Теория вероятностей и математическая статистика
Лектор А.А. Лобузов
Семестр 6 **Лекция 9**

Критерии однородности двух выборок

Рассматриваются случайные выборки

$\mathbf{X}=(X_1, X_2, ..., X_N)$, полученная при независимых наблюдениях случайной величины ξ с функцией распределения $F_\xi(x)$;

$\mathbf{Y}=(Y_1, Y_2, ..., Y_M)$, полученная при независимых наблюдениях случайной величины η с функцией распределения $F_\eta(x)$.

Пусть $F_N(x)$ − эмпирическая функция распределения для выборки $\mathbf{X}=(X_1, X_2, ..., X_N)$, $F_M(x)$ − эмпирическая функция распределения для выборки $\mathbf{Y}=(Y_1, Y_2, ..., Y_M)$.

Основная гипотеза: $\mathbf{H}_0=\{F_\xi(x)= F_\eta(x)\}$.

В качестве конкурирующей гипотезы чаще всего берётся: $\mathbf{H}_1=\{F_\xi(x)\neq F_\eta(x)\}$.

Критерий однородности Колмогорова-Смирнова

Критерий Колмогорова-Смирнова для проверки однородности двух случайных выборок применяется в случае, когда функции распределения наблюдаемых случайных величины ξ и η непрерывны.

При проверке гипотезы $\mathbf{H}_0 = \{F_\xi(x) = F_\eta(x)\}$ против гипотезы $\mathbf{H}_1 = \{F_\xi(x) \neq F_\eta(x)\}$ в критерии Колмогорова-Смирнова рассматривается статистика

$$D_{N,M}(\mathbf{X},\mathbf{Y}) = \sup\{|F_N(x,\mathbf{X}) - F_M(x,\mathbf{Y})| : -\infty < x < +\infty\},$$

для которой выполняется свойство

$$\mathrm{P}(D_{N,M}(\mathbf{X},\mathbf{Y})\sqrt{\frac{NM}{N+M}} \le z) \xrightarrow[N\to\infty]{} K(z),$$

где

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

Схема проверки гипотезы $\mathbf{H}_0$ при конкурирующей гипотезе $\mathbf{H}_1$ по выборкам $\mathbf{x} = (x_1, x_2, ..., x_N)$ и $\mathbf{y} = (y_1, y_2, ..., y_M)$, полученных при наблюдениях случайных величин ξ и η соответственно, при уровне значимости α:

1. По числовым выборкам $\mathbf{x} = (x_1, x_2, ..., x_N)$ и $\mathbf{y} = (y_1, y_2, ..., y_M)$ строим вариационные ряды $x_{(1)} \le x_{(2)} \le ... \le x_{(N-1)} \le x_{(N)}$ и $y_{(1)} \le y_{(2)} \le ... \le y_{(M-1)} \le y_{(M)}$ соответственно.

2. Находим значение

$$D_{N,M} = \max_{j,k}\{|F_N(x_{(j)}) - F_M(x_{(j)})|, |F_N(x_{(j)} - 0) - F_M(x_{(j)})|,$$

$$|F_N(y_{(k)}) - F_M(y_{(k)})|, |F_N(y_{(k)}) - F_M(y_{(k)} - 0)|\},$$

где $F_N(x_{(j)}) = \frac{j}{N}$, $F_N(x_{(j)} - 0) = \frac{j-1}{N}$, $F_M(y_{(k)}) = \frac{k}{M}$, $F_M(y_{(k)} - 0) = \frac{k-1}{M}$;

$$F_M(x_{(j)})=\begin{cases}0, x_{(j)}<y_{(1)};\\ \frac{k}{M}, y_{(k)}\le x_{(j)}<y_{(k+1)};\\ 1, x_{(j)}\ge y_{(M)};\end{cases}\quad F_N(y_{(k)})=\begin{cases}0, y_{(k)}<x_{(1)};\\ \frac{j}{N}, x_{(j)}\le y_{(k)}<x_{(j+1)};\\ 1, y_{(k)}\ge x_{(N)}.\end{cases}$$

Затем находим $K_{N,M}=D_{N,M}\sqrt{\frac{NM}{N+M}}$.

3. По заданному значению уровня значимости α берем по функции распределения Колмогорова критическое значение k_α.

Далее делается вывод о справедливости гипотезы:

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

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

Иногда рассматривают конкурирующие гипотезы

$\mathbf{H}_1^+=\{\sup\{\mathrm{M}[F_N(x,\mathbf{X})-F_M(x,\mathbf{Y})]:-\infty<x<+\infty\}>0\}$ и

$\mathbf{H}_1^-=\{\inf\{\mathrm{M}[F_N(x,\mathbf{X})-F_M(x,\mathbf{Y})]:-\infty<x<+\infty\}<0\}$.

При проверке гипотезы $\mathbf{H}_0$ против гипотезы $\mathbf{H}_1^+$ берут статистику

$$K_{N,M}^+(\mathbf{X},\mathbf{Y})=D_{N,M}^+(\mathbf{X},\mathbf{Y})\sqrt{\frac{NM}{N+M}},$$

где $D_{N,M}^+(\mathbf{X},\mathbf{Y})=\sup\{[F_N(x,\mathbf{X})-F_M(x,\mathbf{Y})]:-\infty<x<+\infty\}$.

При проверке гипотезы $\mathbf{H}_0$ против гипотезы $\mathbf{H}_1^-$ берут статистику

$$K_{N,M}^-(\mathbf{X},\mathbf{Y})=D_{N,M}^-(\mathbf{X},\mathbf{Y})\sqrt{\frac{NM}{N+M}},$$

где $D_{N,M}^-(\mathbf{X},\mathbf{Y})=-\inf\{[F_N(x,\mathbf{X})-F_M(x,\mathbf{Y})]:-\infty<x<+\infty\}$.

Статистики $K^{+}_{N,M}(\mathbf{X},\mathbf{Y})$, $K^{-}_{N,M}(\mathbf{X},\mathbf{Y})$, а также $K^{+}_{M,N}(\mathbf{X},\mathbf{Y})$, $K^{-}_{M,N}(\mathbf{X},\mathbf{Y})$ имеют одинаковое распределение и для них выполняется свойство

$$\lim_{N\to\infty,M\to\infty} \mathrm{P}(K^{+}_{N,M}(\mathbf{X},\mathbf{Y})\le z) = \lim_{N\to\infty,M\to\infty} \mathrm{P}(K^{-}_{N,M}(\mathbf{X},\mathbf{Y})\le z) =,$$

$$= \lim_{N\to\infty,M\to\infty} \mathrm{P}(K^{+}_{M,N}(\mathbf{X},\mathbf{Y})\le z) = \lim_{N\to\infty,M\to\infty} \mathrm{P}(K^{-}_{M,N}(\mathbf{X},\mathbf{Y})\le z) = 1-e^{-2z^2},\ 0\le z<\infty.$$

Схема проверки гипотезы $\mathbf{H}_0$ при конкурирующей гипотезе $\mathbf{H}^{+}_1$ по выборкам $\mathbf{x}=(x_1, x_2,..., x_N)$ и $\mathbf{y}=(y_1, y_2,..., y_M)$, полученных при наблюдениях случайных величин ξ и η соответственно, при уровне значимости α:

1. По числовым выборкам $\mathbf{x}=(x_1, x_2,..., x_N)$ и $\mathbf{y}=(y_1, y_2,..., y_M)$ строим вариационные ряды $x_{(1)}\le x_{(2)}\le...\le x_{(N-1)}\le x_{(N)}$ и $y_{(1)}\le y_{(2)}\le...\le y_{(M-1)}\le y_{(M)}$ соответственно.

2. Находим значение

$$D^{+}_{N,M} = \max_{j,k}\{(F_N(x_{(j)})-F_M(x_{(j)})),\ (F_N(x_{(j)}-0)-F_M(x_{(j)})),$$
$$(F_N(y_{(k)})-F_M(y_{(k)})),\ (F_N(y_{(k)})-F_M(y_{(k)}-0))\},$$

где $F_N(x_{(j)})=\frac{j}{N}$, $F_N(x_{(j)}-0)=\frac{j-1}{N}$, $F_M(y_{(k)})=\frac{k}{M}$, $F_M(y_{(k)}-0)=\frac{k-1}{M}$;

$$F_M(x_{(j)})=\begin{cases} 0, x_{(j)}<y_{(1)};\\ \frac{k}{M}, y_{(k)}\le x_{(j)}<y_{(k+1)};\\ 1, x_{(j)}\ge y_{(M)};\end{cases} \quad F_N(y_{(k)})=\begin{cases} 0, y_{(k)}<x_{(1)};\\ \frac{j}{N}, x_{(j)}\le y_{(k)}<x_{(j+1)};\\ 1, y_{(k)}\ge x_{(N)}.\end{cases}$$

Затем находим $K^{+}_{N,M}=D^{+}_{N,M}\sqrt{\frac{NM}{N+M}}$.

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

Далее делается вывод о справедливости гипотезы:

если $K^+_{N,M} \le s_\alpha$, то при уровне значимости α принимается основная гипотеза $\mathbf{H}_0$;

если $K^+_{N,M} > s_\alpha$, то при уровне значимости α принимается альтернативная гипотеза $\mathbf{H}^+_1$.

Схема проверки гипотезы $\mathbf{H}_0$ при конкурирующей гипотезе $\mathbf{H}^-_1$ проводится аналогично.

Критерий однородности χ^2

Когда наблюдаемые случайные величины ξ и η дискретны, можно применить критерий однородности, основанный на распределении χ^2.

Пусть дискретные случайные величины ξ и η принимают значения $x^*_1 < x^*_2 < ...$

Схема проверки гипотезы $\mathbf{H}_0 = \{F_\xi(x) = F_\eta(x)\}$ против конкурирующей гипотезы $\mathbf{H}_1 = \{F_\xi(x) \ne F_\eta(x)\}$ по выборкам $\mathbf{x} = (x_1, x_2, ..., x_N)$ и $\mathbf{y} = (y_1, y_2, ..., y_M)$, полученных при наблюдениях случайных величин ξ и η соответственно, при уровне значимости α:

1. По числовым выборкам $\mathbf{x} = (x_1, x_2, ..., x_N)$ и $\mathbf{y} = (y_1, y_2, ..., y_M)$ находим частоты n_{i1} и n_{i2} (n_{i1} – число значений x^*_i,

встречающихся в выборке $\mathbf{x}=(x_1, x_2,..., x_N)$, n_{i2} – число значений x_i^*, встречающихся в выборке $\mathbf{y}=(y_1, y_2,..., y_M)$).

2. Находим выборочное значение критерия

$$\chi_B^2=(N+M)[\sum_{i=1}^{m}(\frac{(n_{i1})^2}{N(n_{i1}+n_{i2})}+\frac{(n_{i2})^2}{M(n_{i1}+n_{i2})})-1] .$$

3. По заданному значению уровня значимости α берем по функции распределения $\chi^2(l)$ с число степеней свободы $l=m-1$ критическое значение $\chi^2_{кр,\alpha}(l)$.

Далее делается вывод о справедливости гипотезы:

если $\chi_B^2 \le \chi^2_{кр,\alpha}(l)$, то при уровне значимости α принимается основная гипотеза $\mathbf{H}_0$;

если $\chi_B^2 > \chi^2_{кр,\alpha}(l)$, то при уровне значимости α принимается альтернативная гипотеза $\mathbf{H}_1^+$.

Критическое значение $\chi^2_{кр,\alpha}(l)=F_{\chi^2(l)}(1-\alpha)$, где $F_{\chi^2(l)}(1-\alpha)$ – функция распределения закона $\chi^2(l)$ с плотностью

$$f(x)=\begin{cases} 0, x\le 0; \\ \dfrac{x^{\frac{l}{2}-1}}{2^{\frac{l}{2}}\Gamma\left(\frac{l}{2}\right)}e^{-\frac{x}{2}}, x>0. \end{cases}$$