**11. STATISZTIKA GYAKORLAT: Nemaraméteres próbák**

# 1 Kolmogorov–Szmirnov próba

## 1.1 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

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

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

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

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

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

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

**Egyoldali alternatíva:**

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

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.

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

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

## 1.2 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:**

$$H_1 : F \ne G.$$

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

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

ahol $k_\alpha$ 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

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

Ezért

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

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

## 2 $\chi^2$-próba

Ld. könyv és képletgyűjtemény. Előnye a Kolmogorov–Szmirnov próbával szemben, hogy diszkrét eloszlásokra is alkalmazható; viszont folytonos eloszlásokat is diszkretizálni kell, így információt vesztünk.

## 3 Feladatok példatár 3.2 részből

- – Szabályosnak tekinthető-e az alábbi 100-szor feldobott dobókocka $\alpha = 0.05$ szignifikanciával?

  | kategória | 1 | 2 | 3 | 4 | 5 | 6 |
  |---|---|---|---|---|---|---|
  | gyakoriság | 24 | 21 | 19 | 12 | 13 | 11 |

  $\chi^2 = 8.72 < \chi^2(5)$, az 5 szab. fokú $\chi^2$ eo. $1 - 0.05 = 0.95$ kvantilise, ezért nem tudjuk elutasítani $H_0$-t, a kocka szabályos.

  – Származhat-e mintánk [0.5,6.5]-ön folytonos egyenletes eloszlásból, ha az ekvidisztans felosztásban a gyakoriságok a fentiek? ($\alpha = 0.05$)

  Megint
  $$H_0: p_i = 1/6, \quad i = 1, \ldots, 6$$
  így ugyanaz a $\chi^2$ adódik.

  Ha 100-elemű mintánk konkrét [0.5,6.5]-beli értékeit is ismernénk, akkor egymintás, kétoldali Kolmogorov–Szmirnov próbát hajthatnánk végre.

- Jó-e a RANDOM generátor? Származhat-e az alábbi 20 szám (az egyszerűség kedvéért nagyság szerint rendezve) [0,1]-en egyenletes eloszlásból?

  0.03, 0.05, 0.08, 0.08, 0.11, 0.12, 0.14, 0.19, 0.23, **0.24**,
  0.31, 0.43, 0.45, 0.56, 0.61, 0.67, 0.69, 0.78, 0.85, 0.88

- Most egymintás, kétoldali Kolmogorov–Szmirnov próbát hajthatunk végre (bár 20 nem kimondottan ‘nagy’):

$$D_{20}(\mathbf{y}) = \max_{i=1}^{20} \max\{|G^*_{20}(y^*_i) - y^*_i)|, |G^*_{20}(y^*_i + 0) - y^*_i|\} =$$
$$= |G^*_{20}(y^*_{10} + 0) - y^*_{10}| = |\frac{10}{20} - 0.24| = 0.26.$$

  Mivel $\sqrt{20}D_{20} = 1.16$, a Kolmogorov-táblázat alapján $1-\alpha$ valamivel kisebb, mint 0.9, $\alpha$ pedig valamivel nagyobb, mint 0.1. Ez a szignifikancia nem elég $H_0$ elutasításához, ezért úgy döntünk, hogy jó a véletlen szám generátor (nagyobb minta azért bizonyítóbb erejű lett volna).

- $\chi^2$-próbával 4 ekvidisztans intervallumra osztjuk [0,1]-et és a

$$H_0 : p_i = 1/4, \quad i = 1, 2, 3, 4$$

  null-hipotézist teszteljük:

| intervallum | 0–0.25 | 0.25–0.5 | 0.5–0.75 | 0.75–1 |
|---|---|---|---|---|
| $\nu_i$ | 10 | 3 | 4 | 3 |
| $20p_i$ | 5 | 5 | 5 | 5 |

$$\chi^2 = \frac{25 + 4 + 1 + 4}{5} = 6.8,$$

  ami valamivel nagyobb, mint a $\chi^2(3)$-eloszlás 0.9 kvantilise, $\alpha$ pedig valamivel (de csak kevéssel) kisebb, mint 0.1. Tehát $\chi^2$-próbával valamivel kisebb szignifikanciával (I. fajú hibával) tudnánk elutasítani $H_0$-t, mint a Kolmogorov–Szmirnov próbával. Azonban ez a szignifikancia sem elég $H_0$ elutasításához, ezért inkább úgy döntünk, hogy jó a véletlen szám generátor (nagyobb minta itt is bizonyítóbb erejű lett volna).

  Pl. $\alpha = 0.05$ mellett, mindkét próbával el kell fogadnunk $H_0$-t.

- Fizetések (példatár 1.1)
  - Normális eo.-t követnek-e becsült paraméterekkel? Nem, mert van egy kiugró érték.
  - Pareto eo.-t követnek-e becsült paraméterekkel? (ld. Hf.)

    Pareto: olasz közgazdász vezette be ezt az eo.-t a javak egyenlőtlen elosztására (XX sz. első fele).