STATISZTIKA 1: nemparaméteres próbákhoz kiegészítő
Dr. Bolla Marianna, 2020.05.05.

Kolmogorov–Szmirnov próba

Egymintás eset (illeszkedésvizsgálat):

$H_0 : n-$elemű fae. mintánk az F folytonos eloszlásfüggvényű eloszlásból származik.

Kétoldali alternatíva:

$$H_1 : \mathbb{P}(X < x) \neq F(x).$$

Legyen

$$D_n = \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)|.$$

Amennyiben $x_1^* \leq \cdots \leq x_n^*$ az $\mathbf{x} = (x_1, \ldots, x_n)$ mintarealizáció rendezett alakja; $y_i = F(x_i)$, $y_i^* = F(x_i^*)$ $(i = 1, \ldots, n)$ a (0,1)-en egyenletesbe transzformált minta, akkor

$$\begin{aligned}
D_n(\mathbf{x}) &= \max_i \max\{|F_n^*(x_i^*) - F(x_i^*)|, |F_n^*(x_i^* + 0) - F(x_i^*)|\} = \\
&= \max_i \max\{|\frac{i-1}{n} - F(x_i^*)|, |\frac{i}{n} - F(x_i^*)|\} \\
&= \max_i \max\{|\frac{i-1}{n} - y_i^*)|, |\frac{i}{n} - y_i^*|\} = \\
&= \max_i \max\{|G_n^*(y_i^*) - G(y_i^*)|, |G_n^*(y_i^* + 0) - G(y_i^*)|\} = D_n(\mathbf{y}),
\end{aligned}$$

ahol G_n^* az y-minta empirikus eof.-e és $G(y) = y$, $0 < y < 1$.

A Kolmogorov-tétel alapján tudjuk, hogy H_0 fennállása esetén

$$\lim_{n \to \infty} \mathbb{P}(\sqrt{n} D_n < z) = K(z), \quad \forall z \in \mathbb{R},$$

ahol

$$K(z) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 z^2} = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2}, \quad z > 0.$$

A Kolmogorov-eof. számunkra fontos értékei az alábbi táblázatban találhatók:

z	1.14	1.22	1.36	1.52	1.63	k_α
$K(z)$	0.85	0.90	0.95	0.98	0.99	$1 - \alpha$

Így "elég nagy" n esetén az α szignifikanciához (I fajú hibához) tartozó kritikus tartomány:

$$\mathcal{X}_k = \{\mathbf{x} : \sqrt{n} D_n(\mathbf{x}) \geq k_\alpha\},$$

ahol k_α a Kolmogorov-eloszlás $1 - \alpha$ kvantilise.

Egyoldali alternatíva:

$$H_0 : \mathbb{P}(X < x) \leq F(x) \qquad H_1 : \mathbb{P}(X < x) > F(x).$$

Legyen

$$D_n^+ = \sup_{x\in\mathbb{R}}(F_n^*(x) - F(x)) = \max_i \max\{(F_n^*(x_i^*) - F(x_i^*), (F_n^*(x_i^* + 0) - F(x_i^*))\}.$$

A Szmirnov-tétel alapján tudjuk, hogy

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}D_n^+ < z\right) = S(z), \quad \forall z \in \mathbb{R},$$

ahol

$$S(z) = 1 - e^{-2z^2}, \quad z > 0.$$

A Szmirnov–eloszlásfüggvény értékei (és inverze is) számolható explicite.

$$\mathcal{X}_k = \{\mathbf{x} \,:\, \sqrt{n}D_n^+(\mathbf{x}) \geq k_\alpha\},$$

ahol k_α a Szmirnov-eloszlás $1 - \alpha$ kvantilise.

Kétmintás eset (homogenitásvizsgálat):

H_0 : n-elemű és m-elemű, egymástól is független fae. mintáink ugyanaból a folytonos eloszlásfüggvényű eloszlásból származnak, azaz

$$H_0 :\, F = G.$$

Kétoldali alternatíva:

$$H_1 :\, F \neq G.$$

Legyen

$$D_{n,m} = \sup_{x\in\mathbb{R}} |F_n^*(x) - G_m^*(x)|,$$

ami megint véges maximum.

A Kolmogorov-tétel alapján tudjuk, hogy H_0 fennállása esetén

$$\lim_{n,m\to\infty} \mathbb{P}(\sqrt{\frac{nm}{n+m}}D_{n,m} < z) = K(z), \quad \forall z \in \mathbb{R}.$$

Ezért

$$\mathcal{X}_k = \{\mathbf{x}, \mathbf{y} \,:\, \sqrt{\frac{nm}{n+m}}D_{n,m}(\mathbf{x}, \mathbf{y}) \geq k_\alpha\},$$

ahol k_α a Kolmogorov-eloszlás $1 - \alpha$ kvantilise.

Egyoldali alternatíva:

$$H_0 : F \leq G \qquad H_1 : F > G.$$

Legyen

$$D_{n,m}^{+} = \sup_{x \in \mathbb{R}} (F_n^*(x) - G_m^*(x)),$$

ami megint véges maximum.

A Szmirnov-tétel alapján tudjuk, hogy H_0 fennállása esetén

$$\lim_{n,m \to \infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}} D_{n,m}^{+} < z\right) = S(z), \quad \forall z \in \mathbb{R}.$$

Ezért

$$\mathcal{X}_k = \{\mathbf{x}, \mathbf{y} : \sqrt{\frac{nm}{n+m}} D_{n,m}^{+}(\mathbf{x}, \mathbf{y}) \geq k_\alpha\},$$

ahol k_α a Szmirnov-eloszlás $1 - \alpha$ kvantilise.

A χ^2-próba előnye a Kolmogorov–Szmirnov próbával szemben, hogy diszkrét eloszlásokra is alkalmazható; viszont folytonos eloszlásokat is diszkretizálni kell, így információt vesztünk.