Math-Net.Ru

Общероссийский математический портал

М. И. Тихомирова, Некоторые характеристики зависимостей в дискретных случайных последовательностях, *Дискрет. матем.*, 2008, том 20, выпуск 1, 151–158

DOI: <https://doi.org/10.4213/dm998>

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: 52.23.180.231

26 апреля 2023 г., 06:37:59

Некоторые характеристики зависимостей в дискретных случайных последовательностях

© 2008 г. М. И. Тихомирова

Для некоторых статистик типа χ^2 , предназначенных для выявления зависимости между случайными величинами исследуемых последовательностей, получены предельные теоремы в случае, когда объем выборки стремится к бесконечности.

1. Введение

Пусть

$$x_1, x_2, \dots, x_t, \dots, \quad x_t \in \{1, 2, \dots, N\}, \quad (1)$$

— последовательность случайных величин. По гипотезе H_0 случайные величины в (1) независимы и

$$\mathbf{P}\{x_t = k\} = p_k > 0, \quad k = 1, 2, \dots, N, \quad \sum_{k=1}^N p_k = 1, \quad t = 1, 2, \dots \quad (2)$$

Для выявления зависимостей между случайными величинами последовательности (1) можно использовать частоты семейств несплошных цепочек следующего вида:

$$x_t(r, m_1, m_2) = (x_t, \dots, x_{t+m_1-1}, x_{t+m_1+r}, \dots, x_{t+m_1+m_2+r-1}), \quad r = 0, 1, 2, \dots \quad (3)$$

Частоты этих цепочек обозначим через $v_I(n) = v_I(n, r, m_1, m_2)$:

$$v_I(n) = v_I(n, r, m_1, m_2) = \sum_{t=1}^n \chi_I(t), \quad I = (i_1, i_2, \dots, i_M), \quad (4)$$

где

$$M = m_1 + m_2, \quad \chi_I(t) = \chi(x_t(r, m_1, m_2) = I), \quad t = 1, 2, \dots, n,$$

$\chi(A)$ — индикатор события A . Рассмотрим нормированные и центрированные частоты

$$\xi_I(n) = \xi_I(n, r, m_1, m_2) = \frac{v_I(n) - na_I^{(0)}(r)}{\sqrt{na_I^{(0)}(r)}}, \quad (5)$$

где

$$a_I^{(0)}(r) = p_{i_1} p_{i_2} \dots p_{i_M}.$$

При гипотезе H_0 эти частоты асимптотически нормальны при $n \rightarrow \infty$ с нулевым вектором математических ожиданий и ковариационной матрицей

$$(B_M)_{I,J} = \lim_{n \rightarrow \infty} \text{cov}(\xi_I(n), \xi_J(n)). \quad (6)$$

Здесь и далее (I, J) -й элемент матрицы C будем обозначать через $(C)_{I,J}$. Используя для $v_I(n)$ представление (4), можно показать, что ковариационная матрица B_M имеет следующий вид

$$B_M = E_M - (2s - 1)F_M + \sum_{k=1}^{s-1} (B_M(k) + B'_M(k)), \quad (7)$$

где E_M — единичная $N^M \times N^M$ матрица,

$$(F_M)_{I,J} = \sqrt{a_I^{(0)} a_J^{(0)}}, \quad (B'_M(k))_{I,J} = (B_M(k))_{J,I},$$

$$(B_M(k))_{I,J} = \mathbf{E} \chi_I(1) \chi_J(k+1) / \sqrt{a_I^{(0)} a_J^{(0)}}, \quad (8)$$

$s = m_1 + m_2 + r$ — длина наименьшей сплошной s -цепочки, содержащей несплошную цепочку (3). Статистику $X_r(n) = X(n, r, m_1, m_2)$ определим равенством

$$X_r(n) = \sum_I \xi_I^2(n) - \sum_{\tilde{I}} \tilde{\xi}_{\tilde{I}}^2(n), \quad (9)$$

где

$$\tilde{\xi}_{\tilde{I}}(n) = \frac{v_{\tilde{I}}(n) - n a_{\tilde{I}}^{(0)}(r)}{\sqrt{n a_{\tilde{I}}^{(0)}(r)}}, \quad \tilde{I} = (i_1, i_2, \dots, i_{M-1}),$$

$$v_{\tilde{I}}(n) = \sum_{i_M=1}^N v_I(n), \quad a_{\tilde{I}}^{(0)}(r) = p_{i_1} p_{i_2} \dots p_{i_{M-1}}, \quad I = (\tilde{I}, i_M).$$

Эту статистику удобно рассматривать как квадратичную форму от величин (5)

$$X_r(n) = \sum_{I,J} (A_M)_{I,J} \xi_I(n) \xi_J(n). \quad (10)$$

Коэффициенты $(A_M)_{I,J}$ определяются равенствами

$$A_M = E_M - G_M, \quad (G_M)_{I,J} = \delta_{i_1 j_1} \dots \delta_{i_{M-1} j_{M-1}} \sqrt{p_{i_M} p_{j_M}}. \quad (11)$$

Предельное распределение статистики (9) при гипотезе H_0 при $n \rightarrow \infty$ совпадает с распределением квадратичной формы

$$X_r = \sum_{I,J} (A_M)_{I,J} \xi_I \xi_J, \quad I = (i_1, \dots, i_M), \quad J = (j_1, \dots, j_M), \quad (12)$$

где величины ξ_I имеют нормальное распределение с нулевым вектором математических ожиданий и ковариационной матрицей (7). Это распределение является χ^2 -распределением с $N^M - N^{M-1}$ степенями свободы (см. теорему 2 в [2]).

Пусть $\bar{p}_k^{(l)}$, $k = 1, \dots, N$, – оценки вероятностей p_k , определяемые равенствами

$$\bar{p}_k^{(l)} = \frac{1}{n} \sum v_I(n), \quad l = 1, \dots, M, \quad (13)$$

где суммирование проводится по всем $I = (i_1, \dots, i_M)$, для которых $i_u \in \{1, 2, \dots, N\}$, $u \in \{u: 1 \leq u \leq M, u \neq l\}$, $i_{u_l} = k$. Статистики, получающиеся из статистики (9) заменой величин $a_I^{(0)}(r)$, $a_{\bar{I}}^{(0)}(r)$ величинами

$$\bar{a}_I = \bar{p}_{i_1}^{(1)} \dots \bar{p}_{i_M}^{(M)}, \quad \bar{a}_{\bar{I}} = \bar{p}_{i_1}^{(1)} \dots \bar{p}_{i_{M-1}}^{(M-1)}, \quad (14)$$

обозначим

$$\bar{X}_r(n) = \bar{X}_r(n, r, m_1, m_2), \quad r = 0, 1, 2, \dots \quad (15)$$

В данной работе рассматриваются предельные распределения статистики $X_r(n)$ при конкурирующих гипотезах. Предельные распределения $\bar{X}_r(n)$ при $r = 0$ (сплошные цепочки) приведены в теореме 3 работы [2]. Предельные теоремы в случае несплошных цепочек в работе [2] получены только для некоторых цепочек длины 3. В данной работе предельные распределения статистик $\bar{X}_r(n)$ получены при гипотезе H_0 в двух случаях, когда $m_1 = m_2 = m$ и $m_1 = m$, $m_2 = 1$ при любых r .

Таким образом, получены предельные распределения двух семейств статистик, которые могут быть использованы для обнаружения зависимости случайных величин исследуемой последовательности.

2. Предельные распределения статистик $X_r(n)$ при конкурирующих гипотезах

Будем предполагать (гипотеза H_1), что последовательность (1) является стационарной эргодической цепью Маркова с матрицей вероятностей переходов $(p_{i,j})$ и стационарным распределением $(p_1, \dots, p_N)$. Центрируем частоты $v_I(n)$ своими математическими ожиданиями при гипотезе H_1 . Положим

$$\xi_I^{(1)} = \frac{v_I(n) - na_I^{(1)}(r)}{\sqrt{na_I^{(1)}(r)}}, \quad (16)$$

где

$$a_I^{(1)}(r) = p_{i_1} p_{i_1 i_2} \dots p_{i_{m_1-1} i_{m_1}} p_{i_{m_1} i_{m_1+1}}^{(r+1)} p_{i_{m_1+1} i_{m_1+2}} \dots p_{i_{M-1} i_M},$$

$$p_{i,j}^{(r+1)} = \mathbf{P}\{x_{t+r+1} = j \mid x_t = i\}, \quad M = m_1 + m_2.$$

Нормированные частоты (5) можно представить в следующем виде:

$$\xi_I(n) = \xi_I^{(1)}(n) \sqrt{a_I^{(1)}(r)/a_I^{(0)}(r)} + \sqrt{n} \Delta_I(r),$$

$$\Delta_I(r) = (a_I^{(1)}(r) - a_I^{(0)}(r))/a_I^{(0)}(r). \quad (17)$$

Подставляя полученные выражения для $\xi_I(n)$ в равенство (10), получим, что

$$X_r(n) = \sum_{I,J} (A_M)_{I,J} \xi_I^{(1)}(n) \xi_J^{(1)}(n) \sqrt{a_I^{(1)}(r) a_J^{(1)}(r) / a_I^{(0)}(r) a_J^{(0)}(r)} + 2\sqrt{n} \sum_I g_I \xi_I^{(1)}(n) + n b_r^2, \quad (18)$$

где

$$g_I = \sum_J (A_M)_{I,J} \sqrt{a_J^{(1)}(r) / a_J^{(0)}(r)} \Delta_J(r), \quad b_r^2 = \sum_{I,J} (A)_{I,J} \Delta_I(r) \Delta_J(r). \quad (19)$$

Известно, что для эргодических цепей Маркова величины $\{\xi_I^{(1)}(n)\}$ при $n \rightarrow \infty$ асимптотически нормальны с нулевым вектором математических ожиданий и ковариационной матрицей $B^{(1)}$ с элементами

$$(B^{(1)})_{I,J} = \lim_{n \rightarrow \infty} \text{cov}_{H_1}(\xi_I^{(1)}(n), \xi_J^{(1)}(n)). \quad (20)$$

Отсюда и из равенств (18) получаем следующее утверждение.

Пусть

$$\zeta_n = (X_r(n) - n b_r^2) / (2\sqrt{n}). \quad (21)$$

Теорема 1. При гипотезе H_1

$$\lim_{n \rightarrow \infty} \mathbf{P}\{\zeta_n < x\} = \mathbf{P}\left\{\sum_I g_I \xi_I^{(1)} < x\right\},$$

где ζ_n , g_I определены равенствами (19), (21), величины $\{\xi_I^{(1)}\}$ нормально распределены с нулевым вектором математических ожиданий и ковариационной матрицей $B^{(1)}$, определенной равенствами (20).

Статистику $X_r(n)$ представим в виде

$$X_r(n) = 2\sqrt{n} \zeta_n + n b_r^2.$$

Из теоремы 1 следует, что

$$X_r(n) \approx n b_r^2 + O(\sqrt{n}).$$

Таким образом, возможность выявления зависимости существенно зависит от параметра b_r^2 , который можно вычислить по второй формуле (19). Для иллюстрации найдем параметр b_r^2 , например, в случае $m_1 = m_2$, $N = 2$ и с дважды стохастической матрицей вероятностей переходов

$$(p_{i,j}) = \begin{pmatrix} 1-p/2 & p/2 \\ p/2 & 1-p/2 \end{pmatrix}.$$

В этом случае (см. [3], с. 447)

$$p_{i,j}^{(r)} = \frac{1}{2} (1 + (-1)^{i+j} (1-p)^r), \quad i, j \in \{1, 2\}.$$

Отсюда

$$p_1 = p_2 = \frac{1}{2}, \quad a_I^{(0)}(r) = \frac{1}{4}, \quad I = (i_1, i_2).$$

Используя равенства

$$A_2 = E_2 - G_2, \quad (E)_{IJ} = \delta_{i_1 j_1} \delta_{i_2 j_2}, \quad (G_2)_{I,J} = \frac{1}{2} \delta_{i_1 j_1},$$

по второй формуле (19) получим, что

$$b_r^2 = 4(1-p)^{2(r+1)}, \quad r = 0, 1, 2, \dots$$

3. Предельные распределения статистики $\bar{X}_r(n)$ при основной гипотезе

В работе [2] (см. теорему 1) было показано, что в случае гипотезы H_0 предельное распределение статистики $\bar{X}_r(n)$ при $n \rightarrow \infty$ совпадает с распределением квадратичной формы

$$\bar{X}_r = \sum_{I,J} (\bar{A}_M)_{I,J} \xi_I \xi_J, \quad I = (i_1, \dots, i_M), \quad J = (j_1, \dots, j_M), \quad (22)$$

где случайные величины $\{\xi_I\}$ нормально распределены с нулевым вектором математических ожиданий и ковариационной матрицей B_M , задаваемой равенством (7),

$$\bar{A}_M = E_M - D_M - G_M + 2(M-1)F_M, \quad M = m_1 + m_2, \quad (23)$$

E_M — единичная матрица размера $N^M \times N^M$,

$$(G_M)_{I,J} = \delta_{i_1 j_1} \dots \delta_{i_{M-1} j_{M-1}} \sqrt{p_{i_M} p_{j_M}},$$

$$(F)_{I,J} = \sqrt{a_I^{(0)}(r) a_J^{(0)}(r)}, \quad (D_M)_{I,J} = \delta_{i_M j_M} \sqrt{a_I^{(0)}(r) a_J^{(0)}(r) / (p_{i_M} p_{j_M})}. \quad (24)$$

Из результатов работы [4] следует, что квадратичная форма (22) имеет χ^2 -распределение, если

$$B_M \bar{A}_M B_M \bar{A}_M = B_M \bar{A}_M. \quad (25)$$

Для проверки этого условия нужно перемножить матрицы, входящие в равенства (7) и (23). Для обозначения матриц такого вида воспользуемся обозначениями, введенными в [2]. Матрицу $Q_M(\pi)$, где

$$\pi_k = \begin{pmatrix} i_1 & \dots & i_k \\ j_1 & \dots & j_k \end{pmatrix}$$

— неполная подстановка, определим равенством

$$(Q_M(\pi))_{I,J} = \delta_{u_{i_1} v_{j_1}} \dots \delta_{u_{i_k} v_{j_k}} \sqrt{\frac{p_{u_1} \dots p_{u_M} p_{v_1} p_{v_M}}{p_{u_{i_1}} \dots p_{u_{i_k}} p_{v_{j_1}} \dots p_{v_{j_k}}}},$$

$$I = (i_1, \dots, i_M), \quad J = (j_1, \dots, j_M). \quad (26)$$

В [2] было показано, что

$$Q_M(\pi')Q_M(\pi'') = Q_M(\pi'\pi''), \quad (27)$$

где $\pi'\pi''$ — произведение неполных подстановок. В этих обозначениях матрицы, входящие в (23), записываются следующим образом:

$$D_M = Q_M \begin{pmatrix} M \\ M \end{pmatrix}, \quad G_M = Q_M \begin{pmatrix} 1 & 2 & \dots & M-1 \\ 1 & 2 & \dots & M-1 \end{pmatrix}, \quad F_m = Q_m(\pi_0),$$

где $\pi_0 = \begin{pmatrix} \emptyset \\ \emptyset \end{pmatrix}$ — неполная подстановка, образованная пустыми множествами. Для любой подстановки π матрицы C_M с элементами вида (26) выполняются равенства

$$F_M C_M = C_M F_M = F_M. \quad (28)$$

Отметим также, что

$$G_M D_M = D_M G_M = F_M, \quad D_M D_M = D_M. \quad (29)$$

Используя равенства (28), нетрудно проверить следующее утверждение.

Лемма 1. Для матриц, входящих в (7) и (23), справедливы следующие равенства: $F_M B_M = B_M F_M = O_M$, $\tilde{A}_M F_M = F_M \tilde{A}_M = (2M-3)F_M$, где $O_M = N^M \times N^M$ матрица из нулей.

Вычисление матриц $B_M(k)$, входящих в (7), с использованием равенств (4) не представляет принципиальных трудностей. Используя обозначения (26), приведем выражения для этих матриц в двух рассматриваемых случаях.

Пусть $m_1 = m_2 = m$, $M = 2m$, $s = 2m + r$. При $r < m$

$$\begin{aligned} B_M(k) &= Q_M \begin{pmatrix} k+1 & \dots & m & m+k+1 & \dots & 2m \\ 1 & \dots & m-k & m+1 & \dots & 2m-k \end{pmatrix}, \quad 1 \leq k \leq r; \\ B_M(k) &= Q_M \begin{pmatrix} k+1 & \dots & m & m+1 & \dots & m+k-r & m+k+1 & \dots & 2m \\ 1 & \dots & m-k & m+r+1-k & \dots & m & m+1 & \dots & 2m-k \end{pmatrix}, \\ &\quad r+1 \leq k \leq m-1; \\ B_M(k) &= Q_M \begin{pmatrix} m+1 & \dots & m+k-r \\ m+r+1-k & \dots & m \end{pmatrix}, \quad m \leq k \leq m+r; \\ B_M(k) &= Q_M \begin{pmatrix} k+1-r & \dots & 2m \\ 1 & \dots & 2m+r-k \end{pmatrix}, \quad m+r+1 \leq k \leq s-1. \end{aligned} \quad (30)$$

При $r \geq m$

$$\begin{aligned} B_M(k) &= Q_M \begin{pmatrix} k+1 & \dots & m & m+k+1 & \dots & 2m \\ 1 & \dots & m-k & m+1 & \dots & 2m-k \end{pmatrix}, \quad 1 \leq k \leq m-1; \\ B_M(k) &= F_M, \quad m \leq k \leq r; \\ B_M(k) &= Q_M \begin{pmatrix} m+1 & \dots & m+k-r \\ m+r+1-k & \dots & m \end{pmatrix}, \quad r+1 \leq k \leq m+r; \\ B_M(k) &= Q_M \begin{pmatrix} k+1-r & \dots & 2m \\ 1 & \dots & 2m+r-k \end{pmatrix}, \quad m+r+1 \leq k \leq s-1. \end{aligned} \quad (31)$$

Пусть теперь $m_1 = m$, $m_2 = 1$, $M = m + 1$, $s = m + r + 1$. При $r < m$

$$\begin{aligned} B_M(k) &= Q_M \begin{pmatrix} k+1 & \dots & m & m+k+1 \\ 1 & \dots & m-k & m+1 \end{pmatrix}, & 1 \leq k \leq r; \\ B_M(k) &= Q_M \begin{pmatrix} k+1 & \dots & m & m+1 \\ 1 & \dots & m-k & m+r+1-k \end{pmatrix}, & r+1 \leq k \leq m-1; \\ B_M(k) &= Q_M \begin{pmatrix} m+1 \\ m+r+1-k \end{pmatrix}, & m \leq k \leq s-1. \end{aligned} \quad (32)$$

При $r \geq m$

$$\begin{aligned} B_M(k) &= Q_M \begin{pmatrix} k+1 & \dots & m & m+k+1 \\ 1 & \dots & m-k & m+1 \end{pmatrix}, & 1 \leq k \leq m-1; \\ B_M(k) &= F_M, & m \leq k \leq r; \\ B_M(k) &= Q_M \begin{pmatrix} m+1 \\ m+r+1-k \end{pmatrix}, & r+1 \leq k \leq s-1. \end{aligned} \quad (33)$$

Для проверки условия (25) удобно использовать следующее утверждение.

Лемма 2. В обоих случаях, $m_1 = m_2 = m$ и $m_1 = m$, $m_2 = 1$, верны равенства

$$B_M(k)G_M = B_M(k), \quad G'_M(B_M(k))' = (B_M(k))', \quad D_M(B_M(k))' = B_M(k)D_M = F.$$

Это равенства легко проверяются с использованием формулы (27).

Назовем матрицы $C_M^{(1)}$ и $C_M^{(2)}$ сравнимыми ($C_M^{(1)} \equiv C_M^{(2)}$), если $C_M^{(1)} - C_M^{(2)} = fF_M$, где f – целое число.

Лемма 3. В условиях леммы 2

$$\bar{A}_M B_M \bar{A}_M \equiv \bar{A}_M.$$

Доказательство. Используя равенства (23) и (7), получаем, что

$$B_M \equiv E_M + \sum_{k=1}^{s-1} (B_M(k) + (B_M(k))'), \quad \bar{A}_M \equiv E_M - D_M - G_M.$$

Отсюда и из леммы 2 следует, что

$$\bar{A}_M B_M \equiv E_M - D_M - G_M + \sum_{k=1}^{s-1} (B_M(k) - D_M B_M(k) - G_M B_M(k)).$$

Умножив обе части этого сравнения справа на $\bar{A}_M$ и воспользовавшись равенствами

$$D_M B_M(k) D_M = D_M F_M = F_M, \quad G_M B_M(k) D_M = G_M F_M = F_M,$$

получаем утверждение леммы.

Теорема 2. Если верна гипотеза H_0 , то при любом $r \geq 0$ распределения статистик $\bar{X}_r(n)$ в случаях $m_1 = m_2 = m$, $M = 2m$ и $m_1 = m$, $m_2 = 1$, $M = m + 1$ сходятся к центральному χ^2 -распределению, число степеней свободы которых при всех r одинаково и равно $N^M - N^{M-1} - (N - 1)$.

По лемме 3

$$\bar{A}_M B_M \bar{A}_M = \bar{A}_M + f F_M,$$

где f – целое число. Умножая это равенство слева на B_M и используя равенство (28), получим равенство (25). Таким образом, квадратичная форма (22) в обоих случаях имеет χ^2 -распределение. Числа степеней свободы равны следу матрицы $B_M \bar{A}_M$ и вычисляются непосредственно в каждом случае.

Список литературы

1. Селиванов Б. И., Чистяков В. П., Последовательный χ^2 -критерий, построенный по s -цепочкам состояний цепи Маркова. *Дискретная математика* (1997) **9**, №4, 127–136.
2. Тихомирова М. И., Чистяков В. П., О некоторых статистиках типа χ^2 , функционально зависящих от оценок неизвестных параметров. *Дискретная математика* (2005) **17**, №1, 22–34.
3. Феллер В., *Введение в теорию вероятностей и ее приложения*, **1**. Мир, Москва, 1984.
4. Dik J. J., de Gunst M. C. M., The distribution of general quadratic form in normal variables. *Stat. Neerl.* (1985) **39**, №1, 14–26.

Статья поступила 20.06.2006.