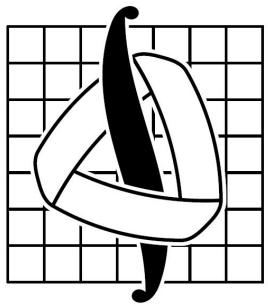
Центральные предельные теоремы для перестановочных случайных величин

Александр Ракитько, МГУ



Международная научная конференция
"Ломоносов-2014"
Москва, 9 сентября

1. Основные определения и понятия.

Определение

Набор случайных величин $\{X_i\}_{i=1}^n$ *называется* **перестановочным**, *если*

$$Law(X_1, \ldots, X_n) = Law(X_{\sigma(1)}, \ldots, X_{\sigma(n)}),$$

где $\sigma \in \mathbb{S}(n)$.

Определение

Бесконечная последовательность $X_1, X_2, \ldots$ *называется* **перестановочной**, *если для каждого* $n \in \mathbb{N}$ *набор* $\{X_i\}_{i=1}^n$ *перестановочен.*

Примеры:

1. Независимые одинаково распределенные случайные величины.
2. Пусть $\{\varepsilon_i\}$ – н.о.р.с.в. Положим $X_i = Y + \varepsilon_i$ или $X_i = Y \cdot \varepsilon_i$.
3. Урновые схемы: выбор без возвращения, схема Пойя. Здесь X_i – цвет шара, извлечённого на $i-$ом шаге.
4. Пусть $n > 1$ и $\sigma : \Omega \to \mathbb{S}(n)$ распределена равномерно. Положим $X_i = \mathbb{I}\{\sigma(i) = i\}$.

Классическая теорема **де Финетти** утверждает:

Теорема

Каждой перестановочной последовательности $\{X_i\}_{i\geq 1}$ *с* $X_1 \in \{0,1\}$ *соответствует единственное распределение* μ *на* $[0,1]$ *такое, что*

$$\mathbb{P}(X_1 = 1, \ldots, X_k = 1, X_{k+1} = 0, \ldots, X_n = 0) = \\ = \int_0^1 p^k (1-p)^{n-k} \mu(dp).$$

В случае конечного набора перестановочных случайных величин теорема де Финетти неверна.

Контрпример:

$$\begin{aligned}\mathbb{P}(X_1 = 0, X_2 = 1) = \mathbb{P}(X_1 = 1, X_2 = 0) &= 1/2,\\ \mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(X_1 = 1, X_2 = 1) &= 0.\end{aligned}$$

Набор $\{X_1, X_2\}$ перестановочен. Но

$$0 = \int_0^1 p^2 \mu(dp) = \int_0^1 (1-p)^2 \mu(dp).$$

Следовательно, подходящей меры μ не существует.

2. ЦПТ для перестановочного массива.

Будем рассматривать массив $\{X_{n,i}, i = 1, \ldots, k_n\}_{n \geq 1}$ случайных величин.

Пусть выполняются следующие условия:

- Набор $\{X_{n,i}, i = 1, \ldots, k_n\}$ перестановочен для каждого $n \in \mathbb{N}$.
- $k_n \in \mathbb{N}$ и $k_n \to \infty$ при $n \to \infty$.
- $\sup_{n \in \mathbb{N}} \mathbb{E} X_{n,1}^4 < \infty$.

Theorem

Пусть выполнены следующие условия

1. $\mathbb{E}X_{n,1}^2 - \mathbb{E}X_{n,1}X_{n,2} \to \sigma^2 > 0$ *при* $n \to \infty$,
2. $\mathrm{cov}(X_{n,1}^2, X_{n,2}^2) + \mathrm{cov}(X_{n,1}X_{n,2}, X_{n,3}X_{n,4}) - -2\mathrm{cov}(X_{n,1}^2, X_{n,1}X_{n,2}) \to 0$ *при* $n \to \infty$.

Тогда для любой возрастающей последовательности $\{m_n\}_{n\geq 1}$ *натуральных чисел такой, что* $m_n/k_n \to \alpha < 1$ *при* $n \to \infty$, *справедливо*

$$m_n^{-1/2} \sum_{i=1}^{m_n} (X_{n,i} - \widehat{\mu}_{k_n}) \to Z_{0,(1-\alpha)\sigma^2} \sim \mathcal{N}(0, (1-\alpha)\sigma^2),$$

где $\widehat{\mu}_{k_n} := k_n^{-1} \sum_{i=1}^{k_n} X_i$.

Рассмотрим $\{Y_i\}_{i=1}^m$ – перестановочный набор, причем

$$\mathbb{E}Y_1 = 0, \quad \mathbb{E}|Y_1|^3 < \infty, \quad \sum_{i=1}^{m} Y_i = 0 \text{ п.н.} \qquad (1)$$

Пусть $\Sigma = (\sigma_{i,j})_{1 \le i,j \le m}$ – ковариационная матрица вектора $(Y_1, \ldots, Y_m)$.

Для функции $h : \mathbb{R}^d \to \mathbb{R}$ и $k \in \mathbb{N}$ положим

$$C_h^{(k)} := \max_{i_1,\ldots,i_d \ge 0, \sum_{j=1}^d i_j = k} \left\| \frac{\partial^k h}{\partial x_1^{i_1} \ldots \partial x_d^{i_d}} \right\|_\infty .$$

Сформулируем вспомогательный результат из [1] (Rollin, 2013):

Лемма

Пусть Y *– вектор, состоящий из перестановочных случайных величин, с ковариационной матрицей* Σ. *Предположим, что условия* (1) *выполнены. Тогда*

$$|\mathbb{E}h(\mathsf{Y}) - \mathbb{E}h(\mathsf{Z})| \leq$$
$$\leq C_h^{(2)} \left[\mathrm{var} \left(\sum_{i=1}^{m} Y_i^2 \right) \right]^{\frac{1}{2}} + 16 m C_h^{(3)} \mathbb{E}|Y_1|^3,$$

где $\mathsf{Z} \sim \mathcal{N}(0, \Sigma)$.

Схема доказательства:

- Переходим к перестановочному набору $\{Y_i := m_n^{-1/2}(X_{n,i} - \widehat{\mu}_{k_n}),\ i = 1, \ldots, k_n\}$.
- Метод Линдеберга.
- С помощью леммы оцениваем

$$|\mathbb{E}h(Y) - \mathbb{E}h(Z)|,$$

где $h(x_1, \ldots, x_{m_n}) = f(x_1 + \ldots + x_{m_n})$, а Z – гауссовский вектор с той же ковариационной структурой, что и Y.

Замечание

Введем $\widehat{\sigma}^2_{k_n} := \frac{1}{k_n}\sum_{i=1}^{k_n}(X_{n,i} - \widehat{\mu}_{k_n})^2$. *Предположим, что*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left((X_{n,1} - \widehat{\mu}_{k_n})/\widehat{\sigma}_{k_n}\right)^4 < \infty.$$

Тогда, для последовательности $(m_n)_{n\in\mathbb{N}}$, *определенной ранее, можно установить следующий вариант центральной предельной теоремы*

$$\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\left(\frac{X_{n,i} - \widehat{\mu}_{k_n}}{\widehat{\sigma}_{k_n}}\right) \xrightarrow{law} Z_{0,1-\alpha} \sim \mathcal{N}(0, 1-\alpha).$$

Интересные результаты можно получать с помощью **мартингального подхода**, предложенного Вебером. Введём фильтрацию

$$\mathfrak{F}_{n,j} := \sigma\{X_{n,1}, \ldots, X_{n,j}, \sum_{i=j+1}^{k_n} X_{n,i}\}.$$

Очевидно, можно переписать

$$\frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} X_{n,i} = \frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} \big(X_{n,i} - \mathbb{E}(X_{n,i}|\mathfrak{F}_{n,(i-1)})\big) + \\ + \frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} \mathbb{E}(X_{n,i}|\mathfrak{F}_{n,(i-1)}).$$

Действительно, в силу измеримости и перестановочности

$$\sum_{i=j}^{k_n} X_{n,i} = \mathbb{E}\Big(\sum_{i=j}^{k_n} X_{n,i} | \mathfrak{F}_{n,(j-1)}\Big) =$$

$$= \sum_{i=j}^{k_n} \mathbb{E}\big(X_{n,j} | \mathfrak{F}_{n,(j-1)}\big) = (k_n - j + 1)\mathbb{E}\big(X_{n,j} | \mathfrak{F}_{n,(j-1)}\big).$$

Таким образом, получаем

$$\mathbb{E}\big(X_{n,j} | \mathfrak{F}_{n,(j-1)}\big) = \frac{1}{k_n - j + 1} \sum_{i=j}^{k_n} X_{n,i}.$$

Приложения.

Пусть $X = (X_1, \ldots, X_n)$ – случайный вектор с $X_k : \Omega \to \{0, 1, \ldots, s\}$, $k = 1, \ldots, n$. Положим

$$\begin{aligned} \mathbb{X} &= \{0,\ldots,s\}^n, \\ \mathbb{Y} &= \{-m,\ldots,0,\ldots,m\}, \end{aligned}$$

здесь $m \in \mathbb{N}$. Предположим, что $Y : \Omega \to \mathbb{Y}$, $f : \mathbb{X} \to \mathbb{Y}$ и задана штрафная функция $\psi : \mathbb{Y} \to \mathbb{R}_+$. Введем функционал ошибки

$$Err(f) := \mathsf{E}|Y - f(X)|\psi(Y).$$

- $Err(f) \to \min\limits_{f} \Longrightarrow f_{opt}$
- Распределение (X, Y) неизвестно:

$$\begin{aligned} f_{opt}(\cdot) &\Longrightarrow \widehat{f}_{PA}(\cdot), \\ \psi(\cdot) &\Longrightarrow \widehat{\psi}(\cdot), \\ Err(\cdot) &\Longrightarrow \widehat{Err}(\cdot). \end{aligned}$$

- K-кроссвалидация. Для $K \in \mathbb{N}$, $(K > 1)$ и $k = 1, \ldots, K$ введем разбиение

$$S_k(N) := \{(k-1)[N/K] + 1, \ldots, k[N/K]\mathbb{I}\{k < K\} + N\mathbb{I}\{k = K\}\}.$$

Удается преобразовать

$$Err(f) = \sum_{i=0}^{2m-1} \sum_{i-m<|y|\leq m} \psi(y)\mathrm{P}(Y = y, |f(X)-y| > i).$$

Тогда в качестве оценки будем использовать

$$\widehat{Err}_K(f_{PA}, \xi_N) := \sum_{i=0}^{2m-1} \sum_{i-m<|y|\leq m} \frac{1}{K} \sum_{k=1}^{K} \widehat{\psi}\big(y, \xi_N(\overline{S_k(N)})\big) \times$$

$$\sum_{j\in S_k(N)} \frac{1}{\sharp S_k(N)} \mathbb{I}\big\{ Y^j = y, |f_{PA}\big(X^j, \xi_N(\overline{S_k(N)})\big) - y| > i \big\}.$$

В случае строго состоятельной оценки $\widehat{\psi}(\cdot)$ в [2] (Булинский, Ракитько, 2014) был установлен критерий, гарантирующий

$$\widehat{Err}_K(f_{PA}, \xi_N) \to Err(f) \text{ п.н., } N \to \infty.$$

Пусть также для оценки $\widehat{\psi}(\cdot)$ справедливо

$$\sqrt{\sharp S_k(N)}(\widehat{\psi}\big(y, \xi_N(S_k(N))\big) - \psi(y)) = O_{\mathrm{P}}(1), \quad N \to \infty,$$

и

$$\sup_{y \in \mathbb{Y},\, N \in \mathbb{N},\, k \in \{1,\ldots,K\}} \mathbb{E}\left(\widehat{\psi}\big(y, \xi_N(S_k(N))\big)\right)^4 < \infty.$$

Теорема

Пусть $(m_N)_{N\in\mathbb{N}}$ – *последовательность натуральных чисел такая, что* $m_N \le q$ *для* $q = [N/K]$ *и*

$$m_N \to \infty, \quad m_N/N \to 0, \quad as \ \ N \to \infty.$$

Выберем такую последовательность положительных чисел $\varepsilon = (\varepsilon_N)_{N\in\mathbb{N}}$, *что* $\varepsilon_N \to 0$ *и* $m_N^{1/2}\varepsilon_N \to \infty$ *as* $N \to \infty$. *Тогда при* $N \to \infty$

$$\sqrt{m_N}\left(\widehat{Err}_K(f_{PA,\varepsilon}, \xi_N) - Err(f)\right) \xrightarrow{law} Z_{0,\sigma^2} \sim \mathcal{N}(0, \sigma^2).$$

Список литературы.

[1] Булинский А.В., Ракитько А.С. (2014). Оценивание небинарного случайного отклика. *Доклады РАН*, **455**(6), 623–627.

[2] Bulinski, A., Butkovsky, O., Sadovnichy, V., Shashkin, A., Yaskov, P., Balatskiy, A., Samokhodskaya, L., Tkachuk, V. (2012). Statistical methods of SNP data analysis and applications. *Open Journal of Statistics.* **2**(1), 73–87.

[3] Röllin, A. (2013). Stein's method in high dimensions with applications. Annales de l'Institut Henri Poincaré, *Probabilités et Statistiques* **49**(2), 529–549.