О равномерной состоятельности непараметрического критерия Неймана

Капаца Дейвид Юрьевич, гр. 622

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

Научный руководитель: д.ф.-м.н., профессор Ермаков М. С.
Рецензент: к.ф.-м.н., с.н.с. Солев В. Н.

Санкт-Петербург, 2022

Введение. Критерии согласия

В работе рассматривается задача проверки гипотезы согласия распределений.

Пусть $X_1, \ldots, X_n$ — независимые, одинаково распределенные случайные величины с ф.р. $F(x)$, $x \in (0, 1]$.

Гипотеза

$$\mathbb{H}_0 : F(x) = F_0(x) = x,$$

Альтернатива

$$\mathbb{H}_1 : F(x) \neq F_0(x).$$

Введение. Метод расстояний

Для создания тестовой статистики критериев согласия обычно применяется *метод расстояний*.

- Берётся псевдометрика $\rho(F, F_0)$, задающая расстояние между гипотезой и неизвестным распределением;
- Тестовая статистика получается подстановкой эмпирической ф.р. $\hat{F}_n(x)$ в функционал $\rho^2(F, F_0)$.
- Примеры: критерии согласия Колмогорова–Смирнова, фон Мизеса, Неймана,...

Таким образом, тестовые статистики по методу расстояний для выборки $X^{(n)} = X_1, \ldots, X_n$ имеют вид

$$T_n(\hat{F}_n) = \rho^2(\hat{F}_n, F_0).$$

Введение. Метод расстояний и альтернативы

В **контексте метода расстояний** при гипотезе

$$\mathbb{H}_0 : F(x) = F_0(x) = x$$

логично рассмотреть непараметрическое множество альтернатив

$$\mathbb{H}_1 : F \in \{F : \rho^2(F, F_0) > b > 0\}.$$

В асимптотической постановке можем допустить динамику (например, «сближение» альтернативы и гипотезы при $n \to \infty$)

$$\mathbb{H}_n : F \in \{F : \rho_n^2(F, F_0) > b_n > 0\}.$$

Введение. Метод расстояний и альтернативы

В **контексте метода расстояний** при гипотезе

$$\mathbb{H}_0 : F(x) = F_0(x) = x$$

логично рассмотреть непараметрическое множество альтернатив

$$\mathbb{H}_1 : F \in \{F : \rho^2(F, F_0) > b > 0\}.$$

В асимптотической постановке можем допустить динамику (например, «сближение» альтернативы и гипотезы при $n \to \infty$)

$$\mathbb{H}_n : F \in \{F : \rho_n^2(F, F_0) > b_n > 0\}.$$

Пусть заданы

- Выборка н.о.р.с.в. $X^{(n)} = (X_1, X_2, \dots, X_n)$ из $F(x)$;
- Плотность $\frac{dF(x)}{dx} = p(x) = 1 + f(x) = 1 + \sum_{j=1}^{\infty} \theta_j \varphi_j(x)$;
- Ортонормированная система функций в $\mathbb{L}_2$: $\{\varphi_j(x)\}_{j=0}^{\infty}$.

Критерии типа Неймана определяются выбором

$$\rho_n^2(F, F_0) = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^2 \left(\int_0^1 \varphi_j(x)\, d\, F(x) \right)^2 = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^2 \theta_j^2 =: T_n(F).$$

Отсюда, подставляя $\hat{F}_n$, получаем тестовую статистику

$$T_n(\hat{F}_n) = \rho_n^2(\hat{F}_n, F_0) = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^2 \bar{\varphi}_j^2(X^{(n)}),$$

где $\bar{\varphi}_j(X^{(n)}) := \frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i)$. Заметим, что $\mathbf{E}_\theta \bar{\varphi}_j = \theta_j$.

Введение. Цели работы

Для критериев типа Неймана

1. Доказать «различимость» гипотезы $\mathbb{H}_0 : F = F_0 = x$ и множеств альтернатив

$$\mathbb{H}_n : F \in \Psi_n = \Big\{ F : T_n(F) > b_n > 0 \Big\};$$

2. Изучить вопрос асимптотической нормальности $T_n(\hat{F}_n)$ при справедливости гипотезы или альтернативы;
3. Найти необходимые и достаточные условия состоятельности (= «различимости») последовательности альтернатив F_n, имеющих заданную скорость сходимости к гипотезе в $\mathbb{L}_2$-норме.

Условия

Тестовые статистики: $T_n(\hat{F}_n) = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^2 \bar{\varphi}_j^2(X^{(n)})$

Условия на $\varkappa_{nj}^2$, $\{\varphi_j(x)\}_{j=1}^{\infty}$, $\boldsymbol{\theta}_n$

K1–K5. Убывание, нормировка, поведение хвоста $\{\varkappa_{nj}^2\}_{j=1}^{\infty}$;

F1. Равномерная ограниченность $\{\varphi_j(x)\}_{j=1}^{\infty}$;

T1$\vee$T2. Ограничение хвостов сумм θ_{nj}^2 при $n \to \infty$:

$$\sum_{j>Ck_n} \theta_{nj}^2 = \begin{cases} O\left(\sum_{j=1}^{Ck_n} \theta_{nj}^2\right); & \text{[T1]} \\ O\left(n^{-1} k_n^{1/2}\right). & \text{[T2]} \end{cases}$$

Здесь $k_n = \sup\left\{k : \sum_{j<k} \varkappa_{nj}^2 \leqslant \frac{1}{2} \sum_{j=1}^{\infty} \varkappa_{nj}^2 < \infty\right\}$.

Оптимальность критериев: состоятельность

K_n — последовательность критериев с уровнями значимости $\alpha_n = \alpha(K_n)$. Обозначим за $\beta(K_n, F_n)$ вероятности ошибок второго рода критериев K_n при альтернативе $F_n \in \Psi_n$. Также, пусть $\beta(K_n, \Psi_n) := \sup_{F_n \in \Psi_n} \beta(K_n, F_n)$.

Определение (Состоятельность)

Последовательность простых альтернатив F_n называется **состоятельной** для K_n, если для любого α, $0 < \alpha < 1$, $\alpha(K_n) = \alpha(1 + o(1))$, имеет место

$$\limsup_{n \to \infty} \beta(K_n, F_n) < 1 - \alpha. \tag{1}$$

Множества альтернатив Ψ_n называются **равномерно состоятельными** для K_n, если

$$\limsup_{n \to \infty} \beta(K_n, \Psi_n) < 1 - \alpha. \tag{2}$$

Постановка задачи

Работа состоит в доказательстве следующего утверждения:

Теорема (Об асимптотической нормальности тестовых статистик)

Пусть выполнены условия K1–K5, F1. Тогда для последовательности критериев K_n имеют место $\alpha(K_n) = \alpha + o(1)$ и

$$\beta(K_n, F_n) = \Phi\left(x_\alpha - T_n(F_n)(2A_n)^{-1/2}\right)(1 + o(1)) \quad (3)$$

равномерно для всех последовательностей F_n, для которых выполняется одно из условий T1 или T2.

Здесь $A_n = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^4$ и по условию K2, $A_n \asymp 1$.
x_α задается уравнением $\Phi(x_\alpha) = 1 - \alpha$, $0 < \alpha < 1$.

Следствия из теоремы об асимптотической нормальности

Следствие 1

Подмножества множеств альтернатив Ψ_n, для которых выполнены условия Т1 или Т2, равномерно состоятельны для критериев K_n при любом $b > 0$.

А также следствие о структуре пространства всех состоятельных альтернатив

Следствие 2

Пусть дана последовательность простых альтернатив F_n, такая, что $\|\boldsymbol{\theta}_n\| = \left\| \frac{dF_n}{dF_0} - 1 \right\| \asymp n^{-1/2} k_n^{1/4}$. Она состоятельна для критериев K_n, если и только если, $T_n(F_n) > b$ для всех $n > n_0$ для некоторого $b > 0$.

Вспомогательные леммы

Пусть выполнены условия теоремы 1.

Лемма (Об оценках)

Имеет место

$$\mathbf{E}_{F_n}\left[T_n(\hat{F}_n)\right] = T_n(F_n) + o(T_n(F_n)),$$

$$\mathbf{D}_{F_n}\left[T_n(\hat{F}_n)\right] = 2A_n + o(T_n^2(F_n)).$$

Используя результаты первой леммы, можем доказать

Лемма (Об асимптотической нормальности при $T_n(F_n) \asymp A_n$)

Пусть $T_n(F_n) \asymp A_n$. *Тогда* P_{F_n}*-распределения тестовых статистик*

$$\left(T_n(\hat{F}_n) - T_n(F_n)\right)(2A_n)^{-1/2}$$

сходятся к стандартному нормальному распределению.

Результаты

Результаты: (*при некоторых условиях)

1. Доказана равномерная состоятельность множеств альтернатив по методу расстояний для статистик типа Неймана;
2. Доказана асимптотическая нормальность тестовой статистики критерия Неймана при альтернативах и гипотезе;
3. Получены необходимые и достаточные условия на состоятельные последовательности альтернатив, сближающихся к гипотезе в $\mathbb{L}_2$ норме.

Приложение

*Ответы на замечания рецензента

- *Автор не всегда делает надлежащие ссылки, когда использует результаты и методы предшествующих работ.*
- *Тестовая статистика выписана только в терминах оценок коэффициентов Фурье плотности, а в терминах функционала от эмпирического распределения в явном виде не выписана...*
- *Используются близкие обозначения* A_n *и* $A_n(\boldsymbol{\theta})$ *для достаточно далеких объектов.*

*Примеры непараметрических тестовых статистик

Выборка н.о.р.с.в. $X^{(n)} = (X_1, X_2, \ldots, X_n) \sim F$, её эмпирическая ф.р. F_n, гипотеза $\mathbb{H}_0 : F(x) = F_0(x)$.

Тест Колмогорова–Смирнова

$$T(X^{(n)}) = \sup_x |F_0(x) - F_n(x)|$$

Тесты Андерсона–Дарлинга и Крамера-фон Мизеса

$$T(X^{(n)}) = \int_{-\infty}^{+\infty} (F_n(x) - F_0(x))^2 w(x)\, dF_0(x)$$

- фон Мизес: $w(x) := 1$
- Андерсон–Дарлинг: $w(x) := [F_0(x)(1 - F_0(x))]^{-1}$

Тест Неймана (Neyman's smooth test, 1937)

$\{\varphi_j(x)\}_{j=1}^{\infty}$ - ортонорм. с.ф. из $\mathbb{L}_2(0,1)$, $\bar{\varphi}_j(X^{(n)}) := \frac{1}{n} \sum_{i=1}^n \varphi_j(X_i)$,

$$\mathbb{H}_0 : \boldsymbol{\theta} = 0, \quad \mathbb{H}_1 : p(x) = 1 + \sum_{j=1}^{\infty} \theta_j \varphi_j(x); \quad T(X^{(n)}; k) = n^2 \sum_{j=1}^{k} \bar{\varphi}_j^2(X^{(n)})$$

*Условия на $\varkappa_{nj}^2$

K1. Для любого n последовательность $\varkappa_{nj}^2$ является убывающей.

K2. Существуют такие константы C_1 и C_2, такие что для любого n

$$C_1 < A_n = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^4 < C_2.$$

K3. Существуют C_1 и $\lambda > 1$, что для любого $\delta > 0$ и n имеет место

$$\varkappa_{n,[(1+\delta)k_n]}^2 < C_1(1+\delta)^{-\lambda}\varkappa_n^2,$$

где $k_n = \sup\left\{k : \sum_{j<k} \varkappa_{nj}^2 \leqslant \frac{1}{2}\sum_{j=1}^{\infty} \varkappa_{nj}^2 < \infty\right\}$.

K4. Имеет место $\varkappa_{n1}^2 \asymp \varkappa_n^2$ при $n \to \infty$. Для любого $c > 1$ существует C такое, что $\varkappa_{n,[ck_n]}^2 \geqslant C\varkappa_n^2$ для любых n.

K5. $k_n = o(n^{2/3})$.

*Условия на $\boldsymbol{\theta}$

Приведём два условия на бесконечномерный вектор $\boldsymbol{\theta}_n$, представляющий из себя вектор коэффициентов Фурье разложения $f(x)$ по ортонормированному базису $\{\varphi_j(x)\}_{j=1}^{\infty}$, где $f(x) = p(x) - 1$.

T1. Существует число $C > 0$ такое, что

$$\sum_{j>Ck_n} \theta_{nj}^2 = O\left(\sum_{j=1}^{Ck_n} \theta_{nj}^2\right);$$

T2. Справедливо для некоторого $C > 0$

$$\sum_{j>Ck_n} \theta_{nj}^2 = O(n^{-1}k_n^{1/2}).$$

Предполагается, что выполнено одно из данных условий.

*Альтернативы $\mathbb{H}_n$: непараметрический подход

Напоминаем гипотезу:

$$\mathbb{H}_0 : F(x) = F_0(x) = x, \quad x \in (0,1)$$

Альтернативы вводятся через плотности распределения:

$$p(x) = 1 + f(x) = \frac{dF(x)}{dx}$$

$\mathbb{H}_n : f \in \Psi_n,\ \Psi_n \subset \mathbb{L}_2(0,1)$, т.е. $f(x) = \sum_{j=1}^{\infty} \theta_j \varphi_j(x)$[a]

[a] $\{\varphi_j(x)\}_{j=0}^{\infty}$ — ортонорм. с.ф. в $\mathbb{L}_2$, $\varphi_0(x) = 1$, φ_j — ограничены

- Параметрические/непараметрические альтернативы
- Практический смысл непараметрических альтернатив
- Насколько широкими могут быть Ψ_n, чтобы критерии оставались «оптимальными»?

*Критерии типа Неймана

Определим критерии проверки гипотез типа Неймана:

$$K_n(\hat{F}_n) = \mathbf{1}_{\left(n\,T_n(\hat{F}_n) - n\sum_{j=1}^{\infty} \varkappa_{nj}^2 > (2A_n)^{1/2} x_\alpha\right)},$$

где x_α задается уравнением $\Phi(x_\alpha) = 1 - \alpha$, $0 < \alpha < 1$.

*Математическое ожидание $T_n(X^{(n)})$

Результаты для «укороченной» версии статистики

$$T_n(X^{(n)}) = 2\sum_{j=1}^{\infty} \varkappa_{nj}^2 \sum_{1\leqslant s<s_1\leqslant n} \varphi_j(X_s)\varphi_j(X_{s_1}).$$

$$\mathbf{E}_\theta T_n(X^{(n)}) = 2\sum_{j=1}^{\infty} \varkappa_{nj}^2 \sum_{1\leqslant s<s_1\leqslant n} \mathbf{E}_\theta \varphi_j(X_s)\varphi_j(X_{s_1}) =$$

$$= 2\sum_{j=1}^{\infty} \varkappa_{nj}^2 \sum_{1\leqslant s<s_1\leqslant n} \theta_j^2 = n(n-1)\sum_{j=1}^{\infty} \varkappa_{nj}^2\theta_j^2 =$$

$$= A_n(\boldsymbol{\theta}) - n\sum_{j=1}^{\infty} \varkappa_{nj}^2\theta_j^2 = A_n(\boldsymbol{\theta}) + o(A_n(\boldsymbol{\theta})).$$

*Статистика типа Неймана: пояснение и метод расстояний

Итак, определили $T_n(\hat{F}_n) = n^2 \sum_{j=1}^{\infty} \varkappa_{nj}^2 \bar{\varphi}_j^2(X^{(n)})$.

Вопрос

Почему они проверяют $\mathbb{H}_0 : \boldsymbol{\theta} = \mathbf{0}$?

$$\mathbf{E}_\theta\left[\varphi_j(X_1)\right] =$$

$$= \int_0^1 \varphi_j(x) p(x)\, dx = \int_0^1 \varphi_j(x) \left(1 + \sum_{l=1}^{\infty} \theta_l \varphi_l(x)\right) dx =$$

$$= 0 + \sum_{l=1}^{\infty} \theta_l (1)_{jl} = \theta_j.$$

- Метод расстояний и статистика типа Неймана

*Дисперсия $T_n(X^{(n)})$

$$T_n(X^{(n)}) - \mathbf{E}_\theta T_n(X^{(n)}) = J_{1n} + 2J_{2n}, \text{ где}$$

$$J_{1n} := 2 \sum_{1 \leqslant s < s_1 \leqslant n} H_n(X_s, X_{s_1}), \tag{4}$$

$$H_n(X_s, X_{s_1}) := \sum_{j=1}^{\infty} \varkappa_{nj}^2 (\varphi_j(X_s) - \theta_j)(\varphi_j(X_{s_1}) - \theta_j), \tag{5}$$

$$J_{2n} := (n-1) \sum_{j=1}^{\infty} \varkappa_{nj}^2 \theta_j \left(\sum_{s=1}^{n} \varphi_j(X_s) - n\theta_j \right) \tag{6}$$

Надо оценить $\mathbf{E}_\theta J_{1n}^2$ и $\mathbf{E}_\theta J_{2n}^2$. При этом ясно, что
$\mathbf{E}_\theta^2 J_{1n} J_{2n} \leqslant \mathbf{E}_\theta J_{1n}^2 J_{2n}^2$.
В итоге получаем, что
$\mathbf{D}_\theta T_n(x^{(n)}) = \mathbf{D}_\theta J_{1n}^2 + o(\mathbf{D}_\theta J_{1n}^2) = 2A_n + o(A_n^2(\boldsymbol{\theta}))$.

*Асимптотическая нормальность $T_n(X^{(n)})$

Доказательство леммы основано на проверке условий теоремы из (P.Hall, 1984) и построении похожих оценок.

Требуется показать, что

$$\lim_{n\to\infty} \left\{ \mathbf{E}_\theta V_n^2(X_1, X_2) + n^{-1}\mathbf{E}_\theta |H_n(X_1, X_2)|^4 \right\} \times \\ \times \left(\mathbf{E}_\theta H_n^2(X_1, X_2) \right)^{-2} = 0,$$

где

$$H_n(X_s, X_{s_1}) := \sum_{j=1}^{\infty} \varkappa_{nj}^2 (\varphi_j(X_s) - \theta_j)(\varphi_j(X_{s_1}) - \theta_j),$$

$$V_n(x, y) := \mathbf{E}_\theta H_n(X_1, x) H_n(X_1, y).$$

*Необходимые вычисления, связанные с $\{\varphi_j\}$

$$(1)_j := \int_0^1 \varphi_j(x)\,dx = 0; \qquad (1)_{jk} := \int_0^1 \varphi_j(x)\varphi_k(x)\,dx; \qquad (1)_{jj} = 1.$$

Непосредственными вычислениями получаем

$$\mathbf{E}_\theta \varphi_j(X_1) = \int_0^1 \varphi_j(x)\tilde{f}(x)\,dx = \int_0^1 \varphi_j(x)\left(1 + \sum_{l=1}^\infty \theta_l \varphi_l(x)\right)dx = 0 + \sum_{l=1}^\infty \theta_l (1)_{jl} = \theta_j.$$

Далее, используя независимость X_1 и X_2, получим

$$\mathbf{E}_\theta \varphi_j(X_1)\varphi_j(X_2) = \mathbf{E}_\theta \varphi_j(X_1) \cdot \mathbf{E}_\theta \varphi_j(X_2) = \theta_j^2.$$

Также потребуется вычислить

$$\mathbf{E}_\theta \varphi_j(X_1)\varphi_k(X_1) = \int_0^1 \varphi_j(x)\varphi_k(x)\left(1 + \sum_{l=1}^\infty \theta_l \varphi_l(x)\right)dx = (1)_{jk} + \sum_{l=1}^\infty \theta_l (1)_{jkl}$$

и следующее выражение

$$\mathbf{E}_\theta \varphi_j^2(X_1) = \int_0^1 \varphi_j^2(x)\tilde{f}(x)\,dx = \int_0^1 \varphi_j^2(x)\left(1 + \sum_{l=1}^\infty \theta_l \varphi_l(x)\right)dx = 1 + \sum_{l=1}^\infty \theta_l (1)_{jjl}.$$