Проверка гипотез о законе распределения

Грауэр Л.В.

Проверка гипотезы о законе распределения

Этапы

- выдвижение гипотезы о виде закона распределения
- оценка параметров распределения
- проверка гипотезы

Критерии согласия

- для простых гипотез
- для сложных гипотез

Критерии согласия для простых гипотез

ξ, $X_{[n]}$

$$H_0 : F_\xi(x) = F_0(x)$$

$$H_1 : F_\xi(x) \neq F_0(x)$$

Критерии согласия, основанные на сравнении

- теоретической плотности распределения и гистограммы
- теоретической и эмпирической функций распределения

Критерий χ^2 для простой гипотезы

Разобьем множество значений ξ на

$\Delta_1, \Delta_2, \ldots \Delta_r$: $\quad \Delta_i = (a_{i-1}, a_i]$, $i = 1, ..., r$

$$p_i = P\{\xi \in \Delta_i | H_0\}$$

$$n_i = num\{X_j \in \Delta_i\}$$

Статистика критерия

$$\chi^2(X_{[n]}) = \sum_{i=1}^{r} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2$$

Теорема К.Пирсона

Если H_0 верна, тогда

$$\chi^2(X_{[n]}) \xrightarrow[n\to\infty]{\mathrm{d}} \zeta \sim \chi^2(r-1)$$

Критическая область и p-value

$\alpha \in (0, 1)$

$V_k = \{\chi^2(X_{[n]}) > \chi^2_{1-\alpha}(r-1)\}$

$p - value = 1 - F_{\chi^2}(\chi^2(X_{[n]}))$

Пример

i	0	1	2	3	4	5	≥ 6
n_i	70	78	34	13	4	1	0

$\xi \sim Pois(1)$?

$n = 200$

$p_0 = 0.368,\ p_1 = 0.368,\ p_2 = 0.184,\ p_3 = 0.061,$
$p_4 = 0.015,\ p_5 = 0.003,\ p_{\geq 6} = 0.001$

i	0	1	2	≥ 3
n_i	70	78	34	18
p_i	0.368	0.368	0.184	0.080
np_i	73.6	73.6	36.8	16
$\frac{(n_i - np_i)^2}{np_i}$	0.18	0.26	0.21	0.25

Критерий Колмогорова для простой гипотезы

Пусть $F_0(x)$ непрерывна на $\mathbb{R}$.

Статистика Колмогорова:

$$D_n(X_{[n]}) = \sup_{x \in R} \left| F_n^*(x) - F_0(x) \right| .$$

Если верна H_0, то

$$D_n(X_{[n]}) \xrightarrow[n\to\infty]{\text{п.н.}} 0$$

Если верна H_1, то

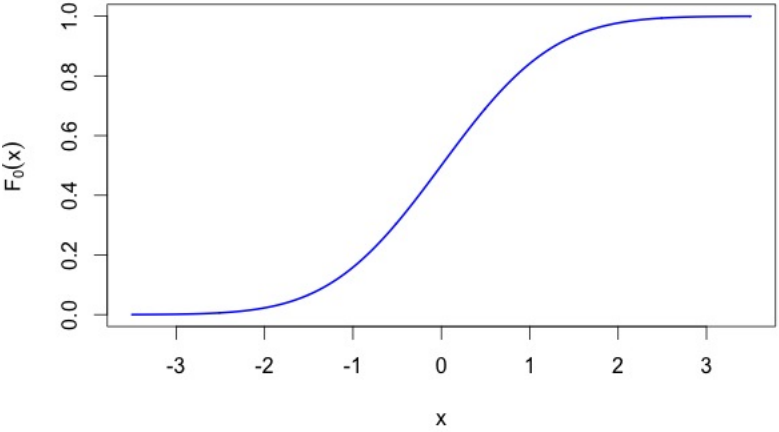
$$D_n(X_{[n]}) \xrightarrow[n\to\infty]{\text{п.н.}} \sup_{x \in R} \left| G(x) - F_0(x) \right| > 0.$$

Теорема Колмогорова

Теорема А.Н. Колмогорова
Если гипотеза H_0 верна, и $F_0(x)$ — непрерывная функция на $\mathbb{R}$, тогда имеет место сходимость:

$$P\{\sqrt{n}D_n(X_{[n]}) < z\} \xrightarrow[n\to\infty]{} K(z) = 1 + 2\sum_{m=1}^{\infty}(-1)^m e^{-2m^2z^2}.$$

$$V_k = \{\sqrt{n}D_n(X_{[n]}) > d_{1-\alpha}\}$$



$F_0(x)$
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
x

Пример

$X_{[n]}$: 0.25, 1.48, 0.32, 0.17, 1.66, 0.29, 0.02, 1.31, 0.12, 3.09

$\xi \sim E(1)$?

$X_{(i)}$	0.02	0.12	0.17	0.25	0.29	0.32	1.31	1.48	1.66	3.09
$F_0(x_{(i)})$	0.02	0.11	0.16	0.22	0.25	0.27	0.73	0.77	0.81	0.95
$F^*(x_{(i)})$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

$D_{10} =$ $\sqrt{10}D_{10} =$

$\alpha = 0.05, \quad d_{1-\alpha} = 1.36$

Критерии согласия для сложных гипотез

ξ, $X_{[n]}$

$$H_0 : F_\xi \in \left\{F(\cdot/\theta) : \theta \in \Theta \subset \mathbb{R}^l\right\}$$

$$H_1 : F_\xi \notin \left\{F(\cdot/\theta) : \theta \in \Theta \subset \mathbb{R}^l\right\}$$

Критерий χ^2 для сложных гипотез

Разобьем множество значений ξ на

$\Delta_1, \Delta_2, \ldots \Delta_k: \quad \Delta_i = (a_{i-1}, a_i],\ i = 1, ..., k$

$$p_i(\theta) = P\{\xi \in \Delta_i | H_0\}$$

$$n_i = num\{X_j \in \Delta_i\}$$

Теорема Фишера

Пусть Θ – открытое множество в $\mathbb{R}^l$.
Пусть выполнены условия:

1. Для любого $\theta \in \Theta$: $\sum_{i=1}^{k} p_i(\theta) = 1$.
2. Для любого $\theta \in \Theta$: $p_i(\theta) > c > 0$ для любого $i = \overline{1,k}$.
3. Для любого $\theta \in \Theta$ существуют и непрерывны производные: $\partial p_i(\theta)/\partial\theta_j$, $\partial^2 p_i(\theta)/(\partial\theta_u \partial\theta_v)$ для любого $i = 1, \ldots, k$, $u, v, j = 1, \ldots, l$.
4. Для любого $\theta \in \Theta$ матрица $\left(\frac{\partial p_i(\theta)}{\partial\theta_j}\right)_{i,j=\overline{1,k}}$ имеет ранг l.

Пусть $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\{n_i\}, \theta)$, *где*

$$L(\{n_i\}, \theta) = \frac{n!}{n_1! \cdot \ldots \cdot n_k!} \prod_{i=1}^{k} p_i^{n_i}(\theta),$$

или $\hat{\theta}$ *— оценка по методу минимума хи-квадрат:*

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(n_i - np_i(\theta))^2}{np_i(\theta)}.$$

Тогда, если гипотеза H_0 *верна, то*

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{k} \frac{(n_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow[n \to \infty]{d} \zeta \sim \chi^2_{k-l-1}.$$

Критическая область и p-value

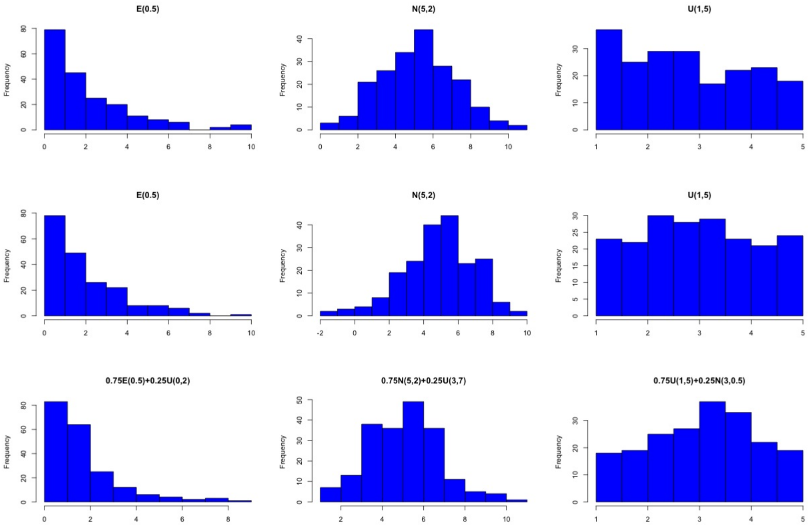
$$\chi^2(\hat{\theta}) = \sum_{i=1}^{k} \frac{(n_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$

$\alpha \in (0, 1)$

$V_k = \{\chi^2(\hat{\theta}) > \chi^2_{1-\alpha}(k - l - 1)\}$

$p - value = 1 - F_{\chi^2}(\chi^2(\hat{\theta}))$

Критерий Колмогорова для сложных гипотез?



Оценки вероятности ошибки 1го рода

Гипотеза	$E(0.5)$	$N(5,2)$	$U(1,5)$
Простая	0.048	0.044	0.048
Сложная	0.004	0.000	0.045

Оценки мощности

Гипотеза	$0.75E + 0.25U$	$0.75N + 0.25U$	$0.75U + 0.25N$
Простая	0.596	0.083	0.521
Сложная	0.028	0.001	0.492

Возможный вариант:
разбить выборку на 2 части:

- построить оценки по 1й части
- проверить гипотезу по 2й части