Теория вероятностей и математическая статистика
Лектор А.А. Лобузов
Семестр 6 Лекция 8

Критерий Колмогорова проверки гипотезы о соответствии выборки заданному закону распределения

Критерий Колмогорова является непараметрическим критерием согласия.

Рассмотрим случайную выборку $\mathbf{X}=(X_1, X_2, ..., X_N)$, состоящую из независимых наблюдений случайной величины ξ.

Основная гипотеза: $\mathbf{H}_0=\{F_\xi(x)=F(x)\}$, где $F(x)$ − непрерывная функция распределения.

Конкурирующая гипотеза: $\mathbf{H}_1=\{F_\xi(x)\neq F(x)\}$.

Проверка гипотезы $\mathbf{H}_0$ основана на теореме Колмогорова.

Рассмотрим эмпирическую функцию распределения

$$F_N(x,\mathbf{X})=F_N(x,X_1,X_2,...,X_N)=\frac{1}{N}\sum_{k=1}^{N}I_{(-\infty,x]}(X_k)=\frac{1}{N}\sum_{k=1}^{N}I_{(X_k\le x]}$$

и статистику Колмогорова $D_N(\mathbf{X})=\sup\{|F_N(x,\mathbf{X})-F(x)|:-\infty<x<+\infty\}$.

Теорема Колмогорова.

Пусть элементы выборки X_k независимы и имеют непрерывную функцию распределения $F(x)$. Тогда $\mathrm{P}(\sqrt{N}D_N(\mathbf{X})\le z)\xrightarrow[N\to\infty]{}K(z)$,

где

$$K(z)=\begin{cases}0, & z\le 0;\\ \sum\limits_{j=-\infty}^{+\infty}(-1)^j e^{-2j^2z^2}, & z>0.\end{cases}$$

Схема проверки гипотезы $\mathbf{H}_0 = \{F_\xi(x) = F(x)\}$ при конкурирующей (альтернативной) гипотезе $\mathbf{H}_1 = \{F_\xi(x) \neq F(x)\}$ по выборке $\mathbf{x} = (x_1, x_2, ..., x_N)$ при уровне значимости α с помощью критерия Колмогорова:

1. По числовой выборке $\mathbf{x} = (x_1, x_2, ..., x_N)$ строим вариационный ряд $x_{(1)} \le x_{(2)} \le ... \le x_{(N-1)} \le x_{(N)}$.

2. Находим значение

$$D_N = \max_{1 \le j \le N} (\max(|F_N(x_{(j)}) - F(x_{(j)})|, |F_N(x_{(j)} - 0) - F(x_{(j)})|)),$$

где $F_N(x_{(j)}) = \frac{j}{N}$, $F_N(x_{(j)} - 0) = \frac{j-1}{N}$.

3. По заданному значению уровня значимости α (часто берут $\alpha = 0{,}05$, $0{,}02$, $0{,}01$, $0{,}005$, $0{,}001$) берем по функции распределения Колмогорова критическое значение k_α (имеются таблицы критических значений).

Далее делается вывод о справедливости гипотезы:

если $\sqrt{N} D_N \le k_\alpha$, то при уровне значимости α принимается основная гипотеза $\mathbf{H}_0$;

если $\sqrt{N} D_N > k_\alpha$, то при уровне значимости α принимается альтернативная гипотеза $\mathbf{H}_1$.

Критическое значение k_α находится из соотношения $K(k_\alpha) = 1 - \alpha$.

Этот алгоритм применяется для негруппированных выборок, на практике он используют при $N \ge 20$.

Рассмотрим конкурирующие гипотезы $\mathbf{H}_1^+ = \{F_\xi(x) \geq F(x)\}$ и $\mathbf{H}_1^- = \{F_\xi(x) \leq F(x)\}$, при этом требуется, что бы существовали такие x, при которых неравенства были строгими.

Для проверки гипотезы $\mathbf{H}_0$ при конкурирующей гипотезе $\mathbf{H}_1^+$ используют статистику Смирнова

$$D_N^+(\mathbf{X}) = \sup\{(F_N(x,\mathbf{X}) - F(x)) : -\infty < x < +\infty\}.$$

Она используется в критерии $\sqrt{N}D_N^+(\mathbf{X})$, для которого справедлива следующая асимптотика $\mathrm{P}(\sqrt{N}D_N^+(\mathbf{X}) \leq z) \xrightarrow[N\to\infty]{} 1 - e^{-2z^2}$, $0 \leq z < \infty$.

Схема проверки гипотезы $\mathbf{H}_0$ при конкурирующей (альтернативной) гипотезе $\mathbf{H}_1^+$ по выборке $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ при уровне значимости α с помощью критерия Смирнова:

1. По числовой выборке $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ строим вариационный ряд $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N-1)} \leq x_{(N)}$.

2. Находим значение

$$D_N^+ = \max_{1\leq j\leq N}(\max((F_N(x_{(j)}) - F(x_{(j)})), (F_N(x_{(j)} - 0) - F(x_{(j)})))),$$

где $F_N(x_{(j)}) = \dfrac{j}{N}$, $F_N(x_{(j)} - 0) = \dfrac{j-1}{N}$.

3. По заданному значению уровня значимости α находим критическое значение $s_\alpha = \sqrt{-\dfrac{1}{2}\ln\alpha}$.

Далее делается вывод о справедливости гипотезы:

если $\sqrt{N}D_N^+ \leq s_\alpha$, то при уровне значимости α принимается основная гипотеза $\mathbf{H}_0$;

если $\sqrt{N}D_N^{+} > s_\alpha$, то при уровне значимости α принимается конкурирующая гипотеза $\mathbf{H}_1^{+}$.

Для проверки гипотезы $\mathbf{H}_0$ при конкурирующей гипотезе $\mathbf{H}_1^{-}$ используют статистику Смирнова

$$D_N^{-}(\mathbf{X}) = \sup\{(F(x) - F_N(x, \mathbf{X})) : -\infty < x < +\infty\}.$$

Она используется в критерии $\sqrt{N}D_N^{-}(\mathbf{X})$, для которого справедлива следующая асимптотика $\mathrm{P}(\sqrt{N}D_N^{-}(\mathbf{X}) \le z) \xrightarrow[N\to\infty]{} 1 - e^{-2z^2}$, $0 \le z < \infty$.

Схема проверки гипотезы $\mathbf{H}_0$ при конкурирующей (альтернативной) гипотезе $\mathbf{H}_1^{-}$ по выборке $\mathbf{x} = (x_1, x_2, ..., x_N)$ при уровне значимости α с помощью критерия Смирнова:

1. По числовой выборке $\mathbf{x} = (x_1, x_2, ..., x_N)$ строим вариационный ряд $x_{(1)} \le x_{(2)} \le ... \le x_{(N-1)} \le x_{(N)}$.

2. Находим значение

$$D_N^{-} = \max_{1\le j\le N}(\max((F(x_{(j)}) - F_N(x_{(j)})), (F(x_{(j)}) - F_N(x_{(j)} - 0)))),$$

где $F_N(x_{(j)}) = \dfrac{j}{N}$, $F_N(x_{(j)} - 0) = \dfrac{j-1}{N}$.

3. По заданному значению уровня значимости α находим критическое значение $s_\alpha = \sqrt{-\dfrac{1}{2}\ln\alpha}$.

Далее делается вывод о справедливости гипотезы:

если $\sqrt{N}D_N^{-} \le s_\alpha$, то при уровне значимости α принимается основная гипотеза $\mathbf{H}_0$;

если $\sqrt{N} D_N^- > s_\alpha$, то при уровне значимости α принимается конкурирующая гипотеза $\mathbf{H}_1^-$.

Следует отметить, что при $N \le 100$ критические значения s_α нужно находить, используя таблицы для точных распределений статистик D_N^+ и D_N^- (см. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики.– М.:Наука, 1983. – 416 с.).

Критерий $\boldsymbol{\omega}^2$ Крамера-Мизеса-Смирнова проверки гипотезы о соответствии выборки заданному закону распределения

Критерий $\boldsymbol{\omega}^2$ Крамера-Мизеса-Смирнова проверки гипотезы $\mathbf{H}_0 = \{F_\xi(x) = F(x)\}$ при конкурирующей (альтернативной) гипотезе $\mathbf{H}_1 = \{F_\xi(x) \ne F(x)\}$ основан на применении статистики

$$\boldsymbol{\omega}_N^2 = \frac{1}{12N^2} + \frac{1}{N}\sum_{j=1}^{N}[F(X_{(j)}) - \frac{2j-1}{2N}]^2,$$

где $X_{(1)}$, $X_{(2)}$, …, $X_{(N)}$ − порядковые статистики случайной выборки $\mathbf{X} = (X_1, X_2, ..., X_N)$. Она используется в критерии $N\boldsymbol{\omega}_N^2$, для которого справедлива следующая асимптотика $\mathrm{P}(N\boldsymbol{\omega}_N^2 \le z) \xrightarrow[N\to\infty]{} a_1(z) =$

$$= \frac{1}{\sqrt{2z}}\sum_{k=0}^{\infty}\frac{\Gamma(k+0{,}5)}{\Gamma(0{,}5)\Gamma(k+1)}\sqrt{4k+1}[I_{-1/4}(\frac{(4k+1)^2}{16z}) - I_{1/4}(\frac{(4k+1)^2}{16z})]exp\{-\frac{(4k+1)^2}{16z}\}$$

где $0 \le z < \infty$, $\Gamma(y)$ − гамма-функция, $I_m(y)$ − модифицированная функция Бесселя.

Значения функции $a_1(z)$ имеются в таблицах (см. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики.– М.:Наука, 1983. – 416 с.).

Схема проверки гипотезы $\mathbf{H}_0 = \{F_\xi(x) = F(x)\}$ при конкурирующей (альтернативной) гипотезе $\mathbf{H}_1 = \{F_\xi(x) \neq F(x)\}$ по выборке $\mathbf{x} = (x_1, x_2, ..., x_N)$ при уровне значимости α с помощью критерия $\boldsymbol{\omega}^2$ Крамера-Мизеса-Смирнова:

4. По числовой выборке $\mathbf{x} = (x_1, x_2, ..., x_N)$ строим вариационный ряд $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(N-1)} \leq x_{(N)}$.

5. Находим значение

$$n\boldsymbol{\omega}_N^2 = \frac{1}{12N} + \sum_{j=1}^{N} [F(X_{(j)}) - \frac{2j-1}{2N}]^2 .$$

6. По заданному значению уровня значимости α берем критическое значение $A_\alpha = a_1^{-1}(1-\alpha)$.

Далее делается вывод о справедливости гипотезы:

если $n\boldsymbol{\omega}_N^2 \leq A_\alpha$, то при уровне значимости α принимается основная гипотеза $\mathbf{H}_0$;

если $n\boldsymbol{\omega}_N^2 > A_\alpha$, то при уровне значимости α принимается альтернативная гипотеза $\mathbf{H}_1$.