

Лекция 7.

Критерии для проверки гипотез о распределении

Простая параметрическая гипотеза

Пусть имеется выборка (X_1, \dots, X_n) объема n из неизвестного распределения \mathcal{F} . Проверяется простая гипотеза $H_0 : \mathcal{F} = \mathcal{F}_1$ против $H_1 : \mathcal{F} \neq \mathcal{F}_1$, где \mathcal{F}_1 - распределение известного типа с известными нами параметрами $\theta = (\theta_1, \dots, \theta_m)$

I. Критерий Колмогорова

Если \mathcal{F}_1 - абсолютно непрерывное распределение с функцией распределения $F(x)$, то применим критерий

$\square K = \sqrt{n} \sup_x |F^*(x) - F(x)|$, где $F^*(x)$ - выборочная функция распределения

То есть используем теорему Колмогорова: если $H_0 : \mathcal{F} = \mathcal{F}_1$, то $K = \sqrt{n} \sup_x |F^*(x) - F(x)| \Rightarrow \mathcal{K}$ - распределение Колмогорова с функцией распределения $F_{\mathcal{K}}(x) = \sum_{j=-\infty}^x (-1)^j e^{-2j^2 x^2}$

Для уровня значимости α находим квантиль t_α такой, что $P(\xi \geq t_\alpha) = \alpha$, где $\xi \in \mathcal{K}$

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{F}_1, & \text{если } K < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{F}_1, & \text{если } K \geq t_\alpha \end{cases}$$

II. Критерий «хи-квадрат» Пирсона

Пусть выборка разбита на k интервалов A_1, A_2, \dots, A_k , $A_i = [a_{i-1}, a_i)$

n_i - соответствующая частота интервала

При распределении \mathcal{F}_1 теоретические вероятности попадания в эти интервалы $p_i = F_{\mathcal{F}_1}(a_i) - F_{\mathcal{F}_1}(a_{i-1})$. Тогда $n'_i = p_i \cdot n$ - теоретические частоты

В качестве статистики критерия выберем $\chi^2_{\text{набл.}} = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i}$

Th. Пирсона. Если $H_0 : \mathcal{F} = \mathcal{F}_1$ верна, то $\chi^2_{\text{набл.}} \Rightarrow \chi^2_{k-1}$ - распределение «хи-квадрат» с $k - 1$ степенями свободы

Критерий: $\square t_\alpha$ - квантиль χ^2_{k-1} уровня α

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} \geq t_\alpha \end{cases}$$

Nota. Часто обозначают $t_\alpha = \chi^2_{\text{теор.}}$

Nota. При этом частота каждого интервала должна быть не меньше 5, а объем выборки - не меньше 50. Число интервалов лучше брать по формуле Стерджесса

Сложная параметрическая гипотеза

Здесь мы будем проверять гипотезу $H_0 : \mathcal{F} \in \mathcal{F}_\theta$ против $H_1 : \mathcal{F} \notin \mathcal{F}_\theta$, где \mathcal{F}_θ - распределение известного типа с неизвестными параметрами

III. Критерий «хи-квадрат» Фишера

Пусть выборка разбита на k интервалов A_1, A_2, \dots, A_k , $A_i = [a_{i-1}, a_i)$

n_i - соответствующая частота интервала A_i

Def. Оценка максимального правдоподобия по частотам называется значения неизвестных параметров, при которых вероятность появления таких частот является максимальной

Пусть $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ - оценка максимального правдоподобия по частотам неизвестных параметров. Тогда теоретические вероятности попадания в интервал считаем по формуле $p_i = F_{\mathcal{F}_\theta}(a_i) - F_{\mathcal{F}_\theta}(a_{i-1})$, теоретическая частота - $n'_i = np_i$

В качестве статистики критерия берется функция $\chi^2_{\text{набл.}} = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i}$

Th. Фишера.

Если $H_0 : \mathcal{F} \in \mathcal{F}_\theta$ верна, то $\chi^2_{\text{набл.}} = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i} \Rightarrow \chi^2_{k-m-1}$ - распределение «хи-квадрат», где m - число параметров неизвестного распределения

$\square t_\alpha$ - квантиль χ^2_{k-m-1} уровня α

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} \geq t_\alpha \end{cases}$$

Nota. Часто в качестве оценки неизвестных параметров берется просто оценка максимального правдоподобия

Ex. Имеется выборка в виде вариационного ряда (5.2; ...; 22.8), $n = 120$, при разбиении на $k = 8$ интервалов получили

A_i	[5.2, 7.4)	[7.4, 9.6)	[9.6, 11.8)	[11.8, 14)	[14, 16.2)	[16.2, 18.4)	[18.4, 20.6)	[20.6, 22.8)	\sum
n_i	12	17	14	13	18	14	13	19	120

Проверить гипотезу о равномерном распределении $H_0 : \mathcal{F} \in U(a, b)$ против $H_1 : \mathcal{F} \notin U(a, b)$ при уровне значимости $\alpha = 0.05$

Дадим оценку параметров методом максимального правдоподобия: $\hat{a} = 5.2$ $\hat{b} = 22.8$

Теоретическая вероятность будет $p'_i = \frac{1}{8}$, теоретическая частота - $n'_i = 15$

$$\chi^2_{\text{набл.}} = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i} = \frac{(12 - 15)^2}{15} + \frac{(17 - 15)^2}{15} + \frac{(14 - 15)^2}{15} + \frac{(13 - 15)^2}{15} + \frac{(18 - 15)^2}{15} + \frac{(14 - 15)^2}{15} + \frac{(13 - 15)^2}{15} + \frac{(19 - 15)^2}{15}$$

$$\frac{(19-15)^2}{15} = 3.2$$

При $\alpha = 0.05$ и $S = k - m - 1 = 5$ квантиль χ_S^2 уровня α равен $t_\alpha = 11.07$

Так как $\chi_{\text{набл.}}^2 < t_\alpha$, нулевая гипотеза о равномерном распределении принимается

Критерии для проверки однородности

IV. Критерий Колмогорова-Смирнова

Пусть имеются 2 независимых выборки (X_1, \dots, X_n) и (Y_1, \dots, Y_m) объемов n и m из неизвестных непрерывных распределений \mathcal{F} и \mathcal{J}

Проверяется $H_0 : \mathcal{F} = \mathcal{J}$ (данные однородны) против $H_1 : \mathcal{F} \neq \mathcal{J}$

В качестве статистики критерия берется функция $K = \sqrt{\frac{nm}{n+m}} \sup_x |F^*(x) - G^*(x)|$, где F^* и G^* - соответствующие выборочные функции распределения

Th. Колмогорова-Смирнова.

Если $H_0 : \mathcal{F} = \mathcal{J}$ верна, то $K \Rightarrow \mathcal{K}$ - распределение Колмогорова

Критерий: t_α - квантиль \mathcal{K} уровня значимости α

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{J}, & \text{если } K < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{J}, & \text{если } K \geq t_\alpha \end{cases}$$

Проверки однородности выборок из нормальных совокупностей

V. Критерий Фишера

Пусть имеют две независимые выборки (X_1, \dots, X_n) и (Y_1, \dots, Y_m) объемов n и m из нормальных распределений $X \in N(a_1, \sigma_1^2)$ и $Y \in N(a_2, \sigma_2^2)$

Проверяется $H_0 : \sigma_1 = \sigma_2$ против $H_1 : \sigma_1 \neq \sigma_2$

В качестве статистики критерия берется функция $K = \frac{S_X^2}{S_Y^2}$, где S_X^2, S_Y^2 - соответствующие исправленные дисперсии, причем $S_X^2 \geq S_Y^2$

Th. Если $H_0 : \sigma_1 = \sigma_2$ верна, то $K = \frac{S_X^2}{S_Y^2} \in F(n-1, m-1)$ - распределение Фишера-Снедекера

По пункту 3 основной теоремы $\frac{(n-1)S^2}{\sigma^2} \in \chi_{n-1}^2$. Если H_0 верна, то $K = \frac{S_X^2}{S_Y^2} =$

$$\frac{(n-1)S_X^2 \sigma_2^2 (m-1)}{\sigma_1^2 (m-1) S_Y^2 (n-1)} = \frac{\frac{\chi_{n-1}^2}{n-1}}{\frac{\chi_{m-1}^2}{m-1}} \in F(n-1, m-1)$$

t_α - квантиль $F(n-1, m-1)$ уровня α

$$\begin{cases} H_0 : \sigma_1 = \sigma_2, & \text{если } K < t_\alpha \\ H_1 : \sigma_1 \neq \sigma_2, & \text{если } K \geq t_\alpha \end{cases}$$

Nota. Здесь критерий согласия работает чуть иным образом: при верной альтернативной гипотезе $K = \frac{S_X^2}{S_Y^2} \xrightarrow{p} \frac{\sigma_1^2}{\sigma_2^2} > 1$

Если нулевая гипотеза отклоняется, то отклоняется общая гипотеза об однородности. А если основная гипотеза принимается, то применяем критерий Стьюдента

VI. Критерий Стьюдента

Пусть (X_1, \dots, X_n) и (Y_1, \dots, Y_m) из нормальных распределений $X \in N(a_1, \sigma^2)$ и $Y \in N(a_2, \sigma^2)$

Проверяется $H_0 : a_1 = a_2$ против $H_1 : a_1 \neq a_2$

$$\text{Th. } \sqrt{\frac{nm}{n+m}} \frac{(\bar{x} - a_1) - (\bar{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \in T_{n+m-2}$$

По пункту 5 основной теоремы считаем, что числитель и знаменатель независимы

$$\sqrt{\frac{nm}{n+m}} \frac{(\bar{x} - a_1) - (\bar{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} = \sqrt{\frac{nm}{n+m}} \frac{\frac{\bar{x} - a_1}{\sigma} - \frac{\bar{y} - a_2}{\sigma}}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2(n+m-2)}}}$$

По пункту 1 основной теоремы $\sqrt{n} \frac{\bar{x} - a_1}{\sigma}, \sqrt{m} \frac{\bar{y} - a_2}{\sigma} \in N(0, 1) \Rightarrow$

$$\frac{\bar{x} - a_1}{\sigma} \in N\left(0, \frac{1}{n}\right), \frac{\bar{y} - a_2}{\sigma} \in N\left(0, \frac{1}{m}\right) \Rightarrow$$

$$\frac{\bar{x} - a_1}{\sigma} - \frac{\bar{y} - a_2}{\sigma} \in N\left(0, \sqrt{\frac{n+m}{nm}}\right) \Rightarrow$$

$$\sqrt{\frac{nm}{n+m}} \left(\frac{\bar{x} - a_1}{\sigma} - \frac{\bar{y} - a_2}{\sigma} \right) \in N(0, 1)$$

По пункту 3 основной теоремы $\frac{(n-1)S_X^2}{\sigma^2} \in \chi_{n-1}^2, \frac{(m-1)S_Y^2}{\sigma^2} \in \chi_{m-1}^2 \Rightarrow$

$$\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2(n+m-2)} \in \frac{\chi_{n+m-2}^2}{n+m-2}$$

$$\text{Из этого } \sqrt{\frac{nm}{n+m}} \frac{(\bar{x} - a_1) - (\bar{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \in \frac{N(0, 1)}{\frac{\chi_{n+m-2}^2}{n+m-2}} = T_{n+m-2}$$

В качестве статистики возьмем $\sqrt{\frac{nm}{n+m}} \frac{(\bar{x} - a_1) - (\bar{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}}$, по теореме при $a_1 = a_2$ получаем,

что $K \in T_{n+m-2}$

Если верна альтернативная гипотеза, то $K \rightarrow \infty$

Критерий: t_α - квантиль $|T_{n+m-2}|$ уровня α

$$\begin{cases} H_0 : a_1 = a_2, & \text{если } K < t_\alpha \\ H_1 : a_1 \neq a_2, & \text{если } K \geq t_\alpha \end{cases}$$

Nota. Если при обоих критериях согласились с нулевой гипотезой, то соглашаемся с гипотезой об однородности выборок

Nota. Критерий хорошо работает, если выборки из нормальных распределении (или очень близких к ним)