Math-Net.Ru

Общероссийский математический портал

А. С. Амбросимов, А. Н. Тимашёв, О рандомизированном подходе к построению критериев проверки равновероятности исходов полиномиальной схемы, *Дискрет. матем.*, 1997, том 9, выпуск 3, 43–51

DOI: <https://doi.org/10.4213/dm480>

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: 18.234.197.8

26 апреля 2023 г., 15:22:14

О рандомизированном подходе к построению критериев проверки равновероятности исходов полиномиальной схемы

© 1997 г. А. С. Амбросимов, А. Н. Тимашёв

Рассматривается наиболее мощный критерий и критерий χ^2 проверки равновероятности исходов полиномиальной схемы в случае, когда альтернативная гипотеза получается путем усреднения всех неравновероятных полиномиальных схем с N исходами по вероятностной мере, порожденной симметричным $(N-1)$ -мерным распределением Дирихле. Получены предельные теоремы о сходимости распределений статистик этих критериев при обеих гипотезах.

Пусть наблюдается выборка $\gamma_1, \dots, \gamma_n$ объема n из полиномиального распределения с N исходами. Требуется выяснить, является она выборкой из равновероятного полиномиального распределения или нет.

Пусть гипотеза H_0 состоит в том, что выборка $\gamma_1, \dots, \gamma_n$ соответствует полиномиальной схеме с N исходами, имеющими одинаковые вероятности, равные $1/N$. В качестве альтернативной гипотезы H_1 , как правило, выбирают гипотезу, соответствующую полиномиальной схеме с N неравновероятными исходами. Однако на практике гипотеза H_1 обычно неизвестна. В этой ситуации возникает вопрос: как проверить, справедлива гипотеза H_0 или нет? Один из возможных ответов на этот вопрос состоит в следующем. Предлагается рассматривать наиболее мощный критерий и критерий χ^2 проверки гипотезы H_1 при альтернативе H_0 в случае, когда гипотеза H_1 получается путем усреднения всех неравновероятных полиномиальных схем с N исходами по вероятностной мере, порожденной симметричным $(N-1)$ -мерным распределением Дирихле с параметром $\nu > 0$, плотность которого имеет вид [1, с. 192]

$$\begin{aligned} p(x_1, \dots, x_N) &= 0, & (x_1, \dots, x_{N-1}) &\notin \Omega_{N-1}, \\ p(x_1, \dots, x_N) &= \frac{\Gamma(N\nu)}{(\Gamma(\nu))^N} \prod_{j=1}^N x_j^{\nu-1}, & (x_1, \dots, x_{N-1}) &\in \Omega_{N-1}. \end{aligned} \quad (1)$$

В равенствах (1) предполагается, что $x_1 + \dots + x_N = 1$, Γ — гамма-функция Эйлера,

$$\Omega_{N-1} = \left\{ (x_1, \dots, x_{N-1}) \mid x_j \geq 0, j = 1, \dots, N-1, \sum_{j=1}^{N-1} x_j \leq 1 \right\}$$

— $(N - 1)$ -мерный симплекс в пространстве $\mathbf{R}^{N-1}$. Другими словами, гипотеза H_1 состоит в том, что выборка $\gamma_1, \dots, \gamma_n$ соответствует рандомизированной полиномиальной схеме со случайным вектором вероятностей исходов, равным $(\xi_1, \dots, \xi_{N-1}, \xi_N)$, $\xi_1 + \dots + \xi_N = 1$, который имеет симметричное распределение Дирихле с параметром $\nu > 0$.

Известно [1, с. 192], что математическое ожидание каждой из случайных величин ξ_j равно

$$\mathbf{E}\xi_j = 1/N, \quad j = 1, 2, \dots, N. \quad (2)$$

Дисперсия каждой из случайных величин ξ_j определяется равенствами

$$\mathbf{D}\xi_j = \frac{N-1}{N^2(N\nu+1)}, \quad j = 1, 2, \dots, N. \quad (3)$$

Из формул (2) и (3), в частности, следует, что при фиксированном $N \geq 2$ и $\nu \rightarrow \infty$ каждая из случайных величин $\xi_1, \dots, \xi_N$ имеет в пределе вырожденное в точке $1/N$ распределение.

Известно также (см. [1]), что все случайные величины $\xi_1, \dots, \xi_N$ одинаково распределены и имеют бета-распределение с плотностью

$$p_1(x) = \frac{\Gamma(N\nu)}{\Gamma(\nu)\Gamma((N-1)\nu)} x^{\nu-1} (1-x)^{(N-1)\nu-1}, \quad (4)$$

где $x \in (0, 1)$.

Заметим, что если $\nu = 1$, то случайный вектор $(\xi_1, \dots, \xi_{N-1})$ имеет равномерное распределение на симплексе Ω_{N-1} с плотностью, равной $(N-1)!$.

В этом случае из (4) получаем

$$p_1(x) = (N-1)(1-x)^{N-2}, \quad x \in [0, 1].$$

Статистика ξ наиболее мощного критерия проверки гипотезы H_1 при альтернативе H_0 имеет вид

$$\xi = \int \dots \int_{\Omega_{N-1}} \pi(x_1, \dots, x_N) p(x_1, \dots, x_N) dx_1 \dots dx_{N-1}, \quad (5)$$

где

$$x_1 + \dots + x_N = 1, \quad \pi(x_1, \dots, x_N) = N^n \prod_{j=1}^N x_j^{h_j}, \quad (6)$$

h_j — частота появления j -го исхода в выборке $\gamma_1, \dots, \gamma_n$, $j = 1, 2, \dots, N$.

Подставляя в (5) значения $p(x_1, \dots, x_N)$ из (1) и $\pi(x_1, \dots, x_N)$ из (6) и вычисляя получившийся интеграл Дирихле, находим, что

$$\xi = N^n \frac{\Gamma(N\nu)}{(\Gamma(\nu))^N \Gamma(N\nu+n)} \prod_{j=1}^N \Gamma(\nu + h_j). \quad (7)$$

Если, в частности, $\nu = 1$, то из (7) следует, что

$$\xi = \frac{N^n(N-1)!}{(N+n-1)!} \prod_{j=1}^N (h_j!). \quad (8)$$

Формула (8) появилась в работе авторов [4].

Теорема 1. Пусть при $n \rightarrow \infty$ параметры $N \geq 2$ и $\nu > 0$ фиксированы. Если справедлива гипотеза H_1 , то

$$\frac{1}{n} \ln \xi \rightarrow \ln N - H(\xi_1, \dots, \xi_{N-1}, \xi_N)$$

по распределению, где

$$\xi_N = 1 - \sum_{j=1}^{N-1} \xi_j, \quad H(x_1, \dots, x_N) = - \sum_{j=1}^N x_j \ln x_j.$$

Из теоремы 1 следует, что при ее условиях и $x \in [0, \ln N]$

$$\lim_{n \rightarrow \infty} \mathbf{P}\{\ln N - n^{-1} \ln \xi < x\} = \frac{\Gamma(N\nu)}{(\Gamma(\nu))^N} \int_S \cdots \int \prod_{j=1}^N x_j^{\nu-1} dx_1 \dots dx_{N-1},$$

где область интегрирования S определяется условиями

$$(x_1, \dots, x_{N-1}) \in \Omega_{N-1}, \\ H(x_1, \dots, x_N) < x, \quad x_1 + \dots + x_N = 1.$$

Если $\lambda(S)$ — классическая мера Лебега области S в пространстве $\mathbf{R}^{N-1}$, то в случае $\nu = 1$ получаем, что

$$\lim_{n \rightarrow \infty} \mathbf{P}\{\ln N - n^{-1} \ln \xi < x\} = (N-1)! \lambda(S).$$

Для обоснования теоремы 1 достаточно заметить следующее. Известно [7], что при условиях этой теоремы

$$\frac{h_j}{n} \rightarrow \xi_j, \quad j = 1, 2, \dots, N,$$

почти наверное, и поэтому при $n \rightarrow \infty$

$$\sum_{j=1}^N \frac{h_j}{n} \ln \frac{h_j}{n} \rightarrow -H(\xi_1, \dots, \xi_{N-1}, \xi_N) \quad (9)$$

по распределению, где, по определению $0 \ln 0 = 0$ и

$$\xi_N = 1 - \sum_{j=1}^{N-1} \xi_j.$$

Логарифмируя обе части равенства (7) и используя формулу Стирлинга для гамма-функции, путем элементарных преобразований нетрудно убедиться в том, что имеют место равенства

$$\frac{1}{n} \ln \xi - \ln N = \frac{1}{n} \sum_{j=1}^N \Gamma(\nu + h_j) - \ln n + 1 + o(1) = \sum_{j=1}^N \frac{h_j}{n} \ln \frac{h_j}{n} + o(1),$$

где $o(1) \rightarrow 0$ при $n \rightarrow \infty$ с вероятностью 1. Из этих равенств с учетом (9) получаем утверждение теоремы 1.

Заметим, что, как показано в [7], при усреднении полиномиальной схемы с N исходами по вероятностной мере, порожденной симметричным $(N-1)$ -мерным распределением Дирихле с параметром $\nu > 0$, получается перестановочная последовательность, частоты которой $h_1, \dots, h_N$ для первых n наблюдений $\gamma_1, \dots, \gamma_n$, $h_1 + \dots + h_N = n$, имеют распределение Маркова–Пойа, определяемое равенствами

$$\mathbf{P}\{h_j = r_j, 1 \leq j \leq N\} = \frac{n!}{r_1! \dots r_N!} \frac{\Gamma(N\nu)}{(\Gamma(\nu))^N \Gamma(N\nu + n)} \prod_{j=1}^N \Gamma(\nu + r_j). \quad (10)$$

В равенствах (10) $r_1, \dots, r_N$ — целые неотрицательные числа, для которых $r_1 + \dots + r_N = n$.

Распределение вида (10) рассматривалось в [2], где были получены локальные предельные теоремы для случайных величин μ_r , равных числу появлений значения r среди частот $h_1, \dots, h_N$, $r = 0, 1, \dots$, в случае, когда $n, N \rightarrow \infty$ (более подробную библиографию по распределениям вида (10) см. в [7]).

Если, в частности, $\nu = 1$, то из (10) следует, что

$$\mathbf{P}\{h_j = r_j, 1 \leq j \leq N\} = \binom{n + N - 1}{n}^{-1}$$

(так называемая статистика Бозе–Эйнштейна [3, с. 52]).

Известно также (см. [7]), что при фиксированных значениях n и N и $\nu \rightarrow \infty$

$$\mathbf{P}\{h_j = r_j, 1 \leq j \leq N\} = \frac{n! N^{-n}}{r_1! \dots r_N!} (1 + o(1)). \quad (11)$$

Формула (11) свидетельствует о близости гипотез H_0 и H_1 при больших значениях параметра ν .

Если же при фиксированных значениях n и N $\nu \rightarrow 0+$, то из (10) следует, что

$$\mathbf{P}\{h_j = r_j, 1 \leq j \leq N\} = o(1),$$

если $r_j < n$, $j = 1, 2, \dots, N$, и

$$\mathbf{P}\{h_j = r_j, 1 \leq j \leq N\} = \frac{1}{N} (1 + o(1)),$$

если найдется значение j_0 , такое, что $r_{j_0} = n$ (при этом $r_j = 0$ при $j \neq j_0$).

Статистика ξ наиболее мощного критерия проверки гипотезы H_1 при альтернативе H_0 , определенная в (7), при больших значениях n и N вычисляется достаточно

сложно. Поэтому, в частности, представляет интерес критерий проверки гипотезы H_1 при альтернативе H_0 , основанный на статистике

$$\chi^2 = \frac{N}{n} \sum_{j=1}^N (h_j - n/N)^2. \quad (12)$$

Хорошо известно, что статистика χ^2 при $n \rightarrow \infty$ и фиксированном $N \geq 2$ в случае, когда справедлива гипотеза H_0 имеет распределение χ^2 с $N - 1$ степенями свободы.

Теорема 2. При условиях теоремы 1 и $x \in [0, 1 - 1/N]$

$$\lim_{n \rightarrow \infty} \mathbf{P}\{\chi^2/(nN) < x\} = \frac{\Gamma(N\nu)}{(\Gamma(\nu))^N} \int \dots \int \prod_{j=1}^N x_j^{\nu-1} dx_1 \dots dx_{N-1}, \quad (13)$$

где $x_1 + \dots + x_N = 1$ и область интегрирования определяется условиями

$$(x_1, \dots, x_{N-1}) \in \Omega_{N-1}, \\ \sum_{j=1}^N \left(x_j - \frac{1}{N}\right)^2 < x.$$

Для доказательства теоремы 2 с учетом сказанного выше достаточно заметить, что

$$\frac{\chi^2}{nN} = \sum_{j=1}^N \left(\frac{h_j}{n} - \frac{1}{N}\right)^2 \rightarrow \sum_{j=1}^N \left(\xi_j - \frac{1}{N}\right)^2$$

по распределению, и поэтому предельное значение вероятности $\mathbf{P}\{\chi^2/(nN) < x\}$ равно вероятности

$$\mathbf{P}\left\{\sum_{j=1}^N \left(\xi_j - \frac{1}{N}\right)^2 < x\right\},$$

которая совпадает с выражением в правой части формулы (13).

Следствие 1. При условиях теоремы 2 и $N = 2$

$$\lim_{n \rightarrow \infty} \mathbf{P}\{\chi^2/(2n) < x\} = \frac{\Gamma(2\nu)}{(\Gamma(\nu))^2} \sqrt{2x} \int_0^1 \left(\frac{1}{4} - \frac{x}{2} + 2xz - 2xz^2\right)^{\nu-1} dz,$$

где $x \in [0, 1/2]$.

В частности, для случая $\nu = 1$

$$\lim_{n \rightarrow \infty} \mathbf{P}\{\chi^2/(2n) < x\} = \sqrt{2x}$$

при $0 \leq x \leq 1/2$.

Следствие 2. При условиях теоремы 2 и $\nu = 1$

$$\lim_{n \rightarrow \infty} \mathbf{P}\{\chi^2/(nN) < x\} = (N-1)! \lambda(B_{N-1}),$$

где λ — классическая мера Лебега в пространстве $\mathbf{R}^{N-1}$ и B_{N-1} — множество точек $(x_1, \dots, x_{N-1}) \in \Omega_{N-1}$, для которых

$$\sum_{j=1}^N \left(x_j - \frac{1}{N}\right)^2 < x, \quad x_1 + \dots + x_N = 1.$$

В частности, для случая $N = 3$ получаем, что $\lambda(B_2) = \pi x / \sqrt{3}$, если $0 \leq x \leq 1/6$, и

$$\lambda(B_2) = \frac{\pi x}{\sqrt{3}} - \frac{x\sqrt{3}}{2} \arccos \frac{1-3x}{3x} + \frac{1}{2} \sqrt{2x - \frac{1}{3}},$$

если $1/6 \leq x \leq 2/3$.

Заметим, что при фиксированных значениях n и N и $\nu \rightarrow 0+$ в случаях, когда справедлива гипотеза H_1 , распределение случайной величины χ^2 сходится к вырожденному в точке $n(N-1)$ распределению.

Рассмотрим теперь вопрос о распределении статистики $\ln \xi$ при условии справедливости гипотезы H_0 . Из формулы (7) получаем, что

$$\ln \xi = n \ln N + \ln \Gamma(N\nu) - N \ln \Gamma(\nu) - \ln \Gamma(N\nu + n) + \sum_{j=1}^N \ln \Gamma(\nu + h_j). \quad (14)$$

Далее, исходя из (14), устанавливается сходимость по распределению при $n \rightarrow \infty$ статистики $\ln \xi$ (соответствующим образом центрированной и нормированной) при условии справедливости гипотезы H_0 к закону χ^2 с $N-1$ степенями свободы (если $N \geq 2$ фиксировано) и к стандартному нормальному закону (если $N \rightarrow \infty$ так, что $N \ln N = o(n^{1/3})$). Эти результаты в случае $\nu = 1$ были сформулированы авторами в [4]. Заметим, что утверждения теорем 1 и 2 сформулированы в [6].

Теорема 3. Пусть $n \rightarrow \infty$ и $\nu = \nu(n)$ — произвольная положительная функция от n , $N \geq 2$ фиксировано. Если справедлива гипотеза H_0 , то статистика

$$2(1 + N\nu/n) \left(\ln \xi - \ln \Gamma(N\nu) + N \ln \Gamma(\nu) - \frac{N-1}{2} \ln(2\pi) \right. \\ \left. + N \left(\nu - \frac{1}{2} \right) \ln N + \frac{N-1}{2} \ln(N\nu + n) \right)$$

имеет в пределе χ^2 -распределение с $N-1$ степенями свободы.

Доказательство. Согласно равенству (14) исследуем предельное распределение случайной величины

$$\eta = \sum_{j=1}^N \ln \Gamma(\nu + h_j).$$

Используя формулу Стирлинга для гамма-функции, получаем, что

$$\eta = \sum_{j=1}^N \ln \left(\sqrt{\frac{2\pi}{\nu + h_j}} \left(\frac{\nu + h_j}{e} \right)^{\nu + h_j} \right) + o(1).$$

Здесь и далее $o(1) \rightarrow 0$ по вероятности при $n \rightarrow \infty$.

Полагаем

$$\tau_j = \frac{h_j - n/N}{\sqrt{n/N}}, \quad j = 1, 2, \dots, N. \quad (15)$$

Тогда

$$\begin{aligned} \sum_{j=1}^N \tau_j &= 0, \\ \eta &= \frac{N}{2} \ln(2\pi) - \nu N - n \\ &\quad + \sum_{j=1}^N (\nu + n/N + \tau_j \sqrt{n/N - 1/2}) \ln(\nu + n/N + \tau_j \sqrt{n/N}) + o(1) \\ &= \frac{N}{2} \ln(2\pi) - \nu N - n + \frac{n}{2(\nu N + n)} \sum_{j=1}^N \tau_j^2 (1 + o(1)) \\ &\quad + (\nu N + n - N/2) \ln(\nu + n/N) + o(1). \end{aligned}$$

По теореме Пирсона статистика

$$\sum_{j=1}^N \tau_j^2$$

в пределе при $n \rightarrow \infty$ имеет распределение χ^2 с $N - 1$ степенями свободы. Отсюда и из равенства (14) получаем, что случайная величина

$$\begin{aligned} \frac{2(\nu N + n)}{n} \left(\ln \xi - n \ln N - \ln \Gamma(N\nu) + N \ln \Gamma(\nu) + \ln \Gamma(N\nu + n) \right. \\ \left. - \frac{N}{2} \ln(2\pi) + \nu N + n - (\nu N + n - N/2) \ln(\nu + n/N) \right) \end{aligned}$$

в пределе при $n \rightarrow \infty$ имеет распределение χ^2 с $N - 1$ степенями свободы. Остается применить формулу Стирлинга для $\ln \Gamma(N\nu + n)$. Теорема 3 доказана.

Теорема 4. Пусть $n, N \rightarrow \infty$ так, что $N \ln N = o(n^{1/3})$, $\nu = \nu(n, N)$ — произвольная положительная функция. Если справедлива гипотеза H_0 , то статистика

$$\begin{aligned} \sqrt{2/N}(1 + N\nu/n) \left(\ln \xi - \ln \Gamma(N\nu) + N \ln \Gamma(\nu) - \frac{N-1}{2} \ln(2\pi) \right. \\ \left. + N(\nu - 1/2) \ln N + \frac{N-1}{2} \ln(N\nu + n) \right) - \sqrt{N/2} \end{aligned}$$

распределена в пределе по стандартному нормальному закону.

Доказательство. Исследуем статистику

$$\eta = \sum_{j=1}^N \ln \Gamma(\nu + h_j).$$

Из формулы Стирлинга для гамма-функции с учетом (15) получаем, что

$$\begin{aligned} \eta = & \frac{N}{2} \ln(2\pi) - \nu N - n \\ & + \sum_{j=1}^N (\nu + n/N + \tau_j \sqrt{n/N} - 1/2) \ln(\nu + n/N + \tau_j + \sqrt{n/N}) \\ & + \sum_{j=1}^N \frac{\theta_j}{12(\nu + h_j)}, \end{aligned}$$

где $|\theta_j| < 1$, $j = 1, 2, \dots, N$.

Из леммы 2 из [5] получаем оценку

$$\mathbf{P}\{|h_j - n/N| \geq \sqrt{Mn(\ln N)/N}\} = O(N^{-M/2}), \quad (16)$$

при $j = 1, 2, \dots, N$ и любом фиксированном $M > 0$.

Из этих оценок и условия $N \ln N = o(n^{1/3})$ теоремы 4 вытекает сходимость по вероятности к нулю суммы

$$\sum_{j=1}^N \frac{\theta_j}{12(\nu + h_j)}.$$

Далее, справедливы равенства

$$\begin{aligned} \eta = & \frac{N}{2} \ln(2\pi) - \nu N - n \\ & + \sum_{j=1}^N (\nu + n/N - 1/2 + \tau_j \sqrt{n/N}) \left(\ln(\nu + n/N) + \ln \left(1 + \frac{\tau_j \sqrt{n/N}}{\nu + n/N} \right) \right) + o(1) \\ = & \frac{N}{2} \ln(2\pi) - \nu N - n + (N\nu + n - N/2) \ln(\nu + n/N) \\ & + \sum_{j=1}^N (\nu + n/N - 1/2 + \tau_j \sqrt{n/N}) \left(\frac{\tau_j \sqrt{n/N}}{\nu + n/N} - \frac{1}{2} \tau_j^2 \frac{n/N}{(\nu + n/N)^2} \right. \\ & \left. + O \left(\frac{\tau_j^3}{(\nu + n/N)^3} \left(\frac{n}{N} \right)^{3/2} \right) \right) + o(1). \end{aligned}$$

Из оценки (16) вытекает, что

$$\sum_{j=1}^N (h_j - 1/2 + \nu) O \left(\frac{\tau_j^3}{(\nu + n/N)^3} \left(\frac{n}{N} \right)^{3/2} \right) \rightarrow 0$$

по вероятности при $n, N \rightarrow \infty$.

Таким образом,

$$\eta = \frac{N}{2} \ln(2\pi) - \nu N - n + (\nu N + n - N/2) \ln(\nu + n/N) + \frac{1}{2} \sum_{j=1}^N \tau_j^2 + o(1).$$

В рассматриваемых условиях статистика

$$\sum_{j=1}^N \tau_j^2$$

при естественном нормировании и центрировании имеет в пределе при $n, N \rightarrow \infty$ стандартное нормальное распределение. Тем самым теорема 4 доказана.

Список литературы

1. Уилкс С. *Математическая статистика*. Наука, Москва, 1967.
2. Chen Wen Chen. Some local limit theorems in the symmetric Dirichlet-multinomial urn models. *Ann. Inst. Statist. Math.* (1981) **33**, №3, 405–415.
3. Феллер В. *Введение в теорию вероятностей и ее приложения*, т. 1. Мир, Москва, 1964.
4. Амбросимов А. С., Тимашев А. Н. Предельные теоремы для статистики наиболее мощного критерия, связанного с мультиномиальным распределением. В кн.: *“Тезисы докладов третьей международной конференции по теории вероятностей и математической статистики”*, т. 1. Вильнюс, 1981, с. 7–8.
5. Амбросимов А. С. О состоятельности метода максимального правдоподобия в задаче упорядочения исходов полиномиальной схемы в соответствии с их вероятностями. *Дискретная математика* (1995) **7**, №2, 118–130.
6. Тимашев А. Н. О стохастическом подходе к построению критериев проверки равновероятности исходов полиномиальной схемы. В кн.: *“Тезисы докладов второй Всероссийской школы-коллоквиума по стохастическим методам.”* ТВП, Москва, 1995, с. 134–135.
7. Ивченко Г. И., Медведев Ю. И. Об урновой модели Маркова–Поя: от 1917 г. до наших дней. *Обзорные прикладной и промышленной математики* (1996) **3**, №2, 484–511.

Статья поступила 25.10.1995.

Переработанный вариант поступил 05.07.1996.