(\bar{p}^0)$ при уровне значимости ω , $0 < \omega < 1$, задается асимптотическим критическим множеством $\{\overset{\circ}{X}_n^2 > \chi_{1-\omega, N-1}^2\}$, где статистика $\overset{\circ}{X}_n^2 = X_n^2(\bar{p}^0)$ определена в (16). Этот критерий является состоятельным, и его мощность $W_n(\bar{p})$ при близких альтернативах вида $\bar{p}(n) = \bar{p}^0 + \bar{\beta}/\sqrt{n}$, $\bar{\beta} \neq \bar{0}$, удовлетворяет асимптотическому соотношению

$$W_n(\bar{p}(n)) = \mathbf{P}\{\overset{\circ}{X}_n^2 > \chi_{1-\omega, N-1}^2 \mid \bar{p}(n)\} \rightarrow 1 - F_{N-1}(\chi_{1-\omega, N-1}^2; \lambda),$$

где $F_{N-1}(t; \lambda)$ — функция распределения закона $\chi^2(N-1; \lambda)$ и λ дано в (22).

Замечание 8. Изложенная теория доверительного оценивания и проверки гипотез для МП-модели в схеме серий (при условиях (А)) является аналогом соответствующей теории для полиномиальной схемы и переходит в нее при условии, что $d = n/\alpha \rightarrow 0$, так как в этом случае распределение МП($n; \bar{\alpha}$) переходит в полиномиальное распределение $M(n; \bar{p})$ при $\bar{p} = \bar{\alpha}/\alpha$ (см. замечание к теореме 3 в [1]).

3.5.2. Рассмотрим, наконец, гипотезу однородности H , т. е. гипотезу о том, что два независимых вектора $\bar{\eta}_i(n)$, $i = 1, 2$, имеют одно и то же распределение МП($n; \bar{\alpha}$) с некоторым (неизвестным) параметром $\bar{\alpha}$. По-прежнему предполагая выполненными условия (А) схемы серий, находим, что для статистик $\bar{p}_i^* = \bar{\eta}_i(n)$, $i = 1, 2$, выполняется теорема 3. Следовательно, ввиду их независимости

$$\mathcal{L}(\sqrt{n}(\bar{p}_{1, N-1}^* - \bar{p}_{2, N-1}^*) \mid H) \rightarrow \mathcal{N}(\bar{0}, 2\Sigma_{N-1}).$$

Отсюда, для статистики

$$Y_n^2 = \frac{n}{2}(\bar{p}_{1,N-1}^* - \bar{p}_{2,N-1}^*)' \Sigma_{N-1}^{-1}(\bar{p}_{1,N-1}^* - \bar{p}_{2,N-1}^*) = \frac{n}{2(1+d)} \sum_{j=1}^N \frac{(p_{1j}^* - p_{2j}^*)^2}{p_j} \quad (23)$$

получаем соотношение

$$\mathcal{L}(Y_n^2 | H) \rightarrow \chi^2(N-1). \quad (24)$$

Этим еще не решается вопрос о построении критерия, так как в выражение (23) входят неизвестные параметры p_j . Однако предельное соотношение (24) останется в силе, если параметры p_j заменить их состоятельными оценками $(p_{1j}^* + p_{2j}^*)/2$, т. е. если в качестве тестовой статистики взять статистику

$$\dot{Y}_n^2 = \frac{n}{1+d} \sum_{j=1}^N \frac{(p_{1j}^* - p_{2j}^*)^2}{p_{1j}^* + p_{2j}^*}.$$

В итоге получаем, что

$$\mathbf{P}\{\dot{Y}_n^2 > \chi_{1-\omega, N-1}^2\} \rightarrow \omega,$$

т. е. критическое множество $\{\dot{Y}_n^2 > \chi_{1-\omega, N-1}^2\}$ задает искомый критерий согласия для гипотезы однородности с асимптотическим уровнем значимости ω .

Построенный критерий может быть использован для проверки гипотезы о том, что две МП-урны имеют одинаковый состав (одинаковые численности шаров соответствующих цветов).

3.6. Критерий согласия в МП-модели для случая многих выборов

В этом разделе мы кратко обсудим классический подход к задаче проверки гипотез в МП-модели, когда параметры модели предполагаются фиксированными, а число наблюдений над основной переменной (достаточной статистикой модели) достаточно велико.

Итак, пусть наблюдаемая переменная $\bar{\eta}(n)$ имеет распределение МП($n; \alpha \bar{p}$), где $\alpha > 0$ известно, а $\bar{p} \in \mathcal{P}_N$ является неизвестным параметром, относительно которого требуется проверить простую гипотезу $H_0: \bar{p} = \bar{p}^0$. Предположим, что имеется K независимых наблюдений $\bar{\eta}_i$, $i = 1, \dots, K$, над $\bar{\eta}(n)$ и будем считать, что $K \rightarrow \infty$. Таким образом, мы находимся в условиях классической многовыборочной схемы, в рамках которой несложно рассчитать стандартный критерий хи-квадрат для проверки гипотезы H_0 .

Именно, во-первых, как показано в разделе 1.3, статистика

$$\bar{p}^*(K) = \frac{1}{nK} \sum_{i=1}^K \bar{\eta}_i$$

несмещенно и состоятельно (при $K \rightarrow \infty$) оценивает параметр $\bar{p}^0$, если справедлива гипотеза H_0 , и во-вторых, по центральной предельной теореме эта статистика является при H_0 асимптотически нормальной со средним $\bar{p}^0$ и дисперсионной матрицей, указанной в разделе 1.3 (где надо заменить символ $\bar{\theta}$ на $\bar{p}$).

Отсюда следует, что подходящей тестовой статистикой в данном случае является статистика

$$X^2(K) = \frac{K}{e(\alpha, n)} \sum_{j=1}^N \frac{(p_j^*(K) - p_j^0)^2}{p_j^0}, \quad (25)$$

имеющая при гипотезе H_0 предельное (при $K \rightarrow \infty$) распределение хи-квадрат $\chi^2(N-1)$ (сравни со статистикой $\hat{X}_n^2$ предыдущего раздела). В итоге находим, что критерий согласия для гипотезы H_0 асимптотически задается критическим множеством $\{X^2(K) > \chi_{1-\omega, N-1}^2\}$, где ω — заданный уровень значимости.

Описанный подход уже рассматривался в литературе (см. последнюю работу [4] на эту тему, где имеется и соответствующая библиография). Отметим лишь, что в [4] рассматривается класс тестовых статистик вида (в наших обозначениях)

$$T_\varphi = \frac{2K}{e(\alpha, n)\varphi''(1)} \sum_{j=1}^N p_j^0 \varphi(p_j^*(K)/p_j^0),$$

где функция φ удовлетворяет условиям регулярности, обеспечивающим сходимость

$$\mathcal{L}(T_\varphi | H_0) \rightarrow \chi^2(N-1)$$

при $K \rightarrow \infty$, и находятся дополнительные условия на φ , при которых первые два момента статистики T_φ совпадают с аналогичными моментами предельного распределения вплоть до членов порядка K^{-1} .

Отметим также близкую по тематике работу [5], в которой рассматриваются некоторые вопросы предельных (в различных условиях) распределений стандартных тестовых статистик в задаче различения равновероятной полиномиальной схемы и смеси всех полиномиальных схем (с N исходами) по симметрической мере Дирихле $\mathcal{D}(\nu, \dots, \nu)$, т.е. гипотез $H_0(p^0)$ при $\bar{p}^0 = (1/N, \dots, 1/N)$ и $H_1(\bar{\alpha})$ при $\bar{\alpha} = (\nu, \dots, \nu)$ в наших обозначениях.

Список литературы

1. Ивченко Г. И., Медведев Ю. И., Об урновой схеме Маркова–Пойа: от 1917 до наших дней. *Обозрение приклад. и промышл. матем.* (1996) **3**, №4, 484–511.
2. Рао С. Р., *Линейные статистические методы и их применения*. Наука, Москва, 1968.
3. Шеффе Х., *Дисперсионный анализ*. Физматгиз, Москва, 1963.
4. Menendez M. L., Morales D., Pardo P., Vajda I., About divergence-based goodness-of-fit tests in the Dirichlet-multinomial model. *Commun. Statist. Theory and Math.* (1996) **25**, №5, 1119–1133.
5. Амбросимов А. С., Тимашев А. Н., О рандомизированном подходе к построению критериев проверки равновероятности исходов полиномиальной схемы. *Дискретная математика* (1997) **9**, №3, 43–51.

Статья поступила 22.06.1999.